Privacy Motivated Models for Deep Learning-based Human Action Recognition in Intelligent Video Surveillance

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

M.Sc.

Mickael Willy Jacky Cormier

aus Dieppe, Frankreich

Tag der mündlichen Prüfung:	03.11.2025
Erster Gutachter:	Prof. Dr.-Ing. habil. Jürgen Beyerer
Zweiter Gutachter:	Prof. Dr. Rikke Gade

Abstract

Human action recognition in surveillance settings plays a central role in ensuring public safety, occupational security, and forensic investigation. In contrast to appearance-based approaches that rely on storing sensitive visual data, this thesis focuses on skeleton-based representations of human motion and posture, thereby reducing privacy concerns while maintaining high discriminative power in complex, multi-camera environments. To support this research direction, multiple new datasets, methodological contributions, and benchmarks are introduced.

First, the UPAR dataset is proposed, harmonizing semantic attributes across four existing datasets and contributing over 3.3 million binary annotations. Its extension, UPAR-Pose, introduces more than 71,000 joint pose-and-attribute annotations, enabling large-scale studies of multitask learning that jointly address pedestrian attribute recognition and two-dimensional human pose estimation. Experiments with a multitask framework demonstrate that shared feature representations improve attribute recognition and action analysis while remaining computationally efficient compared to independent models.

Second, to address the challenges of adverse illumination, the Low-Light Visible-Infrared Paired - Pose (LLVIP-P)-Pose benchmark is presented, providing more than 26,000 paired visible and thermal pose annotations. Building on this, the Unified Pedestrian Pose Estimation in Thermal Imaging (UPPET) dataset is introduced as the largest thermal human pose dataset to date, comprising 33,654 images with 118,924 annotations. UPPET harmonizes multiple thermal datasets with consistent keypoints and evaluation protocols, supporting robust cross-sensor benchmarking for surveillance and occupational safety.

On the methodological side, two complementary multitask models are explored. One combines two-dimensional pose estimation with pedestrian attributes, while the other integrates three-dimensional pose estimation with joint orientation prediction. This dual design improves robustness under domain shifts while keeping each task specialized. A kinematics-aware 3D pose estimation framework further enhances anatomical plausibility by incorporating bone-length regularization, joint-limit constraints, and temporal fusion. The dual-branch architecture predicts root-relative coordinates alongside per-joint orientations, reducing jitter and ambiguity in challenging scenarios. Experiments reveal that joint orientations improve interpretability and robustness only when estimated with sufficient accuracy, highlighting trade-offs between complexity and reliability.

To improve generalization, a range of skeleton-based data augmentation strategies are introduced, including occlusion, truncation, temporal jitter, scale and viewpoint variation, and fragmented identities. These augmentations prove effective for cross-sensor transfer and downstream action recognition.

Finally, a synthetic surveillance benchmark is contributed, rendering naturalistic human motion in a game engine as a privacy-conscious alternative to real-world recordings. This benchmark covers the entire pipeline—from two-dimensional pose estimation and three-dimensional reconstruction to skeleton-based action recognition—allowing systematic analysis of error propagation and component-level evaluation.

Together, these contributions establish a comprehensive framework for skeleton-based action recognition that balances accuracy, robustness, and privacy. The proposed datasets, methods, and benchmarks provide practical guidelines for integrating multitask learning, kinematic constraints, skeleton-based data augmentation, and synthetic data. The results demonstrate the feasibility of privacy-conscious action recognition in realistic surveillance environments and lay the foundation for future research and deployment in ethically aligned, human-centered applications.

Kurzfassung

Die Erkennung menschlicher Aktivitäten in Videoüberwachungsszenarien spielt eine zentrale Rolle für die öffentliche Sicherheit, die Arbeitssicherheit sowie die forensische Analyse. Im Gegensatz zu erscheinungsbasierten Verfahren, die auf die Speicherung sensibler Bilddaten angewiesen sind, konzentriert sich diese Arbeit auf skelettbasierte Repräsentationen von Bewegungen und Körperhaltungen. Dadurch wird der Datenschutz verbessert, ohne auf die diskriminative Aussagekraft in komplexen Umgebungen mit mehreren Kameras zu verzichten. Zur Unterstützung dieses Forschungsansatzes werden in dieser Arbeit mehrere neue Datensätze, methodische Beiträge und Benchmarks vorgestellt.

Zunächst wird der UPAR-Datensatz eingeführt, der semantische Attribute über vier bestehende Datensätze hinweg harmonisiert und mehr als 3,3 Millionen neue binäre Annotationen bereitstellt. Die Erweiterung UPAR-Pose ergänzt über 71.000 kombinierte Posen- und Attributannotationen und ermöglicht damit aussagekräftige Studien zum Multitask-Lernen, die gleichzeitig eine Personenattributerkennung und eine zweidimensionale Posenschätzung adressieren. Experimente mit einem Multitask-Framework zeigen, dass gemeinsame Merkmalsrepräsentationen sowohl die Attributerkennung als auch die Handlungsanalyse verbessern und dabei recheneffizienter sind als getrennte Modelle.

Darüber hinaus wird mit LLVIP-P-Pose ein Benchmark für Szenarien mit schwierigen Lichtverhältnissen vorgestellt, der mehr als 26.000 gepaarte RGB- und thermische Posenannotationen enthält. Aufbauend darauf wird UPPET eingeführt – der mit 33.654 Bildern und 118.924 Annotationen bislang größte thermische Datensatz für die Erkennung menschlicher Posen. UPPET vereinheitlicht mehrere bestehende thermische Datensätze durch eine

konsistente Definition von Posen-Keypoints und führt Evaluationsprotokolle ein, die robuste sensorübergreifende Auswertungen für Überwachungs- und Arbeitssicherheitsszenarien ermöglichen.

Auf methodischer Ebene werden zwei komplementäre Multitask-Modelle entwickelt. Das eine kombiniert die zweidimensionale Posenschätzung mit der Erkennung von Personenattributen, während das andere dreidimensionale Posen mit Gelenkorientierungen integriert. Diese duale Architektur erhöht die Robustheit gegenüber unterschiedlichen Domänen, während die Spezialisierung der einzelnen Aufgaben erhalten bleibt. Ein vorgeschlagener kinematikbewusster 3D-Poseschätzer verbessert zudem die anatomische Plausibilität, indem er Knochenlängen regularisiert, Gelenkwinkel begrenzt und eine zeitliche Fusion einbezieht. Die zweigleisige Architektur liefert sowohl Koordinaten relativ zu einem Wurzelgelenk als auch Gelenkorientierungen und reduziert dadurch Zittern und Mehrdeutigkeiten in schwierigen Szenarien. Experimente zeigen, dass Gelenkorientierungen die Interpretierbarkeit und Robustheit nur dann steigern, wenn sie mit hinreichender Genauigkeit vorhergesagt werden – was die Abwägung zwischen Modellkomplexität und Zuverlässigkeit verdeutlicht.

Zur Verbesserung der Generalisierungsfähigkeit werden verschiedene skelettbasierte Datenaugmentierungsstrategien eingeführt, darunter Verdeckung, Abschneiden, zeitliches Zittern, Skalierungs- und Perspektivvariationen sowie fragmentierte Posen. Diese Methoden erweisen sich als effektiv für sensorübergreifende Szenarien und die nachgelagerte Aktivitätsanalyse.

Abschließend wird ein synthetischer Überwachungsdatensatz präsentiert, der naturgetreue menschliche Bewegungen in einer Computerspiel-Engine rendert und damit eine datenschutzfreundliche Alternative zu realen Videoaufnahmen bietet. Dieser Benchmark deckt die gesamte Pipeline ab – von der zweidimensionalen Posenschätzung über die dreidimensionale Rekonstruktion bis hin zur skelettbasierten Aktivitätserkennung – und erlaubt eine systematische Analyse der Fehlerfortpflanzung sowie komponentenbasierte Evaluationen.

Zusammenfassend etablieren diese Beiträge ein umfassendes Rahmenwerk für die skelettbasierte Aktivitätserkennung, das Genauigkeit, Robustheit und Datenschutz in Einklang bringt. Die vorgeschlagenen Datensätze, Methoden und Benchmarks liefern praxisrelevante Leitlinien für den Einsatz von Multitask-Lernen, kinematischen Nebenbedingungen, skelettbasierter Datenaugmentierung und synthetischen Daten. Die Ergebnisse zeigen die Machbarkeit datenschutzbewusster Handlungserkennung in realistischen Überwachungsumgebungen und legen die Grundlage für zukünftige Forschung und praxisnahe Anwendungen in menschenzentrierten, ethisch verantwortungsvollen Kontexten.

Acknowledgements

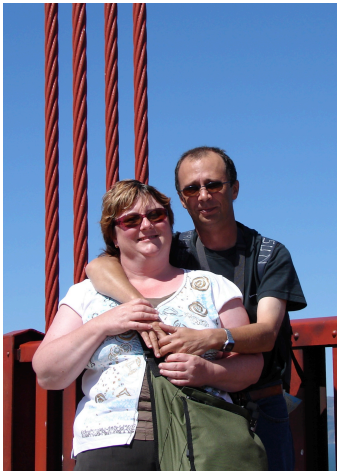
The research conducted over the past five and a half years at the Vision and Fusion Laboratory (IES) of the Karlsruhe Institute of Technology (KIT), which led to this dissertation, would not have been possible without the continuous guidance, encouragement, and inspiration of my advisors, mentors, colleagues, friends, and family. First, I want to express my deepest gratitude to Prof. Dr.-Ing. habil. Jürgen Beyerer for the opportunity to work at IES and for guiding me to grow as a researcher. This thesis would not have been possible without his guidance and support. His passion, patience, and vast knowledge continuously inspired me to pursue my goal. I am grateful to Prof. Dr. Rikke Gade for agreeing to review my thesis and for traveling to attend my defense.

This work results from close collaboration with the Department of Video Exploitation Systems (VID) at Fraunhofer IOSB. I am especially grateful to my friend and colleague Dr.-Ing. Andreas Specker, who has been there for me in good and bad times. Many valuable ideas, papers, and competitions arose from our discussions, and I am glad to continue working together after our dissertations. Thanks also to Dr.-Ing. Daniel Stadler, Dr.-Ing. Jürgen Metzler, and Stefan Wolf for their counsel.

I acknowledge the Polizeipräsidium Mannheim – Projekt Videoschutz and the Polizei Hamburg SP32 for their support, as well as the BMBF Software Campus Program for funding, and especially Dr. Michael Schäffer and Dr. Benjamin Blaß at SHS for their collaboration. I'm also thankful to Prof. Kamal Nasrollahi, Prof. Sergio Escalera, Prof. Julio C. S. Jacques Junior, and Anthony Hoogs for enabling Andreas and me to organize challenges and to host a workshop at the IEEE/CVF Winter Conference on Applications of Computer Vision during my PhD.

I would also like to thank my family and friends for the great moments spent together; I would not have made it without their support and understanding. My mother Valérie and my father Pascal, who is unfortunately no longer with us and could not share this milestone with us. Thank you for your guidance and encouragement. Thanks to many others over the years, especially to Dr.-Ing. Michael Teutsch, Dr.-Ing. Monica Haurilet, and Dr.-Ing. Chengchao Qu, with whom I learned so much about computer vision and deep learning, inspiring and enabling me to pursue my PhD. To my students, past and present, I offer gratitude for their theses, implementations, and annotations.

Finally, my deepest thanks to my wife, Yuliia Cormier, my anchor—supporting, encouraging, and grounding me when times were toughest. I couldn't have done this without you.



Dedicated to my parents Valérie and Pascal.

Contents

Appendix

1 Introduction

This thesis investigates privacy-motivated models for robust human analysis in intelligent video surveillance under a top-down setting with available detections and tracks. Based on available tracks from ealier staged, the aim of this work is to recognize actions from person in the field of view of a camera and to generate a notification, if a salient activity is recognized. The focus lies on three interconnected components—2D Human Pose Estimation (2D-HPE), root-relative 3D Human Pose Estimation (3D-HPE), and Skeleton Based Action Recognition (SBAR)—and on their interactions under realistic constraints. Two new real-image benchmarks, Unified Pedestrian Attribute Recognition and Pose Estimation (UPAR-Pose) (RGB) and UPPET (thermal), are introduced for 2D-HPE, with thermal imagery and pose-centric evaluation supporting privacy-aware study. In addition, a synthetic surveillance dataset, GTA Real World Surveillance (GTA-RWS), is proposed in which real human motion is retargeted into the environment to obtain realistic, contact-rich actions for SBAR under dense occlusions and oblique viewpoints, while avoiding collection of identifiable personal data. The work defines unified, privacy-conscious protocols and metrics, derives task-specific subsets from GTA-RWS for 2D-HPE, 3D-HPE, and SBAR, and conducts a system-level evaluation to quantify accuracy and cross-stage effects across these components.

The remainder of this chapter is organized as follows. Section 1.1 presents the motivation. Section 1.2 outlines key challenges, including those emerging from image acquisition, human pose estimation and real-world constraints. The main contributions, including UPAR-Pose, UPPET, and GTA-RWS, are summarized in Section 1.3. Finally, Section 1.4 describes the thesis outline.

1.1 Motivation

In the field of computer vision, AI-based human action recognition in intelligent video surveillance is increasingly shaped by the dual imperative of functionality and privacy compliance. With the proliferation of high-resolution cameras and real-time analytics infrastructure in urban environments, there is growing demand for systems that can detect and respond to abnormal or threatening behavior in public spaces such as train stations, city centers, and large events. At the same time, these systems must navigate stringent regulatory environments, particularly in the European Union, where the General Data Protection Regulation (GDPR) and the AI Act, now in full legal effect as of 2025, establish strict constraints on data collection, processing, and biometric identification [Eur25]. Furthermore, video surveillance from law enforcement, is subject to further limitations [Bun25]. Such measures are subject to strict requirements, including necessity, proportionality, and a clear legal basis. To address these challenges, recent surveillance architectures increasingly rely on abstracted representations of human bodies. Rather than processing raw visual features such as faces or full silhouettes, these systems typically follow a three-stage pipeline: 2D-HPE, which aims at extracting keypoint from the image of a person and represent them as kinematics skeleton, optionally 3D pose reconstruction, and skeleton-based action recognition [Gol23, Gol22, Cor22b, Cor24a, Dua22b]. This layered structure is designed to reduce identifiability while retaining semantically rich movement information that can be used to classify behaviors such as fighting, falling, or fleeing. Crucially, this design philosophy aligns with the AI Act's risk-based framework, which designates real-time biometric surveillance as "high-risk" and mandates transparency, oversight, and harm minimization.

However, as Schabacher observes, architectural abstraction does not eliminate the political and normative stakes of surveillance: it transforms them [Sch23a]. The shift to skeletal data and movement trajectories introduces new forms of opacity. Categories such as "aggression" or "suspicious behavior" are not universal but contextually and institutionally defined. When such categories are encoded into technical systems, they become entangled with

statistical learning mechanisms that derive behavioral norms from datasets whose assumptions often remain unexamined. Schabacher calls this recursive logic "entangled," emphasizing how technical classification becomes inseparable from institutional judgment. Surveillance, in this context, is not merely observational, it becomes a mode of codifying and enforcing implicit behavioral expectations. This is also pointed out by privacy advocates who fear that such technologies could be disproportionately enforced against homeless and underprivileged populations [Red23].

This entanglement is evident in current real-world deployments. Since 2018, the city of Mannheim has operated an AI-enhanced video surveillance assistance system under a legal pilot regime [Kes23]. The system captures live footage, extracts pose skeletons, and evaluates them against learned models of salient behavior. The original video is only accessed for human verification, only when activities are flagged. Reports cite reduced emergency response times and increased public sense of safety [Her23, Hei24a, Hei24b, Lan24]. Correspondingly, a 2013 survey across Europe—covering Germany as well—found that 81% of respondents agreed that CCTV systems help reduce crime [You13]. Furthermore, a 2017 survey conducted in Germany revealed that 81% of respondents associate increased video surveillance with greater personal security [Sta17]. This widespread perception underscores the strong public acceptance for effective safety measures in public spaces, which surveillance systems aim to address. Nonetheless, this acceptance coexists with persistent privacy concerns and debates around civil liberties [Sch23b].

Similarly, Hamburg launched its IVBeo pilot in 2023, which enabled successful intervention in a physical assault case where no emergency call had been made [Pol23a]. These systems avoid traditional biometrics and facial recognition, aiming to maintain situational awareness while respecting privacy law [Gol23].

Yet, these efforts reveal a persistent paradox. On one hand, training and evaluating action recognition models requires access to representative surveillance data. On the other hand, legal and ethical frameworks prohibit unrestricted data collection in public spaces. The GDPR and the AI Act enforce strict data

minimization principles, which limit not only the use but also the type and amount of data that can be collected.

Projects in Mannheim and Hamburg operate as controlled test environments. They are deployed in officially designated *crime hotspots* [Lan21, Pol23b, Pol23c], a legal classification that defines the specific conditions required for enhanced surveillance. While this framework enables limited data acquisition under close regulatory oversight, it has still drawn criticism from civil society organizations and privacy advocates [Lul23, Sch25, Ham25]. Importantly, these deployments adhere to strict legal frameworks, operate under official pilot regimes, and exemplify a responsible approach to AI-surveillance. In contrast, other implementations have raised legal and privacy concerns: in France, several retail chains were found using AI-based surveillance systems without proper authorization [Str24]. In Reims, local authorities deployed AI-powered cameras in public spaces under questionable oversight [Str23], and in Germany, the Federal Constitutional Court has emphasized the illegality of certain unsupervised surveillance practices [Bun23]. These cases and more [Lib24, Lib25] highlight the risks associated with uncontrolled deployment and the potential for privacy violations.

A key concern involves the assumed anonymity of skeletal data. While pose skeletons abstract away visual identity markers such as faces or clothing, they are not inherently anonymous. Gait recognition—the identification of individuals based on movement patterns—has advanced considerably and is now actively researched in defense and intelligence domains, including in IARPA-funded projects in the United States [Liu24, Liu25a]. When skeletal data is combined with contextual metadata such as time, location, and body proportions, there is a growing fear that individuals could become re-identifiable [Sch25].

In real-time deployments, these systems typically retain the image and pose data only temporarily, often for a few minutes, before discarding it. The more serious privacy concern arises during the training phase, where large volumes of annotated skeletal data together with clear images are needed to build and evaluate recognition models. If such data is collected from real individuals

in uncontrolled public settings, there is a fear that such capabilities could enable long-term re-identification and behavioral profiling of individuals [Lib24, Lib25]. Even without capturing facial features or explicit biometric markers, recurring patterns in movement and context may, over time, compromise anonymity and undermine the privacy-preserving intentions of these systems.

At the same time, there are pragmatic limitations. Annotated surveillance datasets are expensive to produce, especially for nuanced human actions under real-world conditions. As shown in [Cor21b], the cost of manual annotation-particularly at the scale needed for robust model training-, is often prohibitive and resource-intensive. This creates pressure to either reuse publicly available datasets or rely on synthetic data, both of which raise issues of realism, domain transferability, and ethical applicability.

This thesis builds upon the architectural principles demonstrated in recent pilot deployments in Mannheim and Hamburg, which use skeleton-based action recognition to reduce reliance on raw biometric data. Motivated by privacy concerns, this work evaluates the generalizability of 2D-HPE models across RGB as well as thermal imaging modalities. In addition, it investigates a multitask learning approach that jointly performs 2D pose estimation and pedestrian attribute recognition, enabling the extraction of soft biometric features such as clothing and semantic attributes. This multitasking strategy supports suspect description and rapid intervention while maintaining privacy-preserving characteristics.

Additionally, the thesis employs normalized root-relative 3D pose representations and skeleton-based action recognition pipelines to minimize raw biometric data processing.

To address challenges related to data scarcity and privacy, this work explores the use of publicly available datasets alongside realistically rendered synthetic motion data. While synthetic data cannot fully replace real-world variability, it allows robust model training without overreliance on sensitive recordings. The envisioned system targets public safety contexts in smart city environments, where rapid response to abnormal human activity is critical but must

occur within strict legal and ethical constraints. Use cases include crowd monitoring at transit hubs, large events, or other designated public spaces, while respecting the right to peaceful assembly by restricting deployment during demonstrations or protests.

Designed as a modular, real-time analytics component, the system integrates with existing surveillance infrastructure, maintaining auditability and minimizing human access to raw video data. Transparent logging and strict access controls ensure accountability and regulatory compliance. The central aim is to reconcile performance, scalability, and privacy. Rather than treating privacy as a constraint, it is a foundational design principle embedded in both the architecture and the data pipeline. Technical contributions include demonstrating competitive action recognition accuracy under privacy-aware constraints, evaluating generalizability across synthetic and real datasets, and proposing audit-friendly designs aligned with the AI Act.

No claim is made that these video surveillance techniques constitute a solution to all societal problems. Rather, this research focuses on understanding specific challenges, potential risks, and requirements for transparency and explainability, contributing to knowledge about how AI-based monitoring can be responsibly and ethically deployed.

1.2 Challenges

Although research on 2D-HPE, 3D-HPE and SBAR has advanced significantly, reliable operation in surveillance environments remains difficult. The complexity arises from diverse camera characteristics, uncontrolled environments, and inherent task limitations. These challenges can be categorized into three groups: those emerging from image acquisition, those specific to 2D-HPE and SBAR, and those arising in real-world deployment. Each group is discussed in the following subsections.

1.2.1 Challenges Emerging from Image Acquisition

Difficulties at the acquisition stage stem from the physical properties of cameras, their placement, and environmental influences that determine signal quality. Large-scale surveillance networks combine heterogeneous devices with different optics, frame rates, compression schemes, and calibration states. This variability introduces inter-camera inconsistencies that models must handle without explicit recalibration. Moreover, unlike controlled laboratory settings, deployed systems operate continuously and must cope with day–night cycles, seasonal conditions, and artificial lighting. Reflections, glare (Figure 1.1), and saturation reduce visibility, while steep mounting positions needed for wide-area coverage amplify occlusion and scale variation, as illustrated in Figure 1.2.

Figure 1.1: Illumination-related challenges – Glare from reflective surfaces (*e.g.*, mobile devices), lens distortion, strong shadows, and structural elements such as walls can obscure human shapes or cause false positive detections.

Figure 1.2: Challenges emerging from image acquisition – Urban scenes introduce occlusions from trams, cables, and other people. Varying person sizes due to camera angles and misleading objects (*e.g.*, strollers, bicycles) reduce accuracy. The model, not trained in-domain, shows strong keypoint predictions on visible, unoccluded persons but confuses left/right sides and misattributes keypoints to non-human objects.

In the following specific challenges emerging from image acquisition are described.

Low spatial resolution: In wide-angle coverage, individuals often occupy only a few pixels. Compression and downscaling compound this loss of detail, hindering joint localization and limb association. The reduced spatial precision diminishes the discriminative value of subtle motion cues, which are critical for distinguishing between fine-grained actions.

Viewpoints: Steep or oblique viewpoints distort body proportions and conceal key joints. Wrists, ankles, or shoulders may remain permanently occluded, producing incomplete skeletons. Such systematic blind spots propagate through the recognition pipeline and limit accuracy, especially for actions dependent on extremity movement.

Illumination: Transitions between daylight, artificial lighting, and adverse weather alter brightness and contrast. Overexposure from headlights or specular reflections produces glare, while fog or overcast skies yield low contrast.

These shifts move data distributions away from training conditions, reducing the stability of pose estimation.

Limited frame rate: To manage storage and bandwidth, many cameras operate at reduced or variable frame rates. The resulting motion blur and temporal aliasing compromise the detection of rapid or subtle gestures. Temporal smoothing cannot reliably compensate for missing or inconsistent frames.

Camera types: Differences in sensor technology, lenses, shutter mechanisms, and compression pipelines yield diverse noise and appearance profiles. Without consistent calibration, multi-camera fusion becomes unreliable, complicating spatial reasoning and cross-view tracking.

In summary, acquisition-related constraints generate heterogeneous and frequently degraded input signals. These effects impact nearly every stage of the vision pipeline and form a structural barrier for robust 2D-HPE and SBAR in surveillance contexts. Reliable deployment therefore requires methods that can absorb such variability while maintaining stable downstream recognition.

1.2.2 Human Pose Estimation-specific Challenges

Beyond acquisition-related effects, both 2D-HPE and SBAR exhibit limitations that stem from their methodological assumptions. Since action recognition relies on pose sequences, weaknesses in keypoint estimation directly propagate into the higher-level representation.

In the following specific challenges related to Human Pose Estimation (HPE) are described.

Occlusion: Crowded or cluttered environments frequently produce partial or complete occlusions. These may arise from other individuals, static infrastructure, moving objects, or self-occlusion. Missing or uncertain joints fragment skeletons and introduce kinematic inconsistencies, which in turn distort action descriptors (Figure 1.3).

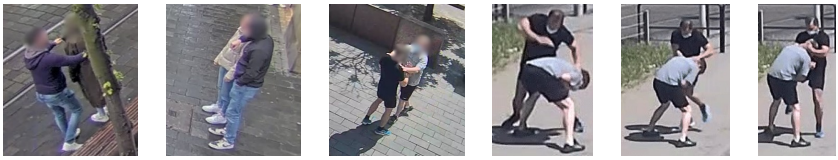


Figure 1.3: Human pose estimation-specific challenges – This figure shows examples of two pairs of individuals captured from different viewpoints. In the first image, the person on the right is occluded by a tree and the other individual. In the second, the farther person is partially hidden behind the closer one. The last four images depict a fight, where heavy mutual and self-occlusion significantly complicates accurate keypoint detection.

Viewpoints: Extreme camera angles reduce the visibility of keypoints and create ambiguities between left and right body parts. Top-down perspectives, common in wide-area surveillance, make head orientation and small joints difficult to estimate. These errors lower pose accuracy and impair downstream recognition (Figure 1.4).

Figure 1.4: Different viewpoints – This figure shows samples for two individuals captured from different point of views. The first two images depict the individuals from behind. The next two from the side and the last two from a steep perspective. Especially, locating keypoints suffers from varying viewpoints due to different appearance and occlusions when seen from the front, back or side.

Small Detections: Performance remains acceptable for large bounding boxes but degrades sharply for subjects smaller than 32^2 pixels. At this scale, joints collapse into a handful of pixels, increasing association errors and limiting the reliability of limb articulation cues. Fine-grained actions are particularly difficult to separate under such conditions.

Imbalanced data distributions: Critical safety-related actions such as falls, struggles, or fights are rare in available datasets. Models trained on such data are biased toward frequent activities such as standing or walking, limiting

generalization to rare but important events. Countermeasures including class rebalancing or curriculum learning can mitigate these effects but add complexity to training.

Overall, occlusion, viewpoint variation, low resolution, and class imbalance constitute core methodological challenges. Their impact is amplified by the lack of surveillance-oriented datasets with consistent annotation standards, which hinders systematic benchmarking and slows progress toward robust deployment.

1.2.3 Real-world Challenges

Operational deployment introduces a different class of difficulties that extend beyond algorithmic performance in controlled experiments. Once systems are embedded into real surveillance networks, they must remain reliable under evolving environments, continuous operation, and tight resource constraints. As illustrated in a single frame in Figure 1.5, tracking quality varies strongly within the same scene: foreground individuals are represented with stable trajectories, whereas persons in the background exhibit fragmented or inconsistent tracks due to scale variation and frequent occlusion.

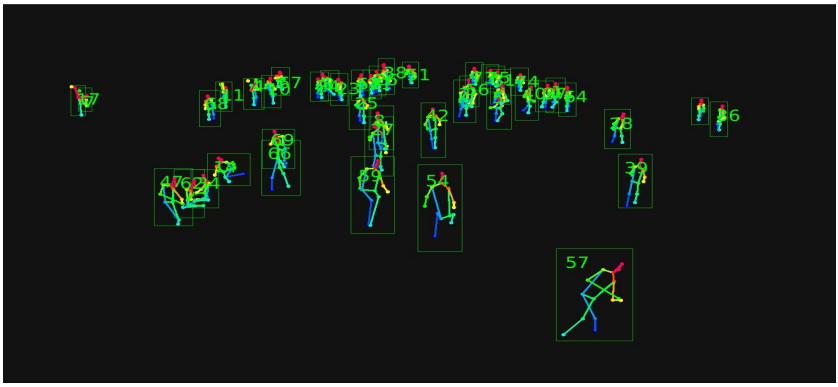


Figure 1.5: Anonymized results of a central station scene – Track results in the foreground are usually stable and poses are predicted with confidence. Background tracks appear more unstable due to smaller boxes and increased occlusion. Temporal consistency partially stabilized background tracks.

Generalization ability: Surveillance environments change over time. Seasonal shifts in clothing, lighting, and weather, as well as temporary factors such as construction work or public events, alter the visual appearance of scenes. In addition, sensors age, cameras are replaced, and layouts of public spaces evolve. These shifts introduce persistent distribution changes that degrade performance if models are not periodically recalibrated or adapted.

Processing pipeline: Unlike isolated benchmarks, real deployments require complete processing pipelines: video ingestion, frame-wise detection, multi-object tracking, pose estimation, sequence construction, and action recognition. Each stage is susceptible to error, and mistakes propagate through the pipeline. A single misaligned bounding box or identity switch can compromise subsequent skeleton extraction, leading to unreliable action predictions. Ensuring temporal consistency under real-time conditions demands careful synchronization, buffering, and error-handling strategies.

Hardware and real-time constraints: Surveillance networks typically consist of hundreds of cameras but operate under constrained budgets. This limits

available compute, memory, and power resources, often to modest GPU capacity or even CPU-only operation. To provide timely situational awareness, the entire pipeline must process streams in real time. These restrictions reduce the feasible input resolution and temporal horizon, and they discourage reliance on large ensembles or computationally expensive post-processing.

In summary, real-world deployment is not only a matter of algorithmic accuracy but also of robustness against long-term distribution shifts, stability of multi-stage pipelines, and efficient operation under resource limitations. Addressing these interdependent constraints is essential for enabling timely and trustworthy responses in safety-critical scenarios.

1.3 Contributions

The vast amounts of video data generated by modern surveillance networks are no longer analyzed effectively by human operators alone. Intelligent surveillance systems based on deep learning enable automated understanding of human behavior, transforming raw video streams into actionable information. This thesis develops a deep learning-based framework for skeleton-based action recognition in multi-camera surveillance networks under realistic conditions. The main contributions are summarized as follows:

- The Unified Pedestrian Attribute Recognition (UPAR) dataset and its pose-extended variant (UPAR-Pose) are introduced as large-scale resources for generalization studies. UPAR consolidates established Pedestrian Attribute Recognition (PAR) datasets, harmonizes annotations for 40 attributes, and contributes 3.3 million new binary labels [Spe23, Cor23, Cor24b]. UPAR-Pose adds 71,015 joint pose-and-attribute annotations [Cor26a], enabling systematic investigations of generalization in both single-task 2D-HPE and multitask 2D-HPE/PAR settings.

- Previous surveillance-oriented human pose estimation has relied predominantly on RGB data, neglecting robustness under adverse illumination and privacy-sensitive conditions [Jia21, Cho18]. To address this limitation, the LLVIP-P benchmark is contributed as an extension of Low-Light Visible-Infrared Paired (LLVIP), adding 26,135 2D pose-annotated person instances [Cor24c]. This benchmark enables rigorous evaluation across thermal–visible modalities.
- A unified thermal pose estimation benchmark (UPPET) is introduced by harmonizing pose annotations across four sources: LLVIP-P, OpenThermalPose (OTP) [Kuz24], CAMEL - Pose (CAMEL-P) [Geb18], and a new contributed industrial dataset (Thermisch-PE (TPE)). Based on the PoseTrack18 15-keypoint topology, UPPET resolves annotation inconsistencies and establishes multi-source protocols for cross-domain and cross-sensor evaluation. This benchmark addresses data scarcity relative to RGB datasets such as COCO [Lin14], domain shifts across environments [Nik21], and sensor heterogeneity, supporting realistic assessment for surveillance and worker-safety applications.
- A multitask 2D-HPE-PAR Model is proposed to provide pedestrian description a the same time with a pose. This model demonstrate that a multitask backbone successfully learns both HPE and soft biometric attributes, achieving competitive performance across tasks [Cor26a].
- An anatomy- and kinematics-aware framework for 3D-HPE is proposed that enforces bone-length and joint-limit constraints and exploits temporal structure to enhance stability and anatomical plausibility. A dual-fusion architecture jointly predicts root-relative 3D joints and per-joint orientations, reducing jitter and producing consistent sequences suitable for downstream tasks [Cor26b].

- Surveillance-oriented augmentation for SBAR is introduced to reflect degradations typically observed in 2D-HPE under operational conditions, including occlusion, truncation, missing keypoints, temporal jitter, scale and viewpoint variation, and identity fragmentation. Skeletal representations are systematically evaluated under these perturbations [Cor24a], resulting in design guidelines and improved robustness in adverse environments.
- A synthetic surveillance benchmark (GTA-RWS) is and its generation pipeline contributed, rendering naturalistic human motion in a game engine as a privacy-conscious alternative to sensitive real-world recordings. Modular subsets cover 2D-HPE, 3D-HPE lifting, and SBAR, with dense frame-level annotations including bounding boxes, 2D and 3D poses, action labels, and tracking identities. The dataset supports systematic evaluation of the full action recognition pipeline, enabling analysis of both individual modules and their interactions. Controlled degradations, such as occlusion, identity switches, and temporal jitter, can be studied to quantify error propagation and assess robustness. While the visual domain gap limits direct applicability for low-level detector training, GTA-RWS provides an ethical, scalable, and reproducible benchmark for surveillance-oriented research. To the best of the author's knowledge, it is the first dataset enabling end-to-end evaluation of the full pipeline and the influence of each component on overall system performance, with realistic movement derived from recordings of real persons.

1.4 Thesis Outline

This thesis is organized as follows. Related literature in 2D-HPE, 3D-HPE, and SBAR, together with closely related topics, is reviewed in Chapter 2. The overall problem formulation and system design adopted throughout the thesis are presented in Chapter 3. Datasets and evaluation protocols for each task, including UPAR, UPAR-Pose, and UPPET, are described in Chapter 4. Baseline methods and training setups for each task are detailed in Chapter 5.

Generalization improvements for 2D-HPE and a multi-task model that jointly performs PAR to improve resource efficiency are introduced in Chapter 6. A 2D-to-3D uplifting model is proposed in Chapter 7 to improve temporal kinematic stability and to estimate joint orientations. Data augmentation targeting real-world conditions to improve SBAR is developed in Chapter 8, and the impact of input sequence modality is analyzed.

The proposed techniques are evaluated against recent state-of-the-art methods in Chapter 9. A synthetic dataset is presented in Chapter 10, created by retargeting real pose sequences into a game engine under surveillance-like layouts, enabling system-level analyses (e.g., the influence of pipeline components on downstream modules). Finally, Chapter 11 concludes the thesis and outlines directions for future research.

A semi-automated annotation tool developed by the author and student assistants is documented in the Appendix. This tool enabled the creation of UPAR, UPAR-Pose, UPPET, and additional datasets.

A NOTE ON IMPLEMENTATION Mickael Cormier is responsible for conception and implementation of the overall framework as presented in this thesis. Parts of this dissertation are based on joint works resulting from very close collaboration with his students Caleb Ng Zhi Yi (Section 4.1.1.3 and Section 4.1.1.3), Jeremy Zolk (Chapter 7 and Section 9.2), Yannik Schmid (Section 8.1) and David Anderlohr (Section 10.1 and Section 10.1.2). Both, Mickael Cormier and the corresponding student, have contributed substantially to this research. While it is difficult to set a precise boundary, Mickael Cormier was rather in charge of the idea whereas the student focused on the implementation.

2 Related Work

This thesis aims at researching SBAR of persons using deep learning techniques. This chapter presents the relevant literature related to this topic and the proposed framework. First, Section 2.1 provides background information on HPE. Next, the literature on SBAR is presented in Section 2.2, where different approaches to the task are described.

2.1 Human Pose Estimation

HPE encompasses the localization of anatomical keypoints in 2D image space and their reconstruction in 3D space. In 2D-HPE, methods infer image-plane keypoints and assemble them into skeletons. In 3D-HPE, methods estimate body-centric or camera-centric joint coordinates, and optionally orientations, from single images or video. For surveillance use, robustness to occlusion, scale, and viewpoint variation, low illumination, and sensor heterogeneity is essential. This section reviews 2D-HPE paradigms (bottom-up, top-down, one-stage) with emphasis on deployment trade-offs, followed by 3D-HPE approaches that either predict 3D poses directly from RGB (single-stage) or uplift 2D keypoints to 3D with temporal models.

2.1.1 2D Human Pose Estimation

2D-HPE addresses the localization of anatomical human landmarks (keypoints) in images or video and the assembly of these landmarks into a skeletal topology. Methods for 2D-HPE follow three principal learning

modalities. Regression-based techniques predict continuous Cartesian coordinates directly for each keypoint [Tos14, Wan24a]. Heatmap-based techniques predict dense likelihood maps for each keypoint type, from which coordinates are derived by a peak-finding step [Xia18]. Classification-based techniques discretize the coordinate space and reformulate localization as a classification problem [Li22b]. Recent instances of this paradigm demonstrate favorable quantization properties and robust handling of uncertainty (*e.g.*, SimCC [Li22b]). These three modalities establish a conceptual bridge to the architectural taxonomies discussed below: heatmap decoders and their discretization biases influence both top-down and one-stage systems, classification-based representations integrate naturally with one-stage and transformer-based set-prediction formulations, and regression strategies underlie several efficient real-time designs.

The remainder of this section reviews dominant 2D-HPE paradigms: bottom-up, top-down, one-stage, and transformer-based methods. It also discusses temporal multi-frame techniques for stabilizing pose predictions, and highlights dataset limitations, including sparse outdoor surveillance and thermal imagery. Algorithmic trade-offs relevant for surveillance—robustness under occlusion, generalization across viewpoints and sensors, computational scalability, and transparency for auditability, are emphasized throughout.

Bottom-up methods: Bottom-up pipelines first detect keypoints of all types across the image and subsequently associate detected keypoints into human instances. Representative works include associative models that employ pairwise affinity fields to link adjacent joints and offset-field formulations that predict displacement vectors between parts and their parent joints [Cao17, New17, Pis16, Pap18]. Part affinity and fine-grained association fields increase robustness to local ambiguities in pair assignments [Cao17, Kre19].

Bottom-up designs exhibit near-constant inference time with respect to the number of persons present and therefore present an attractive computational profile for dense crowds. Practical limitations arise from scale variation and grouping ambiguity: distant subjects produce small keypoint footprints that require multi-resolution features for reliable detection, and severe occlusion increases false associations. Architectural refinements target these failure

modes through multi-resolution feature extraction and adaptive local filters [Che20, Gen21].

Top-down methods: Top-down pipelines decompose the task into person detection followed by single-person pose estimation within cropped regions. This decomposition simplifies pose modeling and yields high localization accuracy in many benchmarks. Strong single-person estimators employ high-resolution backbones and careful heatmap decoding strategies: multi-scale, parallel-resolution networks preserve spatial details [Sun19], while heatmap-postprocessing methods reduce quantization bias and produce sub-pixel corrections [Zha20]. In particular, distribution-aware decoding techniques such as DARK [Zha20] and unbiased decoding procedures [Hua20] substantially reduce discretization errors introduced by heatmap resolution and coordinate transformations, thereby improving reported localization metrics.

Top-down designs retain sensitivity to bounding-box quality and scale linearly in runtime with subject count. Recent extensions address these issues by permitting multiple-instance prediction within a single bounding box and by explicit occlusion modeling [Khi21, Gol19, Sun24, Pur25]. When accurate detectors and batch inference are available, top-down pipelines deliver a practical balance between accuracy and throughput for many deployment scenarios.

One-stage and end-to-end methods: One-stage methods unify detection and pose estimation into a single network that predicts multi-person poses directly from the image, without intermediate proposal or grouping stages. End-to-end set-prediction formulations based on transformers represent a prominent instantiation of this approach: PETR introduced a transformer-based, set-prediction pipeline with keypoint and positional queries that replaces non-maximum suppression, region proposals, and grouping [Shi22]. GroupPose refined this line of work by integrating a lightweight DETR-style decoder with grouped self-attention, yielding a simple and strong end-to-end baseline [Liu23a]. EDPose and its extension EDPose++ further improve end-to-end multi-person pose estimation by incorporating enhanced query interaction and progressive feature refinement, demonstrating competitive accuracy while reducing runtime [Yan23b, Yan25]. The approach presented

in [McN22] models keypoints and poses as objects, allowing object-detection techniques to handle multi-person pose estimation efficiently and improving robustness to crowded or occluded scenes.

Real-time oriented one-stage methods adopt compact detection-inspired backbones and hybrid coordinate-classification schemes to reduce latency. RTMO (a YOLO-style coordinate-classification plus regression approach) targets high-performance real-time multi-person pose estimation and demonstrates favorable throughput/accuracy trade-offs on standard benchmarks [Lu24]. Diffusion-based regression models recast keypoint regression as an iterative sampling and refinement process. DiffusionRegPose improves robustness under heavy occlusion by producing multi-modal pose hypotheses that undergo refinement [Tan24a]. Tokenized transformer models with self-distillation training regimes, such as SDPose, compress transformer representations while preserving pose accuracy [Che24].

One-stage and end-to-end architectures reduce pipeline complexity and eliminate error propagation from intermediate modules. Performance under extreme occlusion and dense crowds continues to lag the best top-down transformer systems in benchmark comparisons [Pur25].

Transformer-based methods: Transformer models have altered the landscape of 2D-HPE by providing flexible mechanisms to model global context and structural relations among joints. The evolution proceeds from hybrid designs toward pure transformer instantiations and thereafter toward more efficient tokenized representations.

Early hybrid efforts integrated transformer blocks with convolutional backbones to enrich global reasoning while retaining local feature extraction: TransPose introduced attention-based modules that augment convolutional features with long-range dependencies [Yan21b]. Subsequently, token-centric models such as TokenPose represented each keypoint as an explicit token and applied multi-head self-attention to model inter-joint relations [Li21c]. ViTPose demonstrated that a pure Vision Transformer backbone, when pre-trained on large image corpora and adapted to the pose task,

yields highly competitive performance [Xu22]. HRFormer generalized multi-resolution parallelism to transformer blocks, harmonizing high-resolution feature preservation with global attention [YUA21]. Attention-based grouping methods such as CenterGroup replaced explicit grouping heuristics with learned attention that assigns keypoints to person centers [Bra21].

Recent research introduced probabilistic modeling in top-down pipelines. ProbPose [Pur25] predicts pose distributions rather than singular estimates, enabling uncertainty-aware reasoning. Large-scale transformer models such as Sapiens generalize robustly to in-the-wild conditions through self-supervised pretraining on massive human image datasets [Khi24a]. Multi-frame architectures such as Poseidon extend transformers with temporal context, adaptive frame weighing, multi-scale fusion, and cross-attention modules to improve video pose estimation [Pac25]. These developments demonstrate the potential of top-down transformer pipelines in surveillance, particularly when temporal coherence or generalization to unseen contexts is required.

Transformer architectures provide superior capacity to model structural constraints and long-range dependencies, which benefits scenes with complex pose interactions. Resource requirements remain substantial for large transformer backbones. Current work therefore emphasizes efficient tokenization, distillation, and grouped attention to reduce compute while preserving accuracy.

Classification-based localization: Classification-based coordinate discretization reframes keypoint localization as a classification task over discrete coordinate bins. SimCC introduced a simple and effective strategy for discretized coordinate representation that alleviates quantization errors and stabilizes training across resolutions [Li22b]. Classification representations integrate well with transformer decoders and one-stage predictors because they avoid sub-pixel decoding heuristics and provide probabilistic outputs that support uncertainty estimation.

Temporal stabilization and efficient video inference: Recent research in 2D-HPE has progressed from single-frame prediction toward leveraging temporal cues from video sequences. Bottom-up methods such as PifPaf [Kre19] detect keypoints independently and associate them to individuals, providing inference time independent of crowd size. OpenPifPaf [Kre21] extends this framework to multi-frame inputs by introducing Temporal Composite Association Fields, enabling spatio-temporal keypoint association and improved stability across frames. Additional approaches, including Deep Dual Consecutive Network [Liu21], Temporal Feature Alignment and Mutual Information Maximization [Liu22b], TEMPO [Cho23], Kinematic-Aware Hierarchical Attention Network [Jin23], SmoothNet [Zen22b], and Deciwatch [Zen22a], exploit temporal context using feature alignment, hierarchical attention, and kinematic cues such as velocity and acceleration to refine pose predictions in crowded or occluded scenes.

Despite these advances, the proposed surveillance-oriented pipeline retains single-frame top-down transformer models as the core estimator. Single-frame top-down architectures provide modular interpretability and explicit keypoint-level confidence measures. This enables transparent, auditable configurations suitable for sensitive deployment contexts. Temporal refinement modules remain optional and can be integrated downstream, preserving computational efficiency, real-time responsiveness, and state-of-the-art localization performance.

An important consideration in practical deployments is the trade-off between accuracy and runtime. Single-frame estimators excel in terms of speed, which is critical for high-density scenes with multiple subjects, but their top-down design can become a bottleneck when processing very crowded frames, as each detected bounding box requires independent forward passes. Furthermore, variations in subject height directly affect the size of bounding boxes and, consequently, the spatial resolution available for keypoint estimation. Shorter subjects or partially occluded individuals may yield smaller crops, increasing localization errors and reducing recognition accuracy. Balancing the frame rate, bounding-box scaling, and optional temporal refinement is

therefore crucial to maintain both high throughput and robust performance in crowded, real-world surveillance scenarios.

Datasets: Large-scale RGB benchmarks such as COCO [Lin14] and MPII [And14] drove architectural progress but remain biased toward eye-level viewpoints and moderate crowd densities. Specialized datasets address occlusion (CrowdPose [Li19b], OCHuman [Zha19a]) and video sequences (Posetrack [And18, Dör22]), yet surveillance scenarios with steep overhead viewpoints, dense crowds, low resolution, and adverse illumination remain underrepresented [Lin23]. Thermal imagery datasets exist in academic literature (OTP [Kuz24], but the corpus of annotated outdoor surveillance-style thermal data remains limited and lacks standardization across keypoint topologies and cross-sensor evaluation protocols.

This thesis addresses these deficits by proposing benchmarks and protocols tailored to outdoor surveillance and thermal modalities, with large-scale harmonized annotations supporting specialization and generalization studies.

Summary: Architectural choices in 2D-HPE involve trade-offs among accuracy, scalability, and transparency. Bottom-up networks scale gracefully with scene density while offering limited interpretable outputs. One-stage models simplify the pipeline at the expense of transparency, as intermediate confidence estimates remain inaccessible. Top-down transformer-based estimators deliver state-of-the-art accuracy while preserving explicit person-centric bounding boxes and joint confidence values. This modular structure facilitates component-level evaluation, logging, and immediate adjustment under sensitive operational protocols. Runtime configuration of detection thresholds and keypoint acceptance rules supports legally and ethically accountable deployment in public safety applications. The datasets and protocols introduced in this thesis create the infrastructure to rigorously evaluate transformer-based systems across modalities, camera types, viewpoints, and environmental conditions, enabling systematic progress in privacy-aware surveillance applications.

2.1.2 3D Human Pose Estimation

3D-HPE addresses the recovery of skeletal configurations from images, videos, or other sensing modalities. Historically, prior to deep learning, methods relied on generative or template-based strategies. Brauer [Bra14] employed the Implicit Shape Model (ISM) to infer 3D joint locations from image evidence. The approach encoded a probabilistic mapping between local appearance descriptors and 3D keypoint positions, aggregating votes across spatially detected features to hypothesize full-body poses. This framework established early principles of multi-hypothesis fusion and structured reasoning that continue to influence contemporary pipelines.

Recent 3D-HPE methods fall roughly into two families. The first focuses on parametric mesh reconstruction, often inspired by Michael Black's SMPL model and its extensions. SMPL and subsequent works [Lop23, Raj22, Pat25, Goz25, Dwi24, Kel23, Xia25] represent human bodies as deformable meshes controlled by low-dimensional pose and shape parameters. These methods enable accurate, anatomically plausible reconstructions and support downstream applications in biomechanics, motion analysis, and avatar synthesis. They are often optimized to produce realistic surface geometry, enforce joint limits, and maintain bone-length consistency.

The second family targets skeleton-based 3D poses, typically expressed as joint coordinates in camera or world space. Approaches vary according to whether poses are inferred directly from monocular images or uplifted from 2D keypoints. Direct regression methods [Zho18, Li20] predict 3D joint locations end-to-end from RGB inputs, often employing temporal convolutions or transformer encoders for video sequences. Multi-person extensions incorporate volumetric heatmaps or detection-based strategies [Fab18, Jia24] to handle crowded scenes in monocular settings. Direct 3D regression from images provides end-to-end predictions but often struggles with depth ambiguities and generalization across viewpoints. An alternative strategy leverages the high reliability of 2D-HPE models: predicted 2D keypoints serve as an intermediate representation, which is subsequently lifted into 3D. This decoupling allows 2D detectors to exploit large-scale datasets and mature architectures,

while lightweight temporal models focus on resolving depth and maintaining temporal consistency.

2D-to-3D Uplifting: Temporal models exploit sequences of 2D skeletons, using recurrent networks, temporal convolutions, or transformer-based architectures [Mar17, Pav19b, Zhe21, Zha22, Zha23b, Zha23a, Cho23, Pen24, Li24]. Early methods employed Long Short-Term Memory (LSTM) networks [Hos18], while Graph Neural Networks (GNNs) modeled joint dependencies explicitly [Ci19, Cai19, Zha19b, Liu20a]. Temporal convolutional networks demonstrated robust performance for video sequences, producing smooth 3D pose trajectories [Mar17, Pav19b, Liu20b, Che21a].

Transformer architectures marked a significant advancement. PoseFormer [Zhe21] introduced the first convolution-free transformer for 3D pose uplifting, establishing a competitive baseline for sequence modeling. PoseFormer V2 [Zha23b] and ContextPoseFormer [Zha23a] extended this line, integrating spatial-temporal attention and multi-scale reasoning. MotionBERT [Zhu23] unified motion representation learning across frames, while MixSTE [Zha22] employed a sequence-to-sequence spatio-temporal encoder to improve temporal coherence. Multi-hypothesis models [Li23] addressed inherent depth ambiguity by producing multiple plausible 3D poses per 2D input, and TCPFormer [Liu25b] introduced trajectory-consistent priors for temporally stable predictions.

Despite strong performance, many methods neglect structural constraints critical for anatomical plausibility. Bone-length regularization, orientation-based representations, and kinematics-aware losses reduce artifacts such as implausible joint rotations or unrealistic limb proportions [Che21a, Lud25a, Pen24]. Hsu et al. [Hsu24] enforced anatomically plausible bone lengths while preserving orientations, and Chen et al. [Che21a] predicted bone direction and length separately, refining poses with a joint shift loss. Temporal smoothness has been further encouraged using velocity and acceleration constraints [Jin23, Pen24].

These developments highlight a trend toward transformer-based 2D-to-3D uplifting models that provide fast, accurate, and temporally stable 3D predictions. Such models enable efficient integration into surveillance pipelines, where sequences of 2D keypoints are rapidly converted to 3D skeletons for downstream SBAR while preserving anatomical realism and temporal coherence.

Multi-Person 3D Pose Estimation: Estimating 3D human pose for multiple individuals introduces additional challenges, including inter-person occlusions, depth ambiguities, and increased computational complexity. Early solutions employed volumetric representations, discretizing 3D space into voxels and predicting per-voxel occupancy or heatmaps for each joint. Fabbri et al. [Fab20] proposed compressed volumetric heatmaps for multi-person 3D pose estimation using the JTA dataset [Fab18], demonstrating that voxel-based representations capture spatial relationships between joints across people while maintaining differentiability for end-to-end learning. Such approaches, however, incur high memory and computational costs, limiting real-time applicability.

Subsequent methods leveraged monocular image sequences and parametric modeling to balance efficiency and accuracy. WorldPose [Jia24] introduced a baseline for multi-person 3D pose estimation in world coordinates, combining detection, per-person keypoint regression, and temporal consistency across frames. While world-coordinate representations are valuable for certain applications, such as multi-camera activity recognition or scene-level reasoning, they rely on accurate camera calibration and absolute depth estimation. In practical real-world surveillance, these conditions are often unavailable or unreliable. Per-person human-centric, root-relative poses provide a robust and deployable alternative. Root-relative representations describe joint configurations relative to the pelvis or torso, offering sufficient information for action recognition and behavior analysis while avoiding the need for precise localization in world space.

Datasets such as JTA provide high-quality multi-person poses in standard scenarios, and WorldPose include more dynamic motions such as kicking or tackling. Nevertheless, both remain limited for typical non-sport surveillance,

where highly dynamic interactions occur, such as falls, fights, or aggressive gestures. The lack of annotated 3D data for these safety- or security-relevant activities motivates the contributions of this thesis, which propose a dataset tailored to those conditions. Without such data, assessing 3D reconstruction accuracy in these challenging contexts remains largely infeasible.

Together, these works highlight the trade-offs between volumetric expressiveness, computational efficiency, and applicability to real-world pipelines. Transformer-based 2D-to-3D uplifting approaches complement these methods by providing fast, human-centric 3D skeletons suitable for downstream tasks such as skeleton-based action recognition, even in monocular, dynamic, and crowded scenes.

Summary: Overall, 3D-HPE has matured from early template-based reasoning to deep learning pipelines capable of accurate single- and multi-person 3D reconstruction. Parametric mesh models ensure anatomical plausibility, while skeleton-based methods —direct regression or 2D-to-3D uplifting— provide modularity and computational efficiency. Human-centric representations, rooted in a per-person local coordinate system, align naturally with real-world surveillance pipelines: they remain robust under monocular input, crowded scenes, and uncalibrated cameras, and supply sufficient information for SBAR without requiring world-coordinate ground truth. Multi-person extensions, transformer-based temporal modeling, and alternative sensing modalities further enable reconstruction in challenging monocular video streams. Transformer-based 2D-to-3D uplifting serves as the core strategy for this thesis, converting sequences of 2D keypoints into accurate 3D skeletons for downstream SBAR while preserving anatomical plausibility and temporal coherence.

Importantly, this thesis contributes by explicitly enforcing anatomical constraints and predicting joint orientations, enhancing the realism of reconstructed skeletons and providing richer input modalities that improve the reliability of downstream action recognition tasks.

2.2 Skeleton-Based Action Recognition

SBAR has become a pivotal approach in human action analysis, leveraging the structured representation of human joints and bones over time. Skeleton-based methods abstract human motion into keypoints, offering intrinsic advantages for privacy-sensitive applications such as surveillance, since identifiable visual information is removed while retaining motion dynamics. Sequences of keypoints obtained from HPE algorithms encode temporal movement patterns while remaining robust to environmental factors including illumination changes, occlusions, and background clutter. The evolution of SBAR has been driven by advances in spatial-temporal modeling, representation learning, and computational efficiency. However, real-world surveillance introduces additional challenges that are often not reflected in benchmark datasets, such as noisy 2D or pseudo-3D keypoints, occlusions, tracking inconsistencies, and variable camera viewpoints. Addressing these factors is essential for deploying robust skeleton-based recognition in unconstrained environments.

CNN-based Skeleton Feature Extraction: Convolutional Neural Network (CNN) have been employed to model skeleton sequences by treating joint coordinates as structured feature maps or tensors over space and time. DDNet is a notable example, using a double-feature double-motion architecture to efficiently encode both joint positions and temporal dynamics [Yan19]. Similarly, PoseC3D introduces a 3D heatmap-based representation of skeleton sequences, allowing the network to capture spatiotemporal features while mitigating the impact of HPE errors [Dua22b]. Both approaches achieve high computational efficiency, with PoseC3D demonstrating robustness across multiple datasets and the ability to process multi-person sequences with minimal additional overhead. These CNN-based encoders provide practical alternatives for real-world surveillance systems where latency and hardware constraints are critical, complementing more complex Graph Convolutional Neural Network (GCN)- and transformer-based models. However, their results remain behind those of GCNs.

GCNs and Topology-Aware Modeling: GCNs have become the de facto standard for modeling the structured dependencies of the human skeleton, where joints serve as nodes and bones as edges. Early spatio-temporal GCNs demonstrated that separate modeling of spatial and temporal dependencies improves recognition performance. CTR-GCN introduced channel-temporal refinement, weighting salient joints and temporal segments to enhance robustness under occlusion and variable viewpoints [Che21b]. HDGCN decomposes the skeleton graph hierarchically, learning both local joint interactions and global skeletal structure through attention-guided aggregation [Lee23]. Hyperformer extends GCNs via a hypergraph self-attention mechanism, capturing higher-order kinematic dependencies and bone connectivity beyond pairwise interactions [Zho22b]. While most GCN-based methods rely on 3D joint coordinates, 2D or pseudo-3D representations such as $(x,y,1)$ can be used effectively, addressing practical challenges in surveillance where accurate world-coordinate 3D reconstruction is often infeasible.

Transformer-Based Temporal Modeling: Transformers provide a flexible framework for capturing long-range temporal dependencies in skeleton sequences without fixed graph topology constraints. SkateFormer applies a skeletal-temporal partitioning strategy, enabling focused attention on critical joints and frames while maintaining computational efficiency [Do24]. HybridFormer follows a two-stage local-to-global design: lightweight local blocks capture spatial and temporal neighborhoods, and global attention blocks subsequently integrate spatial and temporal dimensions, yielding strong performance while preserving efficiency [Zho25]. MotionBERT introduces a unified pretraining paradigm in which a Dual-stream Spatio-Temporal Transformer (DSTformer) is trained to reconstruct complete 3D motion from noisy or partial 2D skeleton sequences. This 2D→3D reconstruction task equips the model with geometric and kinematic priors that transfer effectively to downstream tasks, including action recognition, 3D-HPE, and mesh recovery, with minimal fine-tuning [Zhu23]. Attention-guided contrastive learning, as in MaskCLR, further strengthens resilience to noisy or missing keypoints [Abd24]. Transformer-based architectures can operate on 2D, pseudo-3D, or 3D inputs, making them suitable for diverse surveillance conditions, though their computational demands remain higher than CNN-

or GCN-based methods, which motivates research into hybrid architectures balancing accuracy and efficiency.

Group-Based and Multi-Person Action Recognition: Extending SBAR to multi-person scenarios requires modeling inter-person interactions and relational dynamics, essential for accurate group activity recognition. Graph-based approaches such as spatial-temporal panoramic graphs capture relational dependencies among multiple individuals and objects within the same framework, enabling joint modeling of intra-person, inter-person, and person-object interactions via graph convolutions [Li25]. Hypergraph-based methods extend this capability by modeling higher-order relationships among joints and actors, yielding richer contextual representations for collective behaviors in a unified structure [Wan24b]. COMPOSER implements a multiscale Transformer-based architecture, reasoning compositionally over keypoint-only representations to detect group activities while avoiding scene bias and privacy concerns. It clusters representations across scales and applies attention over these clusters to interpret group interactions [Zho22a]. The Bi-Causal Group Activity Recognition framework employs bidirectional temporal causality to capture influence patterns among group members, improving the modeling of collective activity progression across time [Zha24]. Although these approaches achieve strong performance on benchmark datasets such as Volleyball [Ibr16] and Collective Activity [Cho09], their practical deployment in surveillance systems is hindered by challenges including occlusions, identity tracking errors, and increased computational complexity in dynamic and crowded environments. Consequently, single-person action recognition models that process 2D or pseudo-3D skeletons efficiently remain the most viable option for many real-world surveillance applications. Nevertheless, group-based methods provide a conceptual foundation for future systems aiming to interpret complex social interactions in unconstrained, multi-agent scenes.

Efficient and Real-World Systems: Practical surveillance applications impose strict constraints on computational resources and latency, necessitating efficient skeleton-based action recognition frameworks. PYSKL provides a modular toolbox that facilitates development, benchmarking, and

reproducible evaluation of skeleton-based models. It implements multiple architectures and supports a wide set of skeleton benchmarks and pre-trained models, enabling fair comparison of methods and good-practice defaults [Dua22a]. UNIK targets real-world deployment by learning transferable spatio-temporal representations and an optimal dependency matrix. The work demonstrates improved cross-dataset generalization after pretraining on a large pose-derived corpus (Posetics) which was introduced in the same work, and reports transfer gains on several downstream benchmarks [Yan21a]. SkeletonMAE proposes a graph-based masked autoencoder pretraining scheme for skeleton sequences that reconstructs masked joints and topological information. Pretraining with this objective yields encoders that generalize across datasets and are more robust to incomplete or noisy keypoint input [Yan23a]. Complementing these approaches, the Unified Keypoint-based Action Recognition Framework via Structured Keypoint Pooling introduces Structured Keypoint Pooling, a cascaded sparse aggregation over keypoints that treats time-series keypoints as point-cloud inputs and improves robustness to detection and tracking errors through its pooling-switching training trick. The method attains state-of-the-art recognition accuracy while running at very high inference rates on modern GPUs [Hac23]. Collectively, these works exemplify strategies for high accuracy and operational efficiency under imperfect 2D-HPE, detection failures, and tracking inconsistencies. A critical gap remains in systematically quantifying how upstream pipeline components influence downstream recognition, which this thesis addresses.

Summary CTR-GCN remains a strong baseline for single-person SBAR in surveillance applications, providing hierarchical graph modeling and channel-temporal refinement that emphasizes salient joints and temporal segments [Che21b]. Its computational efficiency, robustness to noisy 2D and pseudo-3D inputs, and compatibility with real-world surveillance conditions establish it as a practical reference for evaluating novel methods. More advanced graph-based approaches, such as HDGCN, Hyperformer, and HybridFormer, refine structural modeling through hierarchical decomposition, hypergraph attention, or hybrid GCN–transformer integration, thereby improving the representation of complex spatial and temporal dependencies

[Lee23, Zho22b, Zho25]. In parallel, pretraining-based frameworks such as MotionBERT and SkeletonMAE learn generalized motion representations across datasets and tasks, enabling transfer to action recognition and related problems while enhancing robustness to noisy or incomplete input [Zhu23, Yan23a]. Although these advanced models achieve state-of-the-art results in controlled benchmarks, their complexity and computational cost hinder deployment in real-world surveillance scenarios. By systematically examining the effects of upstream pipeline errors and input modalities, this thesis builds on CTR-GCN to bridge the gap between research-oriented performance and deployment-ready surveillance systems.

3 Concept

The thesis focuses on the development of a deep learning–based framework for skeleton-based action recognition in real-world live video within large-scale multi-camera networks. The objective is the provision of a robust system for smart city, surveillance, and safety applications. The methodology comprises modular components with an explicit balance between computational efficiency and privacy protection.

Multiple approaches for action recognition are introduced in Section 2.2. In this work, a modular pipeline is adopted. First, 2D Keypoints are estimated from video through 2D-HPE. In parallel, soft biometric attributes are obtained via PAR in parallel from the same module. Then 3D Joints are subsequently approximated for SBAR for the following reasons:

- **Explainability and interpretability:** Structured, low-dimensional representations (2D Keypoints and 3D Joints) are produced through 2D-HPE, 3D-HPE and SBAR, with intermediate results and decision criteria open to inspection and auditing. Semantic attributes in a human-interpretable form are obtained through PAR.
- **Flexibility:** Operation is supported across sensing modalities (e.g., optical and thermal) and across deployment conditions in heterogeneous multi-camera environments.
- **Deployment and maintainability:** Clear interfaces are specified for the modules. Replacement of components with minimal impact on the remaining system is enabled by abstraction boundaries.
- **Responsiveness:** Low-latency processing is targeted to support live monitoring under soft real-time constraints, with timely notification.

- **Privacy motivation:** While person-related data are not fully avoidable, the use of skeletal representations and soft attributes reduces exposure to identity-bearing imagery and supports bias-aware choices in data processing and model selection.
- **Complementarity:** Soft biometric attributes provide complementary information for skeleton-based action recognition, event confirmation, and incident reporting.

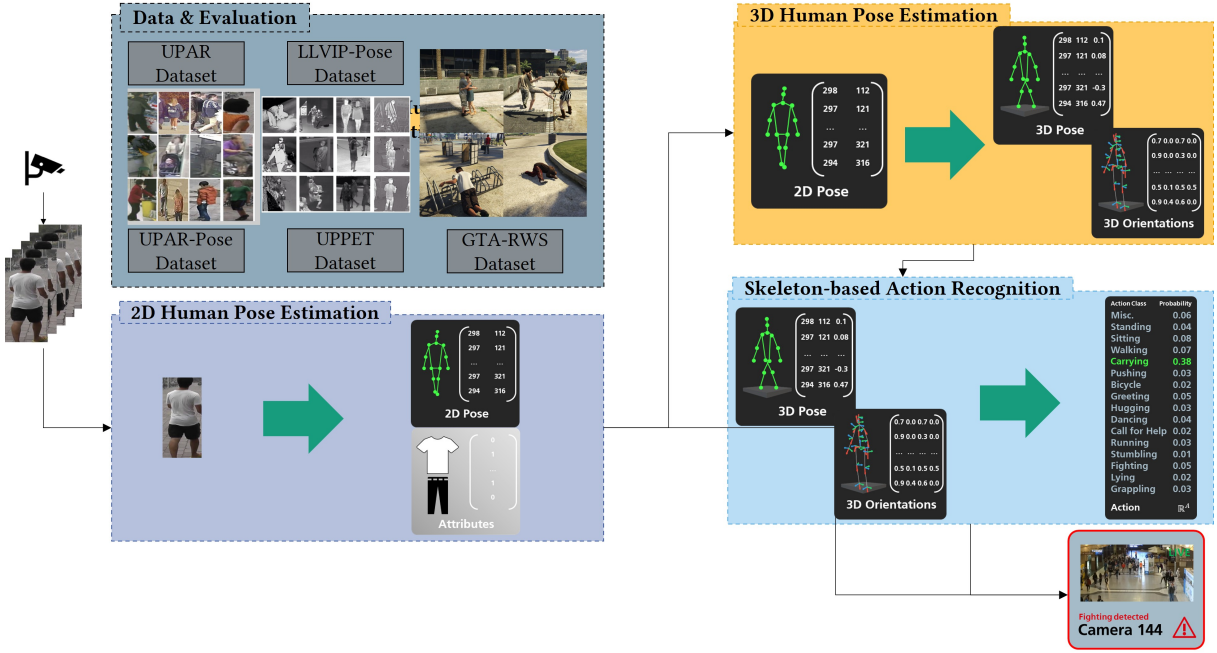


Figure 3.1: Concept Overview – This thesis explores a framework for skeleton-based action recognition using short image crops sequences of a tracked person. The tracks are provided by other works. In this thesis, a 2D-HPE module extracts keypoints and, via an auxiliary head on the same model, person-level semantic attributes. The 2D keypoint sequence is then lifted to 3D joints by a 3D human pose estimation module, with a parallel branch producing 3D joint orientations. Both 3D joint positions and orientations are fed into a skeleton-based action recognition module. Finally, the detected action and its description are combined and sent as a notification when a salient action is observed.

In Figure 3.1, the complete processing framework is shown. It lays out the stages from camera streams or videos to notification generation and situates the methods studied in this thesis. Building on this framework, the long-term vision is robust operation without continuous human monitoring. The goal is a black-screen mode. Only the spatiotemporal segments referenced by a notification are rendered clearly, along with a textual description of the suspect or other relevant actors. Because computation and notification are performed without explicitly using biometric identifiers, and since communication between different systems is often difficult, attribute-based suspect descriptions are more actionable across heterogeneous deployments.

First, tracks and detections provided by an upstream detector–tracker are used as inputs. Although works co-authored by the author of this thesis address detection [Cor21d, Cor21c] and tracking [Spe22a, Spe22b], the scope of this thesis is limited to downstream processing, and detections and tracks are therefore assumed to be available from existing systems (*e.g.*, [Spe24, Sta24]). For each track, person crops are extracted per frame. The HPE module is applied to estimate 2D Keypoints for each frame of the track. In parallel, PAR is applied to the same track crops to extract soft biometric attributes together with confidence measures. The attributes provide a compact, human-interpretable description of the depicted individual and are used during event confirmation and notification.

The single-view lifting module maps sequences of 2D Keypoints to temporally coherent 3D Joint trajectories. A lifting-based formulation is adopted, rather than direct RGB-to-3D regression, to decouple appearance from geometry, leverage robust 2D keypoint detectors trained at scale, and enable calibration-light deployment across cameras. Sequence-level inference imposes temporal and kinematic consistency, improving stability and occlusion recovery. The resulting skeleton tracks contain joint positions in a root-joint-centric coordinate representation, alongside per-joint 3D orientations.

The SBAR module is then applied to the 3D Joint sequences. Graph-based models process joint positions and velocities to produce action posterior scores at the frame or segment level. Notification generation follows. Action scores and spatiotemporal context from tracking (*e.g.*, zones of interest and

dwell times) are fused to trigger a notification when predefined conditions are met. The notification includes timestamps and camera identifiers, minimal visual snippets limited to the referenced segments, and a textual attribute-based description of the relevant actors. No biometric identifiers are included. The user interface remains a black screen until a notification is raised.

Finally, storage, auditing, and inter-system communication are addressed. The system stores 2D Keypoints, 3D Joints, action scores, and attributes, while original imagery is discarded or retained only ephemerally for a few seconds. Because notifications and suspect descriptions rely on attributes rather than biometrics, interoperability with heterogeneous external systems is improved, as textual attribute descriptions are directly actionable across organizational and legal boundaries.

The primary focus of this thesis is on the 2D-HPE, 3D-HPE, and SBAR stages, as they represent the core components of the framework. Nevertheless, research in terms of combining the components in a synthetic environment almost free of biometrics, with realistic motions, is discussed in Chapter 10 to provide necessary context for the application in real-world scenarios. The remainder of this chapter elaborates on the fundamental concepts and principles behind the contributions made to each specific stage in the order of appearance in this thesis.

Evaluation: To develop and evaluate data-driven methods for a specific use case, it is crucial to identify meaningful data and evaluation metrics for the task. When aiming to apply models in a real-world scenario, it is important to learn features that are not overfitted to specific scenes, person appearances, or dataset biases. Measuring generalization performance is only possible if training and test domains originate from different sources. Nevertheless, none of the existing research datasets offer such a scenario [Lin14, And18, Dör22, Zha19a, Li19b, Lin23, Fab21]. Most datasets lack diversity with respect to, e.g., demography, lighting conditions, and clothing variability, which leads to models with poor generalization abilities. Cross-dataset evaluation is difficult due to different skeletal topologies of the human pose annotations in the datasets. To close this research gap, the UPAR dataset [Spe23, Cor23,

Cor24b], followed by UPAR-Pose datasets, are proposed, which harmonize existing datasets for both PAR and 2D-HPE tasks. Furthermore, the UPPET dataset [Cor25] is proposed with annotations for multiple datasets in the thermal spectrum. This allows a large-scale investigation of the generalization performance of PAR and 2D-HPE methods. Furthermore, the LLVIP-P [Cor24c] dataset is proposed to compare 2D-HPE methods with aligned night images in the visible and thermal spectra.

2D Human Pose Estimation: The 2D-HPE model serves as the main feature extractor in the proposed system, producing 2D Keypoints that are refined to 3D and subsequently used for action recognition. High-quality 2D-HPE output is therefore crucial for recognition performance. A top-down approach is adopted due to favorable accuracy, despite drawbacks in inference time. Preliminary work by the author has shown that these latency issues can be addressed efficiently [Cor21d, Cor21c, Cor22a]. Although recent research has proposed heavy CNN- and transformer-based models, performance in real-world surveillance and downstream applications remains underexplored [Cor22b]. Consequently, this thesis systematically evaluates robustness and generalizability in surveillance scenarios across visual and thermal imaging [Cor24c, Cor25]. Building on these findings, a strong single-task 2D-HPE baseline is proposed. A competitive multi-task model is further introduced that jointly performs 2D-HPE and PAR, enabling deployment of a single model instead of two and reducing runtime resource usage.

3D Human Pose Estimation: The 3D-HPE component serves as the lifting stage in the proposed system, mapping sequences of 2D Keypoints to root-relative 3D Joints that subsequently drive action recognition. The intuition is that 3D skeleton sequences are less ambiguous than 2D. Accordingly, the component plays a critical role in delivering temporally consistent and physically plausible motion representations. Single-view, root-relative lifting is used for practicality in unconstrained environments—where calibration parameters or depth measurements are unavailable—and for robustness across heterogeneous camera setups. Recent work has emphasized multi-view or

calibration-dependent approaches, with limited deployment-oriented evaluation in surveillance and downstream applications. A strong single-task 3D-HPE baseline guided by kinematic constraints is proposed. A companion multi-output model is introduced that predicts both 3D Joints and 3D Joint orientations. A constraint-guided prediction framework enforces anatomical and kinematic plausibility. Evaluation targets robustness and generalizability, with emphasis on temporal coherence and physical plausibility. Ablations cover constraint terms and orientation outputs, and latency–accuracy trade-offs are measured to maintain soft real-time operation within the pipeline.

Skeleton Based Action Recognition: The SBAR module constitutes the terminal stage of the pipeline, consuming sequences of 3D Joints and joint orientations to recognize actions in a closed-set setting. Core contributions are computationally efficient augmentation techniques tailored to real-world surveillance artifacts, and a strong graph-based baseline operating on uplifted 3D Joints with optional 3D joint orientations. Prior work largely reports results under controlled conditions with near-perfect 3D inputs. This thesis conducts deployment-oriented assessment in real-world surveillance scenarios. Evaluation covers ground-truth and uplifted inputs (2D-to-3D), alongside realistic jitter, occlusion, joint dropout, and orientation noise. Analyses compare graph- and sequence-based architectures and study temporal aggregation strategies. Ablations quantify the effect of augmentation, orientation channels, and corruption severity, and latency–accuracy trade-offs are measured to sustain soft real-time operation within the pipeline.

Human Action Recognition System: Finally, the integration of the modules into an overall system for real-world applications is outlined. Due to the scarcity of suitable large-scale datasets—caused by annotation effort and privacy regulations—a synthetic dataset, GTA-RWS, is proposed to facilitate end-to-end evaluation of 2D-HPE, 3D-HPE, and SBAR, which are typically assessed on different datasets. The dataset is generated by pose retargeting of real 3D Motion Capture (MoCap) data into the GTA V environment, a platform repeatedly used for synthetic surveillance datasets [Fab18, Koh20, Fab21].

4 Experimental Setup

The experimental setup employed to assess the proposed methods for 2D-HPE (Section 4.1), 3D-HPE (Section 4.2), and SBAR (Section 4.3) is presented in this chapter. Emphasis is placed on fairness, reproducibility, and comparability across tasks and datasets. Standardized data preprocessing and sample definitions are applied where appropriate, and consistent reporting conventions are used for all metrics and splits.

For 2D-HPE, two evaluation settings are adopted to reflect common deployment scenarios. In the specialization setting, training and testing are performed within the same dataset. In the generalization setting, transfer across datasets is targeted, leveraging UPAR-Pose and UPPET together with cross-domain protocols (3-Fold Cross-Validation (3FCV)/4-Fold Cross-Validation (4FCV) and Leave-One-Out Cross-Validation (LOOCV)) detailed in Sections 4.1.3, 4.2.3 and 4.3.3. Cross-validation is employed to obtain robust estimates, with per-domain scoring and averaging over folds as specified in the respective protocol sections.

The chapter is organized by task. For each task, the employed datasets are summarized in Sections 4.1.1, 4.2.1 and 4.3.1. The evaluation measures are defined in Sections 4.1.2, 4.2.2 and 4.3.2. The evaluation protocols, including split definitions and aggregation procedures, are provided in Sections 4.1.3, 4.2.3 and 4.3.3.

4.1 2D Human Pose Estimation

In the following, the experimental setup is presented, which is used to evaluate the proposed methods in the context of this thesis for 2D-HPE. First, an

overview of datasets of the 2D-HPE utilized is provided in Section 4.1.1. In particular, the datasets contributed by the author are highlighted. The evaluation metrics are then discussed in Section 4.1.2, followed by the introduction of the evaluation protocols in Section 4.1.3.

4.1.1 Datasets

An overview of representative 2D-HPE datasets used for training and evaluating models is summarized in Table 4.1. Early efforts in large-scale annotation focused on in-the-wild image collections such as MPII [And14] and COCO [Lin14], which remain widely adopted in the community. Both datasets provide annotations of human skeletons in natural environments, covering a broad range of activities and camera viewpoints. Extensions such as PoseTrack [And18, Dör22] incorporate temporal continuity, thereby enabling the development of pose tracking methods across video sequences. While these resources have proven essential for general-purpose benchmarks, their design is largely constrained to scenarios with favorable conditions: high resolution RGB images, mostly frontal viewpoints, and relatively sparse occlusion. Several datasets have been introduced to specifically address challenges encountered in more difficult conditions. OCHuman [Zha19a] and CrowdPose [Li19b] target crowded environments with high levels of inter-person occlusion, while AI Challenger [Wu17] emphasizes large-scale diversity in everyday settings. Despite these advances, most existing resources still represent human activities in generic contexts and do not capture the unique difficulties posed by surveillance-oriented deployments, where camera viewpoints are steep, occlusion is frequent, and scale variations are significant [Cor22b]. Datasets such as HiEve [Lin23] have made progress in this direction by introducing long surveillance videos in complex urban scenes, yet they remain limited to the RGB modality. At the same time, thermal imaging has recently gained traction for human analysis under low-light and adverse weather conditions for detection [Jia21], tracking [Sta25], and 2D-HPE [Kuz24]. However, current thermal datasets are still relatively small, lack video continuity, and often do not align with the requirements of

multi-camera surveillance evaluation. Furthermore, only a few have human pose annotations [Kuz24, Tan23].

The limited size of 2D human pose estimation datasets is largely due to the difficulty of annotating real data by human annotators. The process is long, tedious, and prone to errors [Cor21a, Cor21b]. To address these limitations, several datasets are introduced in this thesis. These benchmarks explicitly target 2D-HPE in surveillance environments, combining visible-spectrum and thermal modalities. By focusing on realistic camera viewpoints, denser crowd scenarios, and multimodal annotations, they provide a more suitable foundation for evaluating pose estimation models under operational conditions. The contributed datasets are an important step forward in bridging the gap between controlled benchmark datasets and real-world deployment challenges.

Table 4.1: Overview of 2D-HPE datasets – Selection of publicly available datasets for 2D-HPE, grouped into image- and video-based benchmarks.

Dataset	Environment	Number of Skeletons	Keypoints
Images			
COCO [Lin14]	in-the-wild	250K	17
MPII [And14]	in-the-wild	40K	16
AI Challenger [Wu17]	in-the-wild	375K	14
CrowdPose [Li19b]	crowded scenes	80K	14
OCHuman [Zha19a]	crowded scenes	133K	17
OpenThermalPose [Kuz24]	thermal imaging	200K	17
Videos			
PoseTrack (2018/2021) [And18, Dör22]	in-the-wild	153K	15/17
HiEve [Lin23]	surveillance	1M	17

4.1.1.1 UPAR

An objective of this work is to enable privacy-compliant suspect description in surveillance scenarios, motivating the need for PAR alongside pose estimation. To this end, surveillance data is crucial, yet publicly available surveillance-oriented datasets remain scarce. Generalization in real deployments requires diversity in illumination, viewpoints, occlusions, and environments—but such data are rare, and privacy regulations often preclude

sharing raw surveillance footage. Soft biometrics (*e.g.*, age, gender, hair length, clothing colors and lengths, accessories) offer a privacy-friendly alternative that is useful to law enforcement while avoiding direct identity disclosure. However, existing public PAR datasets are limited in scale and diversity. Such corpora typically lack pose annotations, while creating new real-world datasets is costly, labor-intensive, and ethically complex.

Therefore, this thesis adopts a two-stage solution. The first stage introduces UPAR, a unified dataset for PAR and PAR-based person retrieval designed to support realistic generalization. The second stage extends this resource with human pose annotations (UPAR-Pose; see Section 4.1.1.2) to jointly study PAR and 2D-HPE in surveillance conditions.

Publicly available PAR datasets are heavily biased and constrained in scene diversity. Except for Pedestrian Attribute 100K (PA-100K) [Liu17] and Richly Annotated Pedestrian v2 (RAPv2) [Li19a], most public PAR datasets contain fewer than 50,000 annotated pedestrians and focus on either indoor or outdoor settings. For example, Market-1501 (Market-1501) [Den14, Lin19] contains outdoor images captured during summer on a university campus. Attributes are thus biased towards young Asian people wearing light summer clothing and do not reflect real-world diversity. Consequently, findings on such datasets are difficult to transfer to new domains. Moreover, there are pronounced distribution shifts in both attribute frequencies and image content across datasets. Real-world applications require out-of-distribution generalization across datasets with substantially different characteristics. To enable research under such realistic conditions, UPAR is proposed, which unifies multiple PAR datasets at the attribute level and introduces two evaluation protocols focused on generalization (see Section 4.1.3).

UPAR is constructed by combining images from four public datasets, as illustrated in Figure 4.1: PA-100K [Liu17], Pedestrian Attribute (PETA) [Den14], RAPv2 [Li19a], and Market-1501 Attribute [Zhe15, Lin19]. Given the practical and ethical limitations of curating new surveillance datasets—particularly regarding privacy—the author of this thesis argues that existing resources remain underutilized and should be leveraged more effectively via harmonized annotations. These datasets collectively offer broader diversity in

scenarios, ethnicities, and attributes, which means that UPAR is less biased than any single source. Market-1501 Attribute augments Market-1501 with 27 attributes across 32,668 bounding boxes for 1,501 identities captured by five high-resolution and one low-resolution camera. PA-100K contains 100,000 pedestrian images from various outdoor surveillance cameras, with substantial variation in resolution, lighting, and environments, and provides 26 attributes. RAP comes in two versions: Richly Annotated Pedestrian v1 (RAPv1) [Li15] includes 41,585 images with 72 attributes captured over three months in shopping malls; RAPv2 [Li19a] extends this to 84,928 images, 2,589 identities, and 25 scenes, retaining the RAPv1 attributes to support person retrieval and PAR in indoor surveillance scenarios. PETA [Den14] aggregates 19,000 images from 10 datasets (indoor and outdoor), with 61 binary and four multiclass attributes. Although widely used for benchmarking, these datasets share only a small subset of attributes, hindering cross-dataset generalization and evaluation. To address this, attribute definitions are harmonized across the four datasets into a single, cross-compatible scheme.

Figure 4.1: Sample images from the sub-datasets included in the UPAR dataset – Sample images from different PAR datasets [Li15, Li19a, Den14, Lin19]. Each dataset shows different characteristics and scenarios. However, meaningful out-of-distribution evaluation was impossible due to a low number of common attributes across the datasets. The contributed UPAR dataset unifies 40 attributes and thus enables cross-domain investigations.

Attribute selection was guided by relevance for video-surveillance person retrieval, balancing global-scale attributes (*e.g.*, age, gender) and small-scale cues (*e.g.*, glasses). The unified set comprises 40 binary attributes grouped into 12 categories: age, gender, hair length, upper-body clothing length, upper-body clothing color, lower-body clothing length, lower-body clothing color, lower-body clothing type, accessory-backpack, accessory-bag, accessory-glasses, and accessory-hat. After removing images that could not be reliably annotated, UPAR contains 224,737 images with labels for all attributes. The dataset follows the original splits, yielding 148,048 train,

30,830 validation, and 45,859 test images. Dense color annotations are added for 100,000 images (11 unique colors plus "other" and "mixture"), contributing 2.57M new binary color labels. Furthermore, this work adds lower-body clothing length for PETA and RAPv2, glasses for PA-100K, PETA, and Market-1501, age for PETA, and hair length for PA-100K and Market-1501, yielding an additional 0.8M labels. In total, UPAR contributes 3.3M manually labeled and validated new binary annotations. With these additions, this work enables generalization assessment and also supports training and evaluation on the full dataset for specialization studies. Three publications by the author are related to this dataset [Spe23, Cor23, Cor24b]. UPAR has already proven useful for PAR and PAR-based retrieval research in surveillance contexts (*c.f.* [Spe24]).

The contribution is highlighted by the use of the highly specialized Antonn software contributed by the same author and detailed in Appendix A. Furthermore, clear annotation processes are defined and implemented which are illustrated in Appendix B.1.

The final dataset is constructed by merging the sub-datasets with the new labels. For each sub-dataset, UPAR attributes are taken from original annotations where available or from the new additions. Some attributes (*e.g.*, specific colors) are not present in all sub-datasets. To harmonize label dimensionality, the missing attributes are marked as not present for that sub-dataset. Images labeled unknown for attributes that are not visibly determinable (*e.g.*, heavy occlusion, severe truncation), which is particularly frequent in PA-100K, are discarded. The 40 binary attributes and their 12 categories are listed in Table 4.2. All attributes are binary: 0 denotes absence and 1 denotes presence. For mutually exclusive attributes (*e.g.*, gender, clothing length), 0 corresponds to the opposite category (*e.g.*, male, long).

Table 4.2: UPAR attribute annotations – Overview of the attribute annotations provided for the UPAR dataset. Annotations include demographic and material soft biometrics.

Category	Age	Gender	Hair length	Upper-body clothing length	Upper-body clothing color	Lower-body clothing length	Lower-body clothing color	Lower-body clothing type	Backpack	Bag	Glasses	Hat
Attributes	Young Adult Elderly	Female	Short Long Bald	Short	Black Blue Brown Green Grey Orange Pink Purple Red White Yellow Other	Short	Black Blue Brown Green Grey Orange Pink Purple Red White Yellow Other	Trousers&Shorts Skirt&Dress	Backpack	Bag	Normal Sun	Hat

4.1.1.2 UPAR-Pose

The second stage extends UPAR with 2D pose annotations, yielding UPAR-Pose—the first resource to jointly address PAR and 2D-HPE under surveillance conditions. Although PAR is a secondary focus of this dissertation, UPAR-Pose is central to the investigation of 2D-HPE for surveillance and cross-dataset generalization. An publication by the author is related to this dataset [Cor26a].

UPAR-Pose is a curated subset of UPAR. Due to the high cost of pose annotation, 24,268 images from PA-100K are annotated with poses, forming the Pedestrian Attribute and Pose Estimation 24K (PA-24K) subset. Additionally, 17,365 poses for PETA and 37,684 poses for Market-1501 are annotated, for a total of 71,015 joint pose-and-attribute annotations in UPAR-Pose.

Following the UPAR primary-person protocol, pose is annotated per frame for the selected person only, using a standard 2D keypoint scheme. Annotators rely solely on visible evidence in the target frame; severe occlusions are marked accordingly. A similar validation workflow (four-eyes principle and an ambiguity database) ensures consistency across scenes and datasets.

UPAR-Pose supports (1) in-distribution training and testing within a single dataset for upper-bound performance analysis, and (2) out-of-distribution evaluation via leave-one-dataset-out splits to assess robustness to dataset shifts (scene, viewpoint, clothing, ethnicity, season). By unifying PAR and pose labels within comparable surveillance imagery, UPAR-Pose enables studying top-down pose estimation informed by attributes, attribute-conditioned specialization, and transfer between PAR and 2D-HPE. It also provides a realistic benchmark for assessing generalization beyond controlled lab conditions, advancing surveillance-oriented 2D-HPE.

4.1.1.3 LLVIP-Pose

Following UPAR-Pose, a modality shift to paired visible–infrared imagery is adopted to address low-light operation and privacy. Although identity can sometimes be inferred from thermal imagery with sufficient context, pose-based representations provide a second layer of anonymization. Visible-spectrum HPE remains vulnerable to low illumination and adverse weather [Jia21, Cho18], whereas thermal infrared imaging is more robust and has demonstrated advantages in person detection as shown in Figure 4.2. However, it remains underexplored for 2D-HPE.

Figure 4.2: Comparison of a night-time scene captured by an RGB (left) and a thermal camera (right). – In the thermal images, humans are clearly visible with hight contrast, while the RGB image suffers from poor lighting. Images taken from LLVIP [Jia21].

LLVIP-P is introduced as an HPE extension of the LLVIP. The original LLVIP train-test split is retained. After annotation and cleaning, 26,135 person bounding boxes with poses are provided: 6,853 image pairs (18,633 persons) for training and 3,462 image pairs (7,502 persons) for testing. The annotation process is briefly described in Appendix B.2. More details are provided in a publication by the author related to this dataset [Cor24c].

4.1.1.4 UPPET

Thermal HPE remains underexplored primarily due to limited pose-annotated data. Aside from SLP [Liu22a], OTP [Kuz24], and LLVIP-P, most public thermal datasets provide fewer than 10,000 annotated pedestrians, whereas COCO [Lin14] contains roughly 250,000 poses for RGB. Resulting biases in clothing, location, and activity reduce transferability to real-world scenarios. Domain shifts arise from indoor/outdoor differences, weather (sunlight, rain, fog, snow), ambient temperature effects on thermal contrast [Nik21], and sensor quality (bandwidth, resolution, sensitivity). These factors complicate generalization and motivate unified, multi-source evaluation.

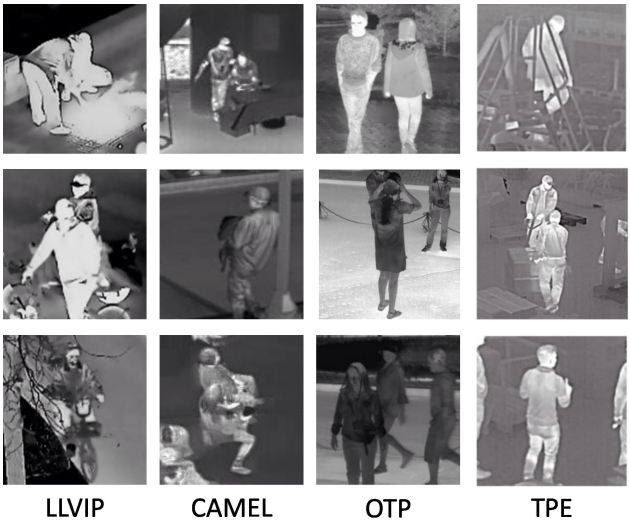


Figure 4.3: Sample images from the sub-datasets included in the UPPET dataset. – The appearance of thermal images varies greatly due to varying sensors and scenes.

Furthermore, rapid progress in thermal sensors has led to heterogeneous cameras with varying resolution, contrast, and noise characteristics; image appearance thus varies markedly by device, as illustrated in Figure 4.3. Training separate HPE models per sensor is impractical at scale, making cross-sensor generalization essential.

As the thermal counterpart to UPAR-Pose, UPPET is introduced to close the gap in thermal HPE resources. UPPET harmonizes annotations across four datasets spanning diverse scenarios and sensors: LLVIP-P, CAMEL-P, OTP, and TPE. An additional publication by the author relates to this dataset [Cor25].

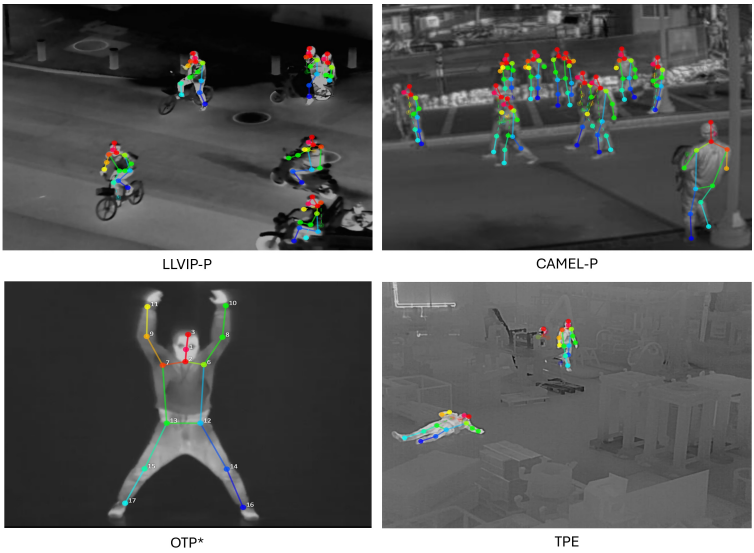


Figure 4.4: Overview of thermal datasets introduced in this work. – (a) LLVIP-P [Cor24c]: same as in Section 4.1.1.3. (b) CAMEL-P: Subset derived from CAMEL [Geb18], with 2,926 images annotated, totaling 25,951 human poses. (c) OTP*: Adapted from OTP [Kuz24] to the PoseTrack18 skeleton topology, with 14,286 pose annotations augmented by adding Nose, Head Top, and Head Bottom keypoints. (d) TPE: Newly collected dataset, comprising 4,321 images and 52,563 pose annotations using a 15-keypoint format.

UPPET is built by combining three public sources— LLVIP-P contributed in this thesis, OTP [Kuz24], and CAMEL [Geb18] (denoted CAMEL-P) originally designed for detection and tracking— and a new industrial dataset (TPE) with bounding boxes, tracks, and poses as contributed in this work. The latter enables research on worker safety by HPE in industrial environments, a scenario not covered by prior datasets. A shared PoseTrack18 topology with 15 keypoints is adopted to unify annotations across datasets. For OTP, which provides 17 COCO keypoints, the head representation is reconciled by annotating Head Top and Head Bottom to match PoseTrack18 (COCO's five head keypoints are collapsed accordingly), reflecting the limited head detail in low-resolution thermal surveillance. For CAMEL, tight person bounding

boxes are first enlarged. Poses are annotated on key frames, subsequently reviewed and interpolated by different annotators, and finally validated. As CAMEL includes varied scenes, four elevated-view sequences (3, 4, 26, 29) are selected and split at person entry/exit points into train and test subsets, yielding CAMEL-P.

The TPE dataset was collected at a single industrial site with volunteer workers. An AXIS Q1952-E 19mm thermal camera was mounted at an elevated position to emulate CCTV. Typical site activities (e.g., carrying objects, forklift operation) were captured, alongside staged incidents (falls, lying down). Volunteers attempted partial and full occlusions, producing challenging poses (kneeling, heavy occlusion). Bounding boxes were first annotated on key frames, interpolated, and, given high initial quality, a separate box review step was omitted. Skeletons were then annotated for selected sequences on key frames and interpolated. All keypoints carry visibility flags (visible/occluded).

For OTP, 14,286 poses are harmonized. For CAMEL-P, 2,926 images yield 25,951 poses: the training set has 2,121 images with 18,259 poses (8.49 poses/image), and the test set 775 images with 7,692 poses (9.93 poses/image). The TPE dataset comprises 14,321 images and 52,563 poses: the training set contains 9,775 images with 38,887 poses (4.50 poses/image), and the test set 4,546 images with 13,676 poses (3.48 poses/image). In total, UPPET includes 33,654 images with 118,924 poses; the training split has 23,291 images with 86,316 poses, and the test split 10,363 images with 32,608 poses, therefore UPPET is the largest thermal human pose estimation dataset to date. UPPET may be used in full or via the proposed protocols for specialization/generalization studies (see Table 4.7).

4.1.2 Evaluation Measures

In this section, the common measures typically used for evaluating the performance of 2D-HPE and PAR are presented. In this section, all evaluation results are presented as percentages, unless indicated otherwise.

4.1.2.1 2D Human Pose Estimation

First introduced with the COCO dataset [Lin14], the OKS metric adapts the concept of Intersection Over Union (IoU) to keypoint localization in 2D-HPE.

Object Keypoint Similarity (OKS). The OKS quantifies the similarity between predicted and ground truth keypoints, normalized by the person's scale. It is defined as

$$\mathrm{OKS} = \frac{\sum_{k=1}^{K} \exp\left(-\frac{d_{L2}(\hat{\mathbf{k}}_s^r, \mathbf{k}_s^r)^2(k)}{2\,s^2\,\sigma_{s_k}^2}\right) \cdot \delta(\mathrm{vis}(k) > 0)}{\sum_{k=1}^{K} \delta(\mathrm{vis}(k) > 0)}, \tag{4.1}$$

where K is the number of annotated keypoints, $d_{L2}(\hat{\mathbf{k}}_s^r, \mathbf{k}_s^r)(k) = \|\hat{\mathbf{K}^r}_k - \mathbf{K^r}_k\|_2$ denotes the Euclidean distance between the predicted and ground-truth locations of keypoint k, $s = \sqrt{h_b \cdot w_b}$ represents the scale of the person given bounding-box height h_b and width w_b, σ_{sk} is a per-keypoint constant accounting for annotation uncertainty, and $\mathrm{vis}(k)$ is the COCO visibility flag. The indicator function $\delta(\cdot)$ ensures that only visible or labeled keypoints contribute.

Average Precision (AP). Based on OKS, detections can be labeled as true positives (TP), false positives (FP), or false negatives (FN) under a threshold $t \in [0,1]$. For a recall level set $\mathrm{R_{set}}$, precision is computed as

$$\mathrm{AP}_t = \frac{\sum_{r \in \mathrm{R_{set}}} \mathrm{TP}_{t,r}}{\sum_{r \in \mathrm{R_{set}}} \left(\mathrm{TP}_{t,r} + \mathrm{FP}_{t,r}\right)}. \tag{4.2}$$

The overall AP is then averaged across the ten standard OKS thresholds $\mathrm{IoU_{set}} = \{0.50, 0.55, \ldots, 0.95\}$:

$$\mathrm{AP} = \frac{1}{|\mathrm{IoU_{set}}|} \sum_{t \in \mathrm{IoU_{set}}} \mathrm{AP}_t. \tag{4.3}$$

Average Recall (AR). Analogously, recall is defined as the proportion of true positives among all relevant instances:

$$\mathrm{AR} = \frac{\sum_{r \in \mathrm{R}_{\mathrm{set}}} \mathrm{TP}_{t,r}}{\sum_{r \in \mathrm{R}_{\mathrm{set}}} \left(\mathrm{TP}_{t,r} + \mathrm{FN}_{t,r} \right)}. \tag{4.4}$$

Variants. In practice, AP and its counterpart AP^{50} (threshold $t = 0.50$) are commonly reported, along with the scale-specific measures AP_S, AP_M, and AP_L, which focus on small ($< 32^2$ pixels), medium ($32^2 \leq$ area $< 96^2$), and large ($\geq 96^2$) persons, respectively.

4.1.2.2 Pedestrian Attribute Recognition

In the evaluation of Person Attribute Recognition (PAR) approaches, two types of metrics are used [Spe24]: label-based and instance-based criteria. These metrics differ in their underlying assumptions and the aspects they prioritize.

Label-based metrics treat attributes as independent entities. Initially, the performance for each attribute is computed separately, and the final evaluation is obtained by averaging these scores across all attributes. In contrast, instance-based metrics account for inter-attribute dependencies. Attributes are inherently correlated, as certain attributes significantly influence the likelihood of others, *e.g.*, skirt correlating with female or short lower-body clothing with short upper-body clothing. Instance-based metrics evaluate attributes collectively for each image and then average the results across multiple samples. This ensures a more meaningful assessment of attribute consistency within a pedestrian image, which is particularly valuable for suspect description. Instead of analyzing individual attributes in isolation, the goal is to capture a comprehensive description of the depicted person. In this thesis, both types of metrics are applied.

Label-based Mean Accuracy (mA): The mA metric represents a label-based evaluation criterion and was originally adapted for the PAR task by [Den14].

In contrast to raw accuracy, mA separates the evaluation of positive and negative samples, addressing the challenge of imbalanced attribute distributions. Without this adjustment, a naive solution of always predicting the absence of attributes would result in artificially high accuracy, as many semantic attributes are infrequent. Specifically, the mA metric calculates the mean recall for positive and negative samples, and then averages these values across all attributes. The formulation is given as:

$$\mathrm{mA} = \frac{1}{2L} \sum_{j=1}^{L} \left(\frac{\mathrm{TP}_j}{\mathrm{P}_j} + \frac{\mathrm{TN}_j}{\mathrm{N}_j} \right), \tag{4.5}$$

where L denotes the total number of attributes, and TP_j, P_j, TN_j, and N_j represent the number of true positives, total positives, true negatives, and total negatives for the j-th attribute, respectively.

Instance-based F1: Instance-based metrics differ from label-based by evaluating attributes collectively for each image. The instance-based F1 metric [Li15] relies on two key measures: instance-based precision $\mathrm{Prec}_{\mathrm{PAR}}$ and recall $\mathrm{Rec}_{\mathrm{PAR}}$. For instance-based metrics, precision and recall are computed per image $\mathbf{I}_i$ and then averaged across the dataset of $\mathrm{N}_{\mathrm{samples}}^{2D}$ images.

Precision measures the fraction of correctly predicted attributes out of all predicted attributes for a given image. It is formally defined as:

$$\mathrm{Prec}_{\mathrm{PAR}} = \frac{1}{\mathrm{N}_{\mathrm{samples}}^{2D}} \sum_{i=1}^{\mathrm{N}_{\mathrm{samples}}^{2D}} \frac{|Y_i \cap f(\mathbf{I}_i)|}{|f(\mathbf{I}_i)|}, \tag{4.6}$$

where Y_i denotes the ground truth positive labels for the i-th image, $f(\mathbf{I}_i)$ returns the predicted positive labels, and $|\cdot|$ signifies set cardinality.

Recall measures the proportion of ground truth attributes correctly identified by the PAR approach. It is expressed as:

$$\mathrm{Rec}_{\mathrm{PAR}} = \frac{1}{\mathrm{N}^{2D}_{\mathrm{samples}}} \sum_{i=1}^{\mathrm{N}^{2D}_{\mathrm{samples}}} \frac{|Y_i \cap f(\mathbf{I}_i)|}{|Y_i|}. \tag{4.7}$$

This calculation is similar to precision but divides the correctly identified attributes by the total number of attributes present in the ground truth.

Finally, the instance-based F1 score $\mathrm{F1}_{\mathrm{PAR}}$ combines precision and recall using their harmonic mean:

$$\mathrm{F1}_{\mathrm{PAR}} = \frac{2 \cdot \mathrm{Prec}_{\mathrm{PAR}} \cdot \mathrm{Rec}_{\mathrm{PAR}}}{\mathrm{Prec}_{\mathrm{PAR}} + \mathrm{Rec}_{\mathrm{PAR}}}. \tag{4.8}$$

4.1.3 Evaluation Protocols

Evaluation protocols for the 2D-HPE and PAR tasks are presented, with assessments organized into two settings: specialization and generalization. Specialization denotes experiments confined to a single dataset, with training and testing drawn from the same source. Generalization, enabled by UPAR-Pose and UPPET, targets the capacity of 2D-HPE and PAR methods to transfer across datasets. A cross-validation framework is employed, using disjoint datasets for training and testing to promote rigorous and reliable assessment.

4.1.3.1 Specialization

For PAR and Top-Down 2D-HPE, evaluation is carried out on person-centric crops derived from bounding boxes. Each bounding box constitutes a single sample, thereby enforcing a consistent, sample-wise protocol. In multi-person imagery, one crop is generated per bounding box; in single-person imagery, the bounding box delineates the pertinent region. By concentrating on the cropped samples, evaluation is standardized across tasks and precise measurement of attribute recognition and pose estimation performance is facilitated.

A summary of the statistics of all 2D-HPE datasets contributed in this thesis is given in Table 4.3. The datasets span a wide range of modalities and crowding levels, from thermal imagery (OTP) to large-scale multi-person surveillance scenes (CAMEL-P, TPE). Single-person datasets such as PA-24K, PETA, and Market-1501 provide harmonized attributes, while UPAR-Pose extends this with unified pose and attribute annotations, enabling joint evaluation of pose estimation and person analysis tasks.

Table 4.3: Contributed 2D-HPE datasets – Statitics of all contributed datasets in this thesis. The UPAR-Pose datasets also provides harmonized PAR annotations for 40 attributes. * means the dataset had already pose annotations which have been extended.

Dataset	Train				Test			
	#Images	#Poses	Avg. #Keypoints	Avg. #Poses	#Images	#Poses	Avg. #Keypoints	Avg. #Poses
OTP [Kuz24]*	4,511	10,537	12.83	2.34	1,579	3,749	12.75	2.37
LLVIP-P	6,854	18,633	12.47	2.74	3,463	7,500	12.14	2.19
CAMEL-P	2,151	18,259	12.53	8.49	775	7,692	12.10	9.93
TPE	9,775	38,887	11.95	4.50	4,546	13,676	11.85	3.81
PA-24K	19,500	19,499	13.07	1.00	4,768	4,768	13.16	1.00
PETA	10,402	10,314	14.21	1.00	6,963	6,909	14.20	1.00
Market-1501	21,226	21,226	13.70	1.00	16.458	16,458	13.85	1.00
UPPET	23,291	86,316	12.29	3.91	10,363	32,608	12.08	3.48
UPAR-Pose	42,826	42,736	13.54	1.00	28,189	28,135	13.82	1.00

4.1.3.2 Generalization

Individual datasets lack the pronounced domain shifts characteristic of real-world deployments. The UPAR-Pose and UPPET datasets were introduced to mitigate this limitation and thereby enable alternative evaluation schemes rooted in domain generalization [Bla11]. Two protocols are defined to assess generalization, distinguished by the volume and diversity of training data permitted. Both protocols rely on cross-validation across three/four sub-domains. The first, 3FCV/4FCV (UPAR-Pose / UPPET), constitutes the more demanding setting, as training is restricted to a single dataset.

The second scheme is LOOCV, which assumes availability of heterogeneous training data from multiple sources. In each of the three/four folds, one sub-domain is reserved for evaluation, while the remaining two/three sub-domains are used for training. The allocation of datasets to the respective splits is reported in Table 4.4 and Table 4.6.

Table 4.4: UPAR-Pose splits – Split definitions for the two UPAR-Pose generalization evaluation schemes. 3FCV protocol is more challenging since only sub-dataset is used for training. In both cases, evaluation is performed on unseen domains.

Split ID	LOOCV		3FCV	
	Training	Evaluation	Training	Evaluation
0	PA-24K, PETA	Market-1501	Market-1501	PA-24K, PETA,
1	Market-1501, PETA	PA-24K	PA-24K	Market-1501, PETA
2	Market-1501, PA-24K	PETA	PETA	Market-1501, PA-24K

Table 4.5: UPAR-Pose split statistics – Statistics of the four splits for each of the two evaluation protocols. Each image represent one person.

Split ID	LOOCV					
	Train			Test		
	#Images	#Poses	Avg. #Keypoints	#Images	#Poses	Avg. #Keypoints
0	29,902	29,813	13.46	16,458	16,458	13.85
1	23,326	23,237	13.94	4,768	4,768	13.16
2	32,424	32,422	12.23	6,963	6,909	14.20
Split ID	3FCV					
	Train			Test		
	#Images	#Poses	Avg. #Keypoints	#Images	#Poses	Avg. #Keypoints
0	12,924	12,923	13.72	11,731	11,677	13.77
1	19,500	19,499	13.07	23,421	23,367	13.95
2	10,402	10,314	14.21	21,226	21,226	13.70

Statistics of training and test images and keypoints per split are provided in Table 4.5 and Table 4.7. The splits follow the original partitions of the constituent datasets to preserve comparability of results.

Table 4.6: UPPET splits – Split definitions for the two UPPET generalization evaluation schemes. 4FCV protocol is more challenging since only sub-dataset is used for training. In both cases, evaluation is performed on unseen domains.

Split ID	LOOCV		4FCV	
	Training	Evaluation	Training	Evaluation
0	OTP, CAMEL-P, TPE	LLVIP-P	LLVIP-P	OTP, CAMEL-P, TPE
1	LLVIP-P, CAMEL-P,TPE	OTP	OTP	LLVIP-P, CAMEL-P, TPE
2	LLVIP-P, OTP, TPE	CAMEL-P	CAMEL-P	LLVIP-P, OTP, TPE
3	LLVIP-P, CAMEL-P, OTP	TPE	TPE	LLVIP-P, CAMEL-P, OTP

Table 4.7: UPPET split statistics – Statistics of the four splits for each of the two evaluation protocols.

Split ID	LOOCV							
	Train				Test			
	#Images	#Poses	Avg. #Keypoints	Avg. #Poses	#Images	#Poses	Avg. #Keypoints	Avg. #Poses
0	16,437	67,683	12.24	4.42	3,463	7,500	12.14	2.19
1	18,780	75,779	12.22	4.31	1,579	3,740	12.75	2.37
2	21,140	68,057	12.23	3.41	775	7,692	12.10	9.93
3	13,516	47,429	12.57	3.52	4,546	13,676	11.85	3.81
Split ID	4FCV							
	Train				Test			
	#Images	#Poses	Avg. #Keypoints	Avg. #Poses	#Images	#Poses	Avg. #Keypoints	Avg. #Poses
0	6,854	18,633	12.47	2.74	6,900	25,108	12.06	4.23
1	4,511	10,537	12.83	2.34	8,784	28,868	11.99	3.70
2	2,151	18,259	12.53	8.49	9,588	24,916	12.07	2.90
3	9,775	38,887	11.95	4.50	5,817	18,932	12.24	3.27

Image totals may not exactly match those of the sub-datasets or their sources. During the construction of UPAR, UPAR-Pose, and UPPET, images with inconsistent annotations or lacking a clear human depiction were removed.

Final scores for both protocols are computed as follows. Metrics are first obtained independently for each test domain. For 3FCV/4FCV, an average across test domains yields the per-split score. In both protocols, the final outcome is the mean over splits. This aggregation assigns equal influence to all evaluation subsets, preventing domination by sub-domains with larger test sets.

4.2 3D Human Pose Estimation

In the following, the experimental setup which is used to evaluate the proposed methods in the context of this thesis for 3D-HPE, is presented. First, an overview of the 3D-HPE datasets utilized is provided in Section 4.2.1. The evaluation metrics are then discussed in Section 4.2.2, followed by the introduction of the evaluation protocols in Section 4.2.3.

4.2.1 Datasets

An overview of video 3D Human Pose Estimation datasets used as benchmarks for training and evaluating monocular 3D pose estimation models, which infer 3D poses from single-camera 2D observations, is provided in Table 4.8. Existing datasets are constructed using two main approaches: marker-based motion capture systems, which rely on optical tracking systems and infrared cameras (e.g., Human3.6M (H36M) [Ion13], 3DPW [Von18], HumanEva [Sig10] and Fit3D [Fie21]), and markerless methods, which use industrial-grade cameras and computer vision algorithms (e.g., MPI-INF-3DHP [Meh17], CMU Panoptic [Joo15], SportsPose [Ing23], FS-Jump3D [Tan24b] and AthletePose3D (AP3D) [Yeu25]). Marker-based systems are precise, however, limited to controlled environments, while markerless systems are more flexible for naturalistic settings, trading off some accuracy. Datasets such as H36M [Ion13] and MPI-INF-3DHP [Meh17] are standard benchmarks for general-purpose motion analysis. H36M provides high-quality marker-based data of daily activities in controlled indoor settings, while MPI-INF-3DHP includes both indoor and outdoor scenes using markerless systems, offering greater diversity. Specialized datasets such as SportsPose [Ing23], ASPset-510 [Nib21], and AIST++ [Li21a] focus on athletic or domain-specific motions. SportsPose is constrained to amateur athletes and fails to capture the high-acceleration dynamics of competitive sports, while AIST++ is restricted to dance movements. Similarly, FS-Jump3D [Tan24b] targets figure skating and has limited applicability beyond this domain. Most existing datasets focus on controlled or specialized settings and rarely address the high-speed, high-acceleration scenarios typical of surveillance or competitive sports. Surveillance applications, in particular, require elevated perspectives and robust 3D-HPE for dynamic scenes involving actions such as kicks, punches, and rapid movements scenarios underrepresented in current datasets.

Table 4.8: Overview of 3D-HPE datasets – The table shows publicly available research datasets for 3D-HPE

Dataset	Environment	Subjects	Keypoints	Poses	Cameras	Markerless	FPS	Frames
Human3.6M [Ion13]	lab	11	26	900K	4	×	50	3.6M
MPI-INF-3DHP [Meh17]	lab & outdoor	8	28	93K	14	✓	25/50	1.3M
3DPW [Von18]	lab & outdoor	7	24	49K	1	×	30	51K
HumanEva-I [Sig10]	lab	6	15	78K	7	×	60	280K
HumanEva-II [Sig10]	lab	6	15	3K	4	×	60	10K
CMU Panoptic [Joo15]	lab	8	18	1.5M	31	✓	30	46.5M
AIST++ [Li21a]	lab	30	17	1.1M	9	✓	60	10.1M
ASPset-510 [Nib21]	outdoor	17	17	110K	3	✓	50	330K
SportsPose [Ing23]	lab & outdoor	24	17	177K	7	✓	90	1.5M
Fit3D [Fie21]	lab	13	25	2.9M	8/12	×	50	2.9M
AthletePose3D [Yeu25]	lab & ice rink	8	55/86	165K	4/8/12	✓	60/120	1.3M

This work focuses on datasets aligned with studying standard human-centric 2D-to-3D uplifting models in dynamic, unstructured settings. Specifically, H36M [Ion13], Fit3D [Fie21], and AP3D [Yeu25] are selected. Furthermore, Harmony4D (H4D) is adapted and used for 3D-HPE. H36M serves as a foundational benchmark with high-quality marker-based data, Fit3D captures dynamic fitness movements relevant to urban surveillance, AP3D introduces markerless, high-acceleration actions across diverse sports, and H4D introduces fight elements relevant to urban surveillance. Despite the limitations raised above, these datasets remain the most suitable options currently available and are therefore employed as benchmarks in this thesis.

4.2.1.1 Human3.6M

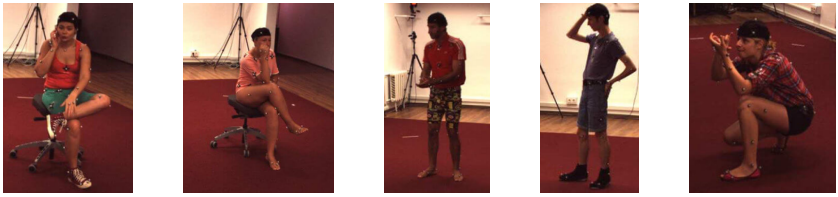


Figure 4.5: Example images from the H36M dataset – The H36M dataset is one of the most widely used benchmarks for 3D human pose estimation. The samples show activities such as sitting and talking on the phone, sitting, standing and talking, standing and thinking, and kneeling while taking a picture, illustrating its coverage of everyday motions in a controlled lab environment.

The H36M dataset [Ion13] is a widely utilized large-scale benchmark for 3D-HPE. As the largest available real-world dataset, it contains 1,376 recorded sequences, encompassing approximately 3.6 million annotated 3D human poses. Each pose is accompanied by four synchronized RGB images, captured from different viewpoints using a multi-camera setup operating at 50 Hz. The dataset includes 11 subjects, with 6 males and 5 females, performing 15 distinct daily activities, including walking, smoking, sitting, greeting, eating, posing, and talking on the phone. Among the 11 subjects, 7 have been annotated with 3D poses. The training split includes data from 5 subjects (S1, S5, S6, S7, and S8), while the test split comprises two subjects (S9 and S11). All sequences were recorded in a controlled indoor environment, using a capture area of 4×3 m. Each sequence focuses on a single individual performing the designated actions, with actors wearing everyday clothing augmented with reflective markers for motion tracking. Sample frames from the dataset are illustrated in Figure 4.5.

Although the dataset is extensive and systematically constructed, it has limitations that restrict its applicability to real-world scenarios. The diversity of participants and activities is constrained, with only 11 subjects and 15 predefined actions, limiting the range of human poses and movements. Furthermore, the controlled indoor setting, with uniform lighting and a plain background, does not reflect the complexity of dynamic, real-world environments. The fixed camera setup offers only a limited range of viewpoints, which may not generalize to scenarios with more varied perspectives. Additionally, the scripted nature of the activities lacks the spontaneity and variability of natural human motion. Finally, there is minimal interaction between participants and their surroundings, such as objects or other individuals, reducing the dataset's relevance for tasks requiring contextual understanding of human action. Despite these limitations, it is still widely used for evaluating 2D-to-3D uplifting methods.

4.2.1.2 AMASS

The Archive of Motion Capture as Surface Shapes (AMASS) dataset [Mah19] is a large-scale resource that consolidates MoCap data from 15 publicly available

datasets into a unified representation. It contains over 40 hours of recorded motion data from 344 subjects, encompassing a wide range of human poses and activities, such as walking, running, jumping, dancing, sitting, and interactions with objects. All motion data are represented using the Skinned Multi-Person Linear (SMPL) body model, a parametric representation that encodes body shape and pose as 3D mesh surface deformations. This standardization ensures compatibility across datasets and simplifies downstream tasks such as 3D-HPE, motion synthesis, and biomechanical analysis. AMASS combines data from diverse sources, capturing a broader range of body types and motion dynamics compared to individual datasets. However, the lack of corresponding RGB recordings limits its use for image-based tasks. Additionally, variations in recording conditions and quality across the contributing datasets may introduce inconsistencies.

Despite these limitations, AMASS is often used in conjunction with other datasets to improve the generalization of machine learning models. A common practice is to project AMASS data into a H36M compatible format, thus suitable for training alongside datasets that include synchronized RGB images. This allows AMASS to serve as a valuable resource for pre-training or multi-dataset training in tasks such as 3D-HPE. For instance, models such as Motion Bidirectional Encoder Representations from Transformers (MotionBERT) [Zhu23] leverage AMASS data during pre-training to enhance their ability to generalize across different datasets and motion patterns, effectively bridging the gap between large-scale MoCap datasets and smaller, video-based datasets.

4.2.1.3 Fit3D

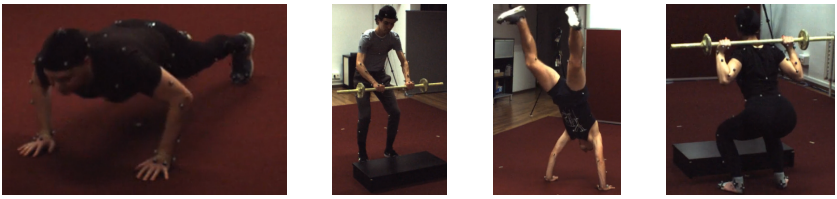


Figure 4.6: Example images from the Fit3D dataset – The Fit3D dataset focuses on fitness-related human motion. The samples depict activities such as push-ups, barbell rows, mule kicks and squats reflecting its emphasis on controlled exercise scenarios with diverse body articulations.

The Fit3D dataset [Fie21], designed to support research in fitness training and motion analysis, provides a large-scale collection of 3D MoCap data synchronized with RGB images. It includes recordings of 11 human subjects with 8 males and 3 females performing various fitness exercises. Among these participants is one licensed fitness instructor, whose movements serve as a reference for correct exercise execution, while the remaining subjects represent trainees with varying levels of skill. The data were captured using a VICON motion capture system equipped with 12 motion cameras and 4 synchronized RGB cameras. Reflective markers were affixed to the skin or clothing of the participants, with all subjects dressed in gym-appropriate, typically form-fitting attire to ensure accurate motion tracking. Several common gym objects were used during the recordings, including two dumbbells, a barbell, a rubber band, and a low-height. The exercises target major muscle groups, such as the arms, legs, back, and abdomen, and are categorized into two types: simple exercises involving basic repetitive movements (e.g., push-ups, squats, dumbbell biceps raises) and compound exercises combining multiple movements (e.g., burpees, which include a push-up and a jump, or clean-and-press routines requiring specific arm trajectories). Each participant performed at least five repetitions for each exercise type. Sample frames from the dataset are illustrated in Figure 4.6.

The dataset covers a range of physical characteristics, with participant heights ranging between 1.55 m and 1.90 m and weights between 60 kg and 110 kg. It includes both physically fit individuals who engage in high levels of physical activity and less trained participants. In total, Fit3D consists of 2,964,236 unique 3D skeletons synchronized with RGB frames. In addition, all participants were scanned in 3D to capture detailed body shape information.

The dataset is officially split into training and validation sets (8 subjects, 2,278,572 images) and a test set (3 subjects, 685,664 images), with all exercise types represented in both subsets. Each video is manually segmented into individual repetitions, resulting in 2,964 annotated timestamps. However, due to the limited availability of ground truth body shape information, which is only provided for the training subset, the official test subset is not used in this thesis. Instead, following [Lud25b], the original training subset is divided into custom splits: 6 subjects (s03, s04, s05, s07, s08, s10) are used for training, with 1 subject each for validation (s09) and testing (s11).

To enable cross-dataset training and evaluation, only the 17 body joints defined in H36M are extracted from the dataset (against 23 in [Lud25b]). Compared to H36M, Fit3D captures a wider range of extreme poses that are more relevant to real-world, uncontrolled scenarios, such as those encountered in public spaces or outdoor environments. Thus, this dataset is particularly valuable for evaluating models under challenging conditions that go beyond basic human activities.

4.2.1.4 AthletePose3D

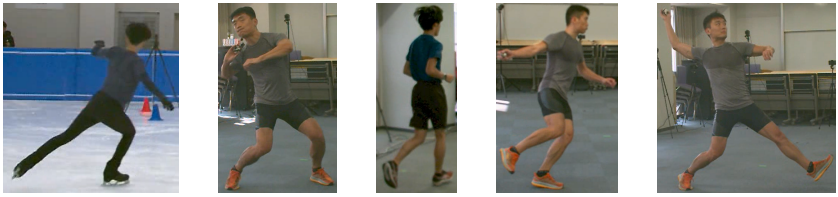


Figure 4.7: Example images from the AP3D dataset – The AP3D dataset is a large-scale benchmark for 3D human pose estimation in sports. The samples highlight its coverage of diverse athletic activities and challenging motion dynamics.

The AP3D dataset [Yeu25] is a recent benchmark designed for 3D-HPE and kinematic validation, focusing on high-speed, high-acceleration athletic movements. It contains approximately 1.3 million frames and 165,000 unique postures, covering 12 sports-specific motion types across three disciplines: running, track and field, and figure skating. These motions include activities such as javelin throw, discus, and figure skating jumps such as Axel, Salchow, and Lutz. Data were collected from 8 athletes, ranging from university-level representatives to national and international competitors. Sample frames from the dataset are illustrated in Figure 4.7. The dataset was recorded using a hardware-synchronized multi-camera system with 4, 8, or 12 high-speed cameras, depending on the complexity of the sport. Running, characterized by rapid limb movements, was recorded at 120 FPS, while track and field and figure skating were recorded at 60 FPS. Videos were captured in 1920×1080 resolution, and the motion capture system was calibrated using a wand-based method, achieving a spatial error of less than 1 mm. Each video sequence is short, reflecting the high-intensity nature of athletic motions, with only 81 frames per sequence used for evaluation. The dataset also provides camera parameters and valid frame indices for 3D-HPE.

AP3D is officially split into training, validation, and test subsets with a 60/20/20 ratio. Compared to general-purpose 3D-HPE datasets such as H36M, AP3D features significantly higher joint speeds and accelerations, reflecting the dynamics of professional sports. This is particularly relevant for evaluating models in sports biomechanics and scenarios requiring robust motion analysis under dynamic conditions. However, its focus on short-duration, controlled sports actions limits its applicability to everyday or prolonged activities. Additionally, the exclusive inclusion of professional athletes may overlook non-professional or recreational movements that occur in broader real-world settings. While these limitations narrow its scope, the high-acceleration and dynamic motions captured in AP3D align well with highly dynamic, high-intensity scenarios, such as fight scenes or other fast-paced activities. For better compatibility with the other datasets

described above, the sequences length is reduced further to 27 frames. In this work, the dataset is extracted with the 17 body joints as defined in H36M.

4.2.1.5 Harmony4D

Figure 4.8: Example images from the H4D dataset – The H4D dataset captures close-contact human interactions in naturalistic settings. The samples illustrate diverse activities including hugging, grappling, sword fighting, karate, and mixed martial arts, highlighting the dataset's focus on dynamic, contact-heavy motion.

H4D [Khi24b] is a large-scale dataset capturing close-contact human interactions across five distinct scenes, featuring pairs of interacting subjects from a pool of 24 participants. It contains high-resolution videos of dynamic activities such as wrestling, dancing, karate, MMA, and fencing. The dataset was recorded using 20 synchronized and calibrated cameras, producing 1.66 million images and 3.32 million annotated human instances. Ground-truth 3D joint positions are provided in COCO topology, along with camera parameters, enabling multi-view analysis and 3D reconstruction. Sample frames from the dataset are illustrated in Figure 4.8.

The dataset is organized into train and test splits, with 169 sequences for training and 39 for testing. Each sequence contains detailed 3D poses for both subjects, which can be projected to 2D across all cameras. In the preprocessed version used for this work, each frame provides two 3D poses and 40 2D poses corresponding to the 20 cameras. The 2D poses are normalized to $[-1,1]$ based on the camera resolution (3840×2160), and 3D poses are similarly scaled using a factor consistent with AMASS. Sequences are further split into fixed-length clips of 27 frames to ensure compatibility with the other datasets, resulting in a standardized format suitable for training and evaluating 3D human pose estimation models.

Compared to previous lab-controlled or single-person datasets, Harmony4D stands out for its in-the-wild capture and focus on dynamic, contact-heavy interactions. This is particularly relevant for studying multi-person 3D pose estimation and interaction modeling in realistic, challenging scenarios.

4.2.2 Evaluation Measures

In the domain of 3D Human Pose Estimation, several metrics are commonly employed to evaluate the performance of models. These metrics assess the consistency and accuracy of predicted joint locations in 3D space compared to ground truth annotations. The most popular evaluation metrics include the MPJPE, the P-MPJPE, the MPJAE, and the MPJVE. Each metric provides unique insights into the performance of a pose estimation model, as detailed below. Additionally, the use of the MPBLE is proposed in this work for better interpretability.

MPJPE: The MPJPE metric, or Mean Per Joint Position Error, is one of the most widely used measures for evaluating 3D Human Pose Estimation. It calculates the Euclidean distance between the predicted joint positions and the ground truth joint positions, averaged across all joints and samples [Ion13]. Formally, it is defined as:

$$\text{MPJPE} = \frac{1}{\text{N}^{3D}_{\text{samples}} \cdot \text{N}_{\text{joints}}} \sum_{p=1}^{\text{N}^{3D}_{\text{samples}}} \sum_{q=1}^{\text{N}_{\text{joints}}} \| \hat{\mathbf{J}}_{p,q} - \mathbf{J}_{p,q} \|_2, \tag{4.9}$$

where $\text{N}^{3D}_{\text{samples}}$ represents the total number of samples, N_{joints} is the total number of joints per pose, $\hat{\mathbf{J}}_{p,q}$ denotes the predicted 3D coordinates of the q-th joint in the p-th sample, and $\mathbf{J}_{p,q}$ refers to the corresponding ground truth joint coordinates.

P-MPJPE: The Procrustes-aligned MPJPE, denoted as P-MPJPE, extends MPJPE by incorporating a rigid alignment step before calculating the error. This alignment uses Procrustes analysis to eliminate differences caused by global translation, rotation, or scale [Zhu23, Sig10]. The metric evaluates how well the predicted pose matches the ground truth after this alignment

and is expressed as:

$$\text{P-MPJPE} = \frac{1}{\text{N}^{3D}_{\text{samples}} \cdot \text{N}_{\text{joints}}} \sum_{p=1}^{\text{N}^{3D}_{\text{samples}}} \sum_{q=1}^{\text{N}_{\text{joints}}} \| \hat{\mathbf{J}}^{\text{aligned}}{}_{p,q} - \mathbf{J}_{p,q} \|_2, \tag{4.10}$$

where $\hat{\mathbf{J}}^{\text{aligned}}{}_{p,q}$ represents the predicted joint coordinates after Procrustes alignment.

MPJAE: The performance of joint rotations is evaluated using the Mean Per Joint Angular Error (MPJAE) [Von18, Lud25b]. The relative rotation of joint q is defined by the rotation matrix $\mathbf{R}_q \in \mathbb{R}^{3\times3}$. For the entire set of N_{joints} joints, the rotation is represented as $\mathbf{R} = (\mathbf{R}_1, \dots, \mathbf{R}_{\text{N}_{\text{joints}}})^T \in \mathbb{R}^{\text{N}_{\text{joints}}\times3\times3}$. The MPJAE measures the geodesic distance between the estimated joint rotations $\tilde{\mathbf{R}}$ and the ground truth rotations $\mathbf{R}$.

For each joint q, the alignment rotation matrix is computed as:

$$\mathbf{R}'_q = \tilde{\mathbf{R}}_q \mathbf{R}_q^T. \tag{4.11}$$

The matrix $\mathbf{R}'_q$ equals the identity matrix $\mathbf{I} \in \mathbb{R}^{3\times3}$ if the estimated and ground truth rotations perfectly match. Otherwise, $\mathbf{R}'_q$ represents the rotation required to align the predicted orientation with the ground truth. The angle of this rotation, denoted φ, is derived from the trace of the matrix ($\text{trace}(\mathbf{R}'_q) = 1 + 2\cos\varphi$) and computed using the arc-cosine function. The error is reported in radians but is converted to degrees for clarity in evaluation tables.

The MPJAE is formally defined as:

$$\text{MPJAE} = \frac{1}{\text{N}_{\text{joints}}} \sum_{q=1}^{\text{N}_{\text{joints}}} \arccos\left(\frac{\text{trace}(\mathbf{R}'_q) - 1}{2}\right), \tag{4.12}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix, $\mathbf{R}_q$ is the ground truth rotation matrix for the q-th joint, and $\tilde{\mathbf{R}}_q$ is the predicted rotation matrix for the q-th joint.

MPJVE: The Mean Per Joint Velocity Error (MPJVE) measures the error in joint velocities between predictions and ground truth [Pav19b]. This metric is particularly relevant for evaluating the temporal consistency of pose estimations in video sequences. It is defined as:

$$\text{MPJVE} = \frac{1}{\text{N}^{3D}_{\text{samples}} \cdot \text{N}_{\text{joints}}} \sum_{p=1}^{\text{N}^{3D}_{\text{samples}}} \sum_{q=1}^{\text{N}_{\text{joints}}} \| \Delta\hat{\mathbf{J}}_{p,q} - \Delta\mathbf{J}_{p,q} \|_2, \tag{4.13}$$

where $\Delta\hat{\mathbf{J}}_{p,q}$ and $\Delta\mathbf{J}_{p,q}$ denote the temporal differences (velocities) of the predicted and ground truth joint positions, respectively.

MPBLE: The Mean Per Bone Length Error (MPBLE) evaluates the structural consistency of predicted poses by comparing the lengths of bones in the predicted skeleton to those in the ground truth. Bone lengths are calculated as the Euclidean norm of the vector connecting two adjacent joints. The metric computes the absolute difference in lengths for all bones, averaged across all bones and samples. It is formally defined as:

$$\text{MPBLE} = \frac{1}{\text{N}^{3D}_{\text{samples}} \cdot \text{N}_{\text{bones}}} \sum_{p=1}^{\text{N}^{3D}_{\text{samples}}} \sum_{b=1}^{\text{N}_{\text{bones}}} \left| \| \hat{\mathbf{B}}_{p,b} \|_2 - \| \mathbf{B}_{p,b} \|_2 \right|, \tag{4.14}$$

where $\hat{\mathbf{B}}_{p,b}$ and $\mathbf{B}_{p,b}$ are the predicted and ground truth bone vectors for the b-th bone in the p-th sample, respectively, $\text{N}^{3D}_{\text{samples}}$ is the total number of samples, and N_{bones} is the total number of bones in the skeleton.

This metric is particularly useful for ensuring anatomically plausible predictions, especially in applications where structural consistency is critical.

4.2.3 Evaluation Protocols

In this thesis, the evaluation protocol is applied separately to each dataset, following the standardized subject splits described above. Predictions are expressed in camera-centered coordinates and root-centered to remove global

translation. Performance is quantified by averaging the error between predicted and ground truth joint positions across all joints and frames. In addition to the raw positional error, some metrics incorporate scale normalization or rigid Procrustes alignment to assess pose accuracy independently of orientation or scale.

4.3 Skeleton-based Action Recognition

This section presents the experimental setup used to evaluate the proposed methods for SBAR. First, an overview of the datasets is provided in Section 4.3.1. The evaluation metrics are then discussed in Section 4.3.2, followed by the evaluation protocols in Section 4.3.3.

4.3.1 Datasets

SBAR datasets provide sequences of 3D joint coordinates over time, either extracted from RGB/depth videos using pose estimation methods or captured with motion capture systems. An overview of widely used SBAR dataset is provided in Table 4.9. Standard benchmarks such as NTU RGB+D [Sha16, Liu19], UCLA [Wan14], and J-HMDB [Jhu13] typically consist of controlled lab recordings or short clips from real-world videos. While these datasets are useful for evaluating general-purpose action recognition models, they are not representative of surveillance scenarios, which often involve long-duration monitoring, rare or anomalous events, and highly unstructured environments. Surveillance-focused datasets such as ShanghaiTech [Liu18], UCF-Crime [Sul18], and VFP290k [An21] (cf. [Gol22] for the latter; co-authored by the author of this thesis) are specifically designed for anomaly detection (binary classification) and are therefore not optimized for skeleton-based action recognition benchmarks.

In this work, the focus is on challenging, real-world human behavior rather than scripted or lab-controlled settings. To this end, Toyota Smarthome (Smarthome) [Das19] is used to evaluate indoor activities performed by

senior participants in unscripted daily routines, providing long-duration sequences with occlusions, variable viewpoints, and class imbalance. UAV-Human [Li21b] complements this with outdoor aerial recordings across urban and natural environments, featuring a large number of subjects and actions under dynamic camera motion, occlusion, and varying lighting conditions. These datasets capture the complexities of real-world behavior in a manner that standard lab-based skeleton datasets cannot, and are thus suitable for evaluating robust models for practical action recognition applications.

Table 4.9: Overview of skeleton-based action recognition datasets – Publicly available datasets for benchmarking 3D skeleton-based action recognition models.

Dataset	Environment	Subjects	Keypoints	Actions	Sequences	Cameras	Modality	Frames
NTU RGB+D [Sha16]	lab	40	25	60	56K	3	RGB+Depth+Infrared+Skeleton	4M
NTU RGB+D 120 [Liu19]	lab/outdoor	106	25	120	114K	3	RGB+Depth+Skeleton	8.2M
Toyota Smarthome [Das19]	home	18	17	31	16K	7	RGB+Depth+Skeleton	16K
UAV-Human [Li21b]	outdoor/urban	119	17	155	67K	UAV cameras	RGB+Depth+Infrared+Skeleton	22K
J-HMDB [Jhu13]	in-the-wild	21	16	21	928	1	RGB+Skeleton	9K
UCLA [Wan14]	lab	8	15	8	1K	3	RGB+Depth+Skeleton	24K

4.3.1.1 UAV-Human

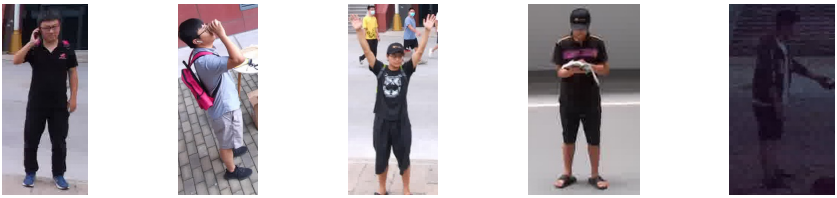


Figure 4.9: Example images from the UAV-Human dataset – The UAV-Human dataset is a large-scale benchmark for human action recognition from aerial viewpoints. The samples illustrate actions such as talking on the phone, drinking, waving hands, reading, and shaking hands at night, highlighting the dataset's challenging conditions with varying viewpoints, distances, and illumination.

The UAV-Human dataset [Li21b] is a large-scale benchmark for human behavior understanding captured by a flying UAV across diverse environments. It comprises 67,428 multimodal video sequences (RGB, depth, infrared, fisheye) featuring 119 subjects performing 155 action classes. It also includes 22,476 frames annotated for 17-joint 2D pose estimation, 41,290 frames for

person re-identification with 1,144 identities, and 22,263 frames for attribute recognition. Data were collected over three months across urban and rural settings under varying lighting, weather, occlusion, and UAV motion conditions, resulting in realistic challenge scenarios. Sample frames from the dataset are illustrated in Figure 4.9. For skeleton-based experiments, we rely on the released 2D keypoint annotations and the authors' official lists for action recognition unless stated otherwise.

4.3.1.2 Toyota Smarthome

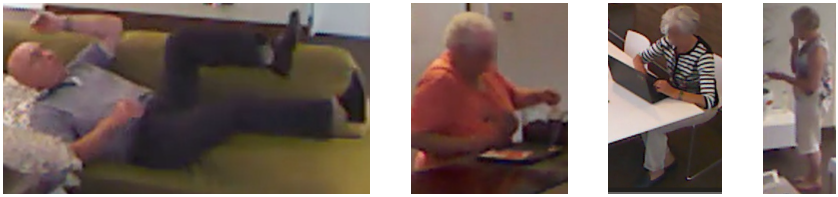


Figure 4.10: Example images from the Smarthome dataset – The Smarthome dataset captures daily living activities of elderly subjects in a real apartment setting. The samples depict actions such as laying down on the couch, making coffee, using a laptop, and taking medication, illustrating the dataset's focus on natural, unscripted behaviors in diverse indoor environments.

The Smarthome dataset [Das19] is a benchmark of daily living activities recorded in a furnished apartment equipped with 7 Kinect v1 cameras. It contains 16,115 video samples across 31 activity classes performed by 18 senior participants aged 60–80 years. Each subject was recorded for approximately eight hours in a single day, without a predefined script, thereby capturing natural daily routines such as drinking, cooking, or cleaning. For privacy, all faces were blurred using an automatic detection method. The dataset provides three modalities: RGB, depth, and 3D skeletons, where skeletons were extracted from RGB using LCR-Net [Rog19].

The dataset poses several challenges for action recognition, including large intra-class variability, class imbalance, occlusions, and variable subject–camera distances. Activity durations vary significantly, from only a few seconds (*e.g.* sit down) to several minutes (*e.g.* prepare meal). Fine-grained

sub-activity labels are also available for composite activities such as making coffee or cooking. Sample frames from the dataset are illustrated in Figure 4.10.

4.3.2 Evaluation Measures

Evaluation focuses on classification performance. The most commonly used metric is **Top-1 Accuracy (ACC)**, while **Mean Average Per-Class Accuracy (MAPCA)** is also reported, especially in class-imbalanced settings.

ACC: This metric measures the proportion of correctly classified samples based on the top predicted class:

$$\mathrm{ACC} = \frac{\mathrm{TP}_c}{\mathrm{N}_{\mathrm{samples}}}, \tag{4.15}$$

where TP_c denotes the number of correctly classified samples and $\mathrm{N}_{\mathrm{samples}}$ the total number of samples in the evaluation set. Top-1 accuracy is widely used as the primary metric for SBAR.

MAPCA: this metric computes the average accuracy across classes, mitigating class imbalance by averaging per-class performance:

$$\mathrm{MAPCA} = \frac{1}{C_{\mathrm{SBAR}}} \sum_{c=1}^{C_{\mathrm{SBAR}}} \frac{\mathrm{TP}_{cc}}{\mathrm{TP}_{cc} + \mathrm{FN}_{cc}}, \tag{4.16}$$

where C_{SBAR} is the number of classes, TP_{cc} the correctly classified samples of class c, and FN_{cc} the misclassified samples of class c.

In summary, Top-1 Accuracy (Acc@1) remains the primary metric for SBAR, while MAPCA is additionally reported when emphasizing robustness to class imbalance (as in Toyota Smarthome).

4.3.3 Evaluation Protocols

Two evaluation protocols are used for skeleton-based action recognition: *Cross-Subject* (CS), where training and test splits are disjoint in terms of individuals, and *Cross-View* (CV), designed to assess generalization across viewpoints. These protocols benchmark model robustness under subject and view variation.

UAV-Human. Two evaluation protocols are defined for skeleton-based action recognition: *Cross-Subject*, where training and test splits are disjoint in terms of individuals; and *Cross-View*, designed to assess generalization across viewpoints. These protocols are widely used to benchmark model robustness under subject and view variation.

Toyota Smarthome. Two official evaluation protocols are defined. The *Cross-Subject* (CS) protocol splits the 18 participants into disjoint training and test groups (11/7). The *Cross-View* (CV) protocol evaluates generalization across camera viewpoints, with CV1 training on camera 1 and testing on camera 2, and CV2 training on multiple cameras (1, 3, 4, 6, 7) while still testing on camera 2. In both cases, mean per-class accuracy is reported as the evaluation metric.

5 Baseline

This chapter formalizes and benchmarks the three core components of the pipeline: 2D-HPE, 3D-HPE, and SBAR. For each task, a concise problem formulation is given, followed by a strong, reproducible baseline to support subsequent evaluations and ablations. As the baseline for 2D-HPE, the heatmap-regression transformer VitPose by Xu et al. [Xu22] is adopted. Details are provided in Section 5.1. For 3D-HPE, the single-view 2D-to-3D lifting transformer named PoseFormer by Zheng et al. [Zhe21] is used to produce temporally coherent, root-centric 3D joints and joint orientations (see Section 5.2). For SBAR, the GCN called CTR-GCN by Chen et al. [Che21b] serves as the baseline classifier over 3D skeleton sequences (see Section 5.3). These baselines offer competitive accuracy, favorable runtime, and publicly available training protocols and pretrained weights, establishing a consistent foundation for comparisons and ablations in later chapters.

5.1 2D Human Pose Estimation

This section introduces the fundamental principle of 2D-HPE as the necessary step before 3D-HPE. The section begins with a formal description of the problem, highlighting the inputs, outputs, and goals of 2D-HPE in Section 5.1.1. Furthermore, a top-down method inspired by Xu et al. [Xu22] is described to establish a baseline for this thesis. Details are presented in Section 5.1.2.

5.1.1 Problem Formulation

Consider a dataset $\mathcal{D}_{2D}$ consisting of $\mathrm{N}^{2D}_{\text{samples}}$ cropped images of single persons and their corresponding 2D keypoint annotations:

$$\mathcal{D}_{2D} = \{(\mathbf{I}^r, \mathbf{K}^r) \mid r = 1, 2, \dots, \mathrm{N}^{2D}_{\text{samples}}\}, \tag{5.1}$$

where $\mathbf{I}^r \in \mathbb{R}^{H \times W \times C}$ represents the r-th cropped image, with height H, width W, and C color channels. The corresponding 2D keypoint annotation is

$$\mathbf{K}^r = \{\mathbf{k}^r_i \mid i = 1, \dots, \mathrm{N}_{\text{keypoints}}\}, \quad \mathbf{k}^r_i = (x^r_i, y^r_i), \tag{5.2}$$

where (x^r_i, y^r_i) are the pixel coordinates of the i-th keypoint in image $\mathbf{I}^r$.

The task of 2D Human Pose Estimation involves learning a mapping function f that predicts the set of 2D keypoints $\hat{\mathbf{K}}^r$ from the input image $\mathbf{I}^r$:

$$\hat{\mathbf{K}}^r = f(\mathbf{I}^r; \phi_{2D}), \tag{5.3}$$

where ϕ_{2D} denotes the learnable parameters.

In most top-down methods, the prediction is obtained from heatmaps. For each keypoint i, a heatmap $\mathbf{H}^r_i \in \mathbb{R}^{h \times w}$ is predicted. $h \times w$ are the height and width of the heatmap respectively, which are downsampled from $H \times W$ usually by a factor of 4. The 2D location $\hat{\mathbf{k}}^r_i = (\hat{x}^r_i, \hat{y}^r_i)$ is recovered as

$$(\hat{x}^r_i, \hat{y}^r_i) = \left(\frac{\hat{u}_i}{w} \cdot W, \frac{\hat{v}_i}{h} \cdot H\right), \quad (\hat{u}_i, \hat{v}_i) = \arg\max_{(u,v)} \mathbf{H}^r_i(u,v), \tag{5.4}$$

where (u,v) are the coordinates of the highest peak in the heatmap.

The model is trained to minimize the difference between the predicted keypoints $\hat{\mathbf{k}}^r_i$ and the ground-truth keypoints $\mathbf{k}^r_i$. After prediction heatmaps coordinates are converted back to image coordinate.

5.1.2 Baseline for 2D Human Pose Estimation

The base framework for all experiments regarding 2D-HPE in this thesis is ViTPose [Xu22]. Proposed as a simple Vision Transformer baseline for top-down 2D-HPE, ViTPose demonstrated through extensive scaling studies that pure Transformer backbones achieve accuracy on par with or above convolutional designs across standard benchmarks. More importantly, Vision Transformers exhibit stronger generalization across domains, which is particularly suitable for the surveillance-oriented and multimodal settings addressed in this work. In line with the common heatmap regression formulation of 2D-HPE, the adopted baseline follows the ViTPose architecture illustrated in Figure 5.1.

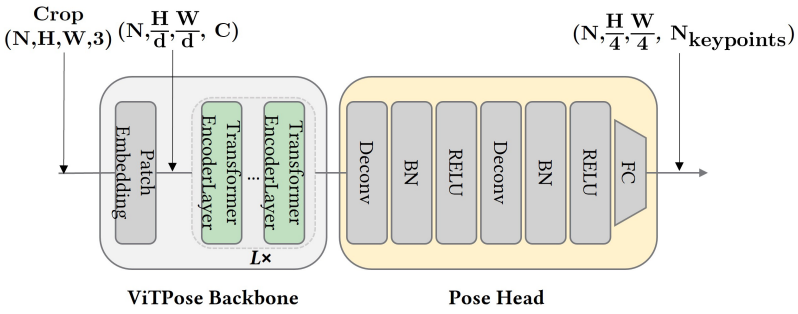


Figure 5.1: VitPose baseline architecture – A croppped person image is embedded into token via patch embedding layer with a downsampling ration of, *e.g.*, $d = 16$ by default. Then the embedded token are processed by several transformer layers. The backbone feature are then passed through a classic decoder composed of two deconvolution blocks and finally a fully connected layer, which delivers a localization heatmap for each keypoint.

A pure Vision Transformer backbone encodes global, spatial dependencies within person crops and a lightweight prediction head decodes per-joint heatmaps. 2D keypoints are obtained from heatmaps via argmax, with optional flip testing during inference. Compared to established CNN-based designs such as HRNet [Wan21], ViTPose replaces multi-branch convolutions

with self-attention while retaining a lightweight head. This design enables more effective modeling of long-range dependencies and yields stronger generalization, while maintaining a favorable accuracy–efficiency trade-off for deployment in surveillance scenarios.

5.1.2.1 Backbone

The development of backbone architectures for monocular 2D human pose estimation has evolved significantly. The framework introduced in [Xia18] demonstrated that generic image backbones combined with a lightweight heatmap regression head can yield strong results. ResNet backbones [He16] were commonly used, later extended by HRNet [Sun19, Wan21] with multi-resolution feature fusion. Hybrid architectures such as HRFormer [YUA21] and TransPose [Yan21b] integrated attention within CNN hierarchies, while hierarchical vision transformers such as PVTv2 [Wan22] introduced scalable Transformer backbones to dense prediction tasks. Preliminary work by the author [Cor21d] adapted PVTv2 by reducing the number of stages to achieve faster inference, showing the impact of architectural simplifications on runtime efficiency. The first prominent fully Transformer-based pose backbone is TokenPose [Li21c], which projects feature maps into a sequence of tokens before applying multi-head self-attention. However, more recent approaches have shifted towards pure Vision Transformer designs, following ViT [Dos20]. ViTPose [Xu22] directly adopts a ViT backbone for human pose estimation, accompanied by a systematic scaling study across backbone capacities. In contrast to hybrid models, ViTPose employs a pure and non-hierarchical Transformer encoder, retaining simplicity while achieving state-of-the-art results across benchmarks. In ViTPose, a cropped person image $I^r \in \mathbb{R}^{H \times W \times 3}$ is embedded into token via a patch embedding layer, *i.e.* $\mathbf{F} \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times C}$, with C is the channel dimension and d, *e.g.* 16 by default, is the downsampling ratio of the patch embedding layer. Then the embedded tokens are processed by several transformer layers, where each consists of a multi-head self-attention (MHSA) Layer and a feed-forward networtk (FFN).

The backbone outputs feature denoted $\mathbf{F_{out}} \in \mathbb{R}^{\frac{H}{d} \times \frac{W}{d} \times C}$ Backbone weights are initialized from ImageNet pretraining, transferring well-established visual representations to the pose estimation task.

Two backbone scales are employed: *ViTPose-Small*, a computationally efficient variant suitable for ablation studies, and *ViTPose-Huge*, providing maximum capacity and state-of-the-art accuracy.

5.1.2.2 Heatmap-Regression Head

The prediction head or decoder follows the heatmap regression paradigm, where each keypoint is represented by a spatial probability distribution rather than direct coordinates. This formulation provides dense supervision and captures spatial uncertainty.

In this work the classic decoder is used. It is composed of two deconvolution blocks, each of which contains one deconvolution layer followed by batch normalization and RELU. Each block upsamples the feature maps by 2 times. Then, a convolution layer with the kernel size 1×1 (FC Layer) is used to obtain the localization heatmaps for the keypoints, *i.e.*,

$$\mathbf{H} = \mathrm{Conv}_{1\times1}\big(\mathrm{Deconv}(\mathrm{Deconv}(\mathbf{F_{out}}))\big) \tag{5.5}$$

where $\mathbf{H} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times \mathrm{N_{keypoints}}}$ denotes the estimated heatmaps (one for each keypoint) and $\mathrm{N_{keypoints}}$ is the number of keypoints to be estimated.

During training, the network is supervised with Gaussian heatmaps centered at ground-truth keypoints. For keypoint $\mathbf{k}_s^r \in \mathbf{K}^r$, the target heatmap is

$$\mathbf{H}_i^{r,\mathrm{gt}}(u, v) = \exp\left(-\frac{(u - x_s^r)^2 + (v - y_s^r)^2}{2\sigma^2}\right), \tag{5.6}$$

where σ controls the spatial spread of the supervision signal.

At inference, 2D joint coordinates are extracted from the predicted heatmap $\mathbf{H}_i^r$ via arg max:

$$\hat{\mathbf{k}}_i^r = \arg\max_{(u,v)} \mathbf{H}_i^r(u,v), \quad \hat{\mathbf{k}}_i^r \in \hat{\mathbf{K}}^r. \tag{5.7}$$

5.1.2.3 Loss Function

The model is optimized using the mean squared error (MSE) loss between predicted and ground-truth heatmaps:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{\text{N}_{\text{keypoints}} \cdot \frac{H}{4} \cdot \frac{W}{4}} \sum_{i=1}^{\text{N}_{\text{keypoints}}} \sum_{u=1}^{\frac{H}{4}} \sum_{v=1}^{\frac{W}{4}} \left\| \mathbf{H}_i^r(u,v) - \mathbf{H}_i^{r,\text{gt}}(u,v) \right\|_2^2. \tag{5.8}$$

This dense supervision encourages the network to place maximal confidence near ground-truth locations while preserving uncertainty in ambiguous regions, such as under occlusion.

5.1.2.4 Implementation Details

All experiments are implemented using the MMPose toolbox [MMP20], which is built on the *PyTorch* deep learning framework [Pas19]. The AdamW optimizer [Los17] is employed in all settings with a weight decay of 0.1. However, due to historical and resource-related considerations, the training regimes differ between the UPAR-Pose and UPPET datasets.

For UPAR-Pose, an extensive training strategy is adopted. Models are trained for 100 epochs with AdamW and an initial learning rate of 1×10^{-3}. Batch sizes are set to 32 for ViTPose-S and reduced to 16 for ViTPose-H due to GPU memory constraints. The learning rate follows a two-stage schedule. First a linear warm-up from 1×10^{-6} to the base learning rate over the first 5 epochs is employed. Then a cosine annealing schedule is utilized for the remaining epochs, decaying the learning rate to a minimum of 1×10^{-5}.

For UPPET, a more conventional regime is applied, largely reflecting the original experimental configuration available at the time. Models are trained for 50 epochs with AdamW, an initial learning rate of 5×10^{-4}, and batch size 16. Step-wise learning rate decay is employed: at epoch 40, the learning rate is reduced by a factor of 0.1, with another reduction applied at epoch 45.

The two training regimes reflect differences in dataset characteristics, model scale, and computational constraints, resulting in distinct optimization schedules for UPAR-Pose and UPPET.

5.2 3D Human Pose Estimation

This section introduces the fundamental principle of 3D-HPE as the necessary step in the feature extraction approach for SBAR. The section begins with a formal description of the problem, highlighting the inputs, outputs, and goals of 3D-HPE in Section 5.2.1. Furthermore, a 2D-3D Uplifting method inspired by Zheng et al. [Zhe21] is described to establish a baseline for this thesis. Details are presented in Section 5.2.2.

5.2.1 Problem Formulation

Notation: For any temporal sequence S, $S^{1:T,i}$ denotes the T frames $\{A^{t,i}\}_{t=1}^{T}$ of sample i. For per-frame variables (e.g., $\mathbf{k}_p^t, \mathbf{J}_q^t$), the sample index i is omitted for readability.

Consider a dataset $\mathcal{D}_{3D}$ consisting of $\mathrm{N}_{\text{samples}}^{3D}$ temporal sequences of 2D keypoints and their corresponding 3D joint annotations:

$$\mathcal{D}_{3D} = \{(\mathbf{K}_{2D}^{1:T,i}, \mathbf{J}^{1:T,i}) \mid i = 1, 2, \dots, \mathrm{N}_{\text{samples}}^{3D}\}, \tag{5.9}$$

where $\mathbf{K}_{2D}^{1:T,i} \in \mathbb{R}^{T \times 2 \times \mathrm{N}_{\text{keypoints}}}$ represents a sequence of T frames of 2D keypoints for the i-th sequence, with each frame containing $\mathrm{N}_{\text{keypoints}}$ keypoints:

$$\mathbf{k}_p^t = (x_p^t, y_p^t), \quad p = 1, \dots, \mathrm{N}_{\text{keypoints}}. \tag{5.10}$$

Similarly, $\mathbf{J}^{1:T,i} \in \mathbb{R}^{T\times 3\times \mathrm{N}_{\text{joints}}}$ denotes the corresponding sequence of 3D joint positions, where each frame contains $\mathrm{N}_{\text{joints}}$ joints:

$$\mathbf{j}_q^t = (X_q^t, Y_q^t, Z_q^t), \quad q = 1, \dots, \mathrm{N}_{\text{joints}}. \tag{5.11}$$

The task of 3D Human Pose Estimation involves learning a mapping function g that predicts the 3D joint sequence $\hat{\mathbf{J}}^{1:T,i}$ from the input 2D keypoint sequence $\mathbf{K}_{2D}^{1:T,i}$:

$$\hat{\mathbf{J}}^{1:T,i} = g(\mathbf{K}_{2D}^{1:T,i}; \phi_{3D}), \tag{5.12}$$

where ϕ_{3D} are the learnable parameters of the model.

To simplify the task, the 3D joint positions are normalized using a root-relative coordinate system, where the root joint's 3D position is subtracted from all other joints. Let the root joint's 3D position for frame t be denoted as:

$$\mathbf{j}_{\text{root}}^t = (X_{\text{root}}^t, Y_{\text{root}}^t, Z_{\text{root}}^t) \tag{5.13}$$

The root-relative position of a joint is defined as:

$$\tilde{\mathbf{j}}_q^t = \mathbf{j}_q^t - \mathbf{j}_{\text{root}}^t = (X_q^t - X_{\text{root}}^t, Y_q^t - Y_{\text{root}}^t, Z_q^t - Z_{\text{root}}^t), \tag{5.14}$$

where $\tilde{\mathbf{j}}_q^t$ represents the position of the q-th joint relative to the root joint for frame t. The entire 3D pose for frame t in the root-relative coordinate system is then:

$$\tilde{\mathbf{J}}^t = \{\tilde{\mathbf{j}}_q^t \mid q = 1, \dots, \mathrm{N}_{\text{joints}}\}. \tag{5.15}$$

The dataset is transformed such that the 3D annotations $\mathbf{J}$ become root-relative positions $\tilde{\mathbf{J}}$. The model is trained to predict $\hat{\tilde{\mathbf{J}}}^{1:T,i}$, the root-relative 3D joints:

$$\hat{\tilde{\mathbf{J}}}^{1:T,i} = g(\mathbf{K}_{2D}^{1:T,i}; \phi_{3D}). \tag{5.16}$$

At inference time, the predicted root-relative 3D joints can be converted back to their absolute 3D positions if the global position of the root joint $\mathbf{J}_{\text{root}}{}^{t}$ is known:

$$\hat{\mathbf{j}}_q^t = \hat{\tilde{\mathbf{j}}}_q^t + \mathbf{j}_{\text{root}}^t. \tag{5.17}$$

The root-relative coordinate system provides several key benefits. By focusing on the relative structure of the pose rather than its absolute position, it simplifies the prediction task and the pose estimation is then invariant to global translations in the 3D space. Additionally, it aligns well with most datasets, where 3D poses are typically provided in a root-relative format. This normalization ensures that the model learns the spatial relationships between body joints effectively, which are critical for accurate 3D pose estimation. For surveillance-oriented SBAR, root-relative encoding is particularly advantageous, as it abstracts away camera placement and scene geometry while preserving motion and articulation cues most relevant for recognizing human actions under unconstrained conditions.

5.2.2 Baseline for 3D Human Pose Estimation

The base framework for all the experiments regarding 3D-HPE in this thesis is the work of Zheng et al. [Zhe21]. The authors were the first to introduce spatial and temporal transformers to 2D-to-3D Uplifting frameworks. The authors carried out a detailed study on several aspects of 3D-HPE transformer models and achieve comparable results to the current state-of-the-art in terms of this task. Since 3D-HPE is a regression task, the baseline model builds upon an uplifting architecture as illustrated in Figure 5.2. It consists of a pure transformer backbone based on the use of a spatial transformer module depicted in yellow, which encodes the local relationships between the 2D keypoints and a temporal transformer module depicted in green, which captures the global dependencies between frames regardless of their distance. The extracted features are then fed to a Fully-Connected (FC) layer to produce 3D coordinates for each joint in each input frame in a single forward

pass. This differs from [Zhe21], since in their work the authors only provide a prediction for the center frame.

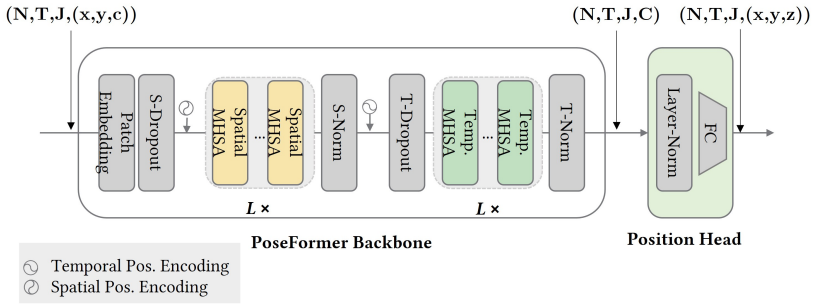


Figure 5.2: Baseline architecture – The backbone extracts feature maps for 2D keypoints sequences input. Each keypoint is a tuple consisting of (x,y,c) where c is the confidence score for the keypoint. This differs from [Zhe21], where only (x,y) are used. An other difference, is that the baseline in this work predicts 3D poses for all frames in the sequence, instead of only for the center frame. This is done with a simple MLP block using one Layer Normalization and one linear layer with output (x,y,z).

5.2.2.1 Backbone

The backbone model encodes rich features representing local and global relationships between the 2D input keypoints. To achieve this, Long Short-Term Memory (LSTM) networks have been proposed in [Hos18], Graph Neural Networks (GNNs) have been used for 2D-to-3D uplifting [Ci19, Cai19, Zha19b, Liu20a], and temporal convolutions are employed in [Mar17, Pav19b, Liu20b, Che21a]. Zheng et al. [Zhe21] introduced the first highly competitive convolution-free transformer network, marking the start of a recent series of pure transformer works for 2D-to-3D uplifting [Zha23b, Zha23a, Zhu23, Li22a, Li23, Zha22, Ein23, Liu23b, Li24, Pen24, Liu25b]. Although simpler than more recent models, it remains competitive in terms of prediction accuracy and runtime.

The baseline backbone model consists of two separate transformers for modeling spatial and temporal information, respectively. Both modules are implemented according to their definitions in [Zhe21]. Those basic modules have been later adapted and combined in different ways. Zhu et al. [Zhu23] propose a Dual-stream Spatio-temporal Transformer (DSTformer) which consists of dual-stream-fusion modules. Each module contains two branches of spatial or temporal MHSA and MLP. While improving accuracy, this comes at the cost of runtime. In an attempt to improve runtime, a frequency positional embedding is proposed in [Zha23b], whereas learnable upsampling tokens in addition to strided transformer are proposed in [Ein23]. The backbone baseline in this thesis is a strong compromise between accuracy and runtime.

Typically, the parameters of the backbone model are initialized with pretrained weights obtained by training the model on the AMASS Dataset [Mah19] for 3D-HPE. Thus, the backbone is already able to produce meaningful features and less data is required for fine-tuning the model. In this thesis, all models with their respective backbones and heads are first pretrained on AMASS as further highlighted in Section 5.2.2.4.

5.2.2.2 Regression Head

In contrast to [Zhe21], where only the center frame is predicted, this work predicts 3D poses for all frames in the sequence. The output of the temporal transformer module is directly processed. A simple MLP block with Layer Normalization and one linear layer is applied, producing the final output which represents the predicted 3D poses for all frames in the sequence.

5.2.2.3 Loss Function

The regression loss $\mathcal{L}_{3d}$ combines Mean Per Joint Position Error (MPJPE), as defined earlier in Equation (4.9), and the Normalized MPJPE (N-MPJPE), referred to as $\mathcal{L}_{\text{scale}}$, with a weighting factor λ_{scale}.

The N-MPJPE is a normalized version of MPJPE, which allows to align the scale of the prediction with the scale of the ground truth 3D pose. This typically allows to ignore differences in the size of the estimated skeleton.

The loss $\mathcal{L}_{3d}$ is then defined as:

$$\mathcal{L}_{3d} = \mathcal{L}_{\text{MPJPE}} + \lambda_{scale} \cdot \mathcal{L}_{\text{scale}}. \tag{5.18}$$

Here, λ_{scale} (*e.g.*, 0.5) is a hyperparameter that controls the contribution of N-MPJPE to the overall regression loss.

5.2.2.4 Implementation Details

The methodology presented in this section is implemented using the *PyTorch* [Pas19] deep learning framework. The PoseFormer backbone [Zhe21] comprises four spatial and four temporal transformers, as outlined in its reference implementation. For input frame selection, a sequence length of 27 frames with a stride of 9 is used, consistent with the configurations detailed in [Zhu23] and [Liu25b].

The model is trained with an initial learning rate of 1.25×10^{-4} and a batch size of 16, utilizing the AdamW optimizer [Los17] for parameter updates. To stabilize training and improve convergence, an exponential learning rate decay with a factor of 0.99 is applied, following the default setup in [Zhu23]. A weight decay parameter of 1×10^{-2} is employed in all experiments to ensure effective regularization and reduce overfitting.

Although the Transformer-based backbone allows flexibility in handling varying input lengths, the initial model is pretrained for 90 epochs on the AMASS dataset using the same input lengths as in subsequent experiments. This step ensures that potential pretraining artifacts do not influence downstream tasks. The AMASS dataset is aligned with the H36M body definition, and camera coordinates are transformed into pixel coordinates. Uncorrupted 2D skeleton keypoints are generated via orthographic projection. The pretraining process

incorporates 3D-specific losses, including the $\mathcal{L}_{3d}$ and a velocity loss, as described in [Zhu23]. Consistent with the reference implementation, data augmentations are applied: horizontal flipping is performed, 15% of keypoints are randomly zeroed, and noise is added to the inputs, sampled from a mixture of Gaussian and uniform distributions. This pretrained model serves as the foundation for all experiments conducted with the baseline.

5.3 Skeleton Based Action Recognition

This section introduces the fundamentals of SBAR, the final stage prior to notification generation. A formal problem statement in Section 5.3.1 defines the inputs, outputs, and objectives. The baseline classifier is CTR-GCN by Chen et al. [Che21b], a spatio-temporal graph convolutional network with channel-wise topology refinement that operates on sequences of root-centric 3D joints produced by the 3D-HPE baseline. Architecture and training details are provided in Section 5.3.2. The model outputs action posterior scores at sequence level.

5.3.1 Problem Formulation

The task of skeleton-based action recognition is defined as follows. A dataset consists of temporal sequences of 3D skeletons, each paired with an action label. Every skeleton sequence contains T frames, and each frame provides the 3D coordinates of a fixed set of body joints. The goal is to classify the entire sequence into one of the predefined action categories.

To capture both spatial structure and temporal dynamics, four complementary input modalities are derived from the joint positions:

- **Joints:** the raw 3D coordinates of all body joints,
- **Bones:** the relative vectors connecting joints to their parent joints,
- **Joint velocities:** the frame-to-frame changes of joint positions,
- **Bone velocities:** the frame-to-frame changes of bone vectors.

Each modality is processed by a dedicated classifier (or *stream*), producing action scores for the sequence. These scores are then fused, typically by averaging, to obtain the final prediction. In this way, the ensemble leverages complementary cues: joints encode absolute structure, bones emphasize part-level relations, and velocities capture motion dynamics.

The objective is to learn these stream-specific classifiers jointly such that, when combined, they can robustly recognize human actions from skeleton sequences in diverse and noisy surveillance settings.

5.3.2 Baseline for Skeleton Based Action Recognition

The base framework for all the experiments regarding SBAR in this thesis is the work of Chen et al. [Che21b] (CTR-GCN). The authors introduced channel-wise topology refinement into spatio-temporal graph convolutional networks, enabling more adaptive modeling of skeletal dynamics compared to fixed-topology GCNs. Their study demonstrated strong performance across several benchmark datasets, establishing CTR-GCN as a competitive and robust baseline for skeleton-based action recognition.

The baseline model builds upon a spatio-temporal GCN architecture. The model processes one type of kinematic feature at a time (*e.g.*, joints, bones, joint velocities, or bone velocities), which is encoded by a CTR-GCN backbone. Within each stream, the backbone encodes local spatial dependencies via graph convolutions with learnable channel-wise topology refinement, while temporal dynamics are modeled by 1D convolutions across frames. Stream-specific features are aggregated through global average pooling, and final action predictions are obtained by a linear classifier head. This differs from earlier GCN-based methods, which relied solely on fixed skeleton graphs and single-stream representations, limiting their ability to capture diverse motion cues.

5.3.2.1 Backbone

CTR-GCN backbone lies at the core of each stream. Each skeleton sequence is represented as input features $\mathbf{X} \in \mathbb{R}^{T \times \mathrm{N}_{\text{joints}} \times d}$, where T is the number of frames, $\mathrm{N}_{\text{joints}}$ is the number of joints, and d is the feature dimension per joint ($d = 3$ for raw 3D coordinates, but may increase after the first projection layer). A spatio-temporal graph convolutional layer combines a temporal convolution with a spatial graph convolution, where the adjacency matrix is refined in a channel-wise manner. Given input $\mathbf{X}$, the output of one CTR-GCN block is

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{A}_k \mathbf{X} \mathbf{W}_k,$$

where K is the number of partitions, $\mathbf{A}_k$ are normalized adjacency matrices with learnable channel-wise refinements, $\mathbf{W}_k \in \mathbb{R}^{d \times d'}$ are trainable weight matrices, and $\mathbf{Y} \in \mathbb{R}^{T \times \mathrm{N}_{\text{joints}} \times d'}$. Stacking multiple such layers progressively increases the receptive field in both space and time, allowing the model to capture higher-order spatial relations and long-term temporal dynamics.

5.3.2.2 Classifier Head

After L_{gcn} spatio-temporal blocks, the final feature tensor has shape $\mathbb{R}^{T \times \mathrm{N}_{\text{joints}} \times d'}$. Global average pooling is applied across time and joints to yield a fixed-dimensional representation:

$$\mathbf{h} = \frac{1}{T \cdot \mathrm{N}_{\text{joints}}} \sum_{t=1}^{T} \sum_{j=1}^{\mathrm{N}_{\text{joints}}} \mathbf{Y}_{t,j}. \tag{5.19}$$

The pooled vector $\mathbf{h} \in \mathbb{R}^{d'}$ is projected to class logits via a linear classifier:

$$\mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b}, \quad \mathbf{z} \in \mathbb{R}^{C_{\text{SBAR}}}. \tag{5.20}$$

The class posterior distribution is obtained via the softmax function:

$$\mathbf{p} = \text{softmax}(\mathbf{z}) \in \Delta^{C_{\text{SBAR}}}. \tag{5.21}$$

5.3.2.3 Loss Function

Training minimizes the cross-entropy loss between predicted posteriors and ground-truth action labels. For a single sample $(\mathbf{X}, y)$, the loss is

$$\mathcal{L}_{CE}(\mathbf{X}, y) = -\log \mathbf{p}_y, \tag{5.22}$$

where $\mathbf{p}_y$ denotes the predicted probability of the ground-truth class y.

5.3.2.4 Implementation Details

The methodology presented in this section is implemented using the *PyTorch* [Pas19] deep learning framework. Training is conducted for 300 epochs with a fixed temporal window length of 64 frames and a batch sizes of 32 for both training and evaluation. The stochastic gradient descent (SGD) optimizer is used with an initial learning rate set to 0.1, and weight decay of 4×10^{-4} is applied for regularization. A learning rate scheduler is used, where the learning rate is reduced by a factor of 0.5 whenever validation performance saturates for 5 consecutive epochs, with a minimum learning rate of 5×10^{-6}. Exponential Moving Average (EMA) of the model parameters is maintained during training to stabilize convergence and improve generalization.

6 2D Human Pose Estimation

In this thesis, the focus lies on top-down 2D-HPE, operating under the assumption that person detections, including tracks, are already available. In this chapter, an emphasis is placed on image-based processing aimed at extracting 2D keypoints in surveillance scenarios. The primary goal is to achieve accurate and robust 2D keypoint detection under diverse conditions, particularly in surveillance-specific settings. For this purpose, datasets tailored to the visible spectrum (UPAR-Pose, as described in Section 4.1.1.2) and thermal imaging (UPPET, as described in Section 4.1.1.4) have been contributed and are utilized. These datasets enable an evaluation of the generalization abilities of state-of-the-art models for both modalities, providing novel insights into their robustness under real-world conditions—an aspect previously not sufficiently addressed in the literature.

While the main focus lies on 2D-HPE, a multitask learning approach is proposed within the visible spectrum, where 2D-HPE is combined with PAR. This multitask configuration is explored as an auxiliary direction to investigate whether joint optimization may further enhance performance, support generalization, or improve resource efficiency. At the same time, it provides an opportunity to study potential trade-offs between spatially detailed localization and globally pooled semantic abstraction.

To this end, various design choices are systematically analyzed, including the comparison between single-task and multitask learning (Section 6.1), as well as a Data Augmentation (DA) strategy (Section 6.2). Through this analysis, a deeper understanding of the conditions that affect the accuracy and generalization abilities of 2D-HPE under surveillance-specific constraints is provided. Furthermore, the potential advantages of using a single backbone for related

tasks are investigated, highlighting possibilities for improving resource efficiency while maintaining focus on the core task of 2D-HPE.

6.1 Multitask Learning with PAR

The potential of multitask learning for 2D-HPE and PAR is investigated, motivated by both practical and scientific considerations. The research question focuses on whether a shared representation may be learned that simultaneously supports fine-grained spatial localization and global semantic classification, and whether such a representation would improve computational efficiency by avoiding separate models for each task [Cor26a]. The motivating hypothesis is that structural cues from human poses could inform attribute recognition, for instance, by constraining the plausible positions of clothing items or accessories (*e.g.*, pants aligned with leg keypoints, a backpack potentially shifting the position of core keypoints). In contrast, attribute information might provide additional context to refine pose estimation.

To explore these hypotheses, a multitask learning configuration is developed in which an early fusion strategy using a shared encoder is adopted, as illustrated in Figure 6.1. The shared encoder produces a joint representation that is subsequently processed by task-specific heads: a heatmap regression head for 2D-HPE and a global average pooling plus classification head for PAR. This configuration allows both tasks to be optimized jointly while requiring only a single forward pass, thereby reducing runtime and resource consumption.

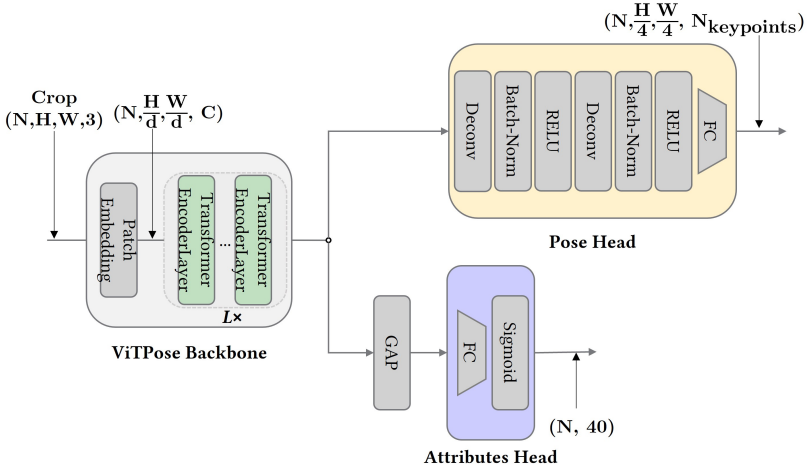


Figure 6.1: Multitask model – Multitask learning configuration with early fusion. A single encoder extracts a shared representation, which is subsequently processed by two task-specific heads: a heatmap regression head for 2D-HPE and a global average pooling plus classification head for PAR. This design enables joint optimization of both tasks while requiring only one forward pass, reducing computational cost and memory usage.

The results shown in Tables 6.1 and 6.2 indicate a clear asymmetry in task interactions. Attribute recognition consistently benefits from the shared representation, achieving notable improvements in mA and F1 scores across datasets. In contrast, 2D-HPE experiences a modest reduction in accuracy relative to its single-task baseline, particularly under cross-protocol evaluation. Despite this reduction, the observed AP values remain above 90% for most datasets, substantially higher than current benchmarks on COCO for equivalent backbone sizes.

This pattern reflects both a systematic advantage for PAR and a representational trade-off for 2D-HPE under multitask optimization. The key insights are that (i) a shared encoder can significantly enhance attribute recognition without severely compromising 2D pose estimation, (ii) pose estimation exhibits some sensitivity to domain shifts when trained jointly with other tasks,

and (iii) the interplay between structural (pose) and semantic (attribute) features provides a meaningful signal for multitask learning. Building on these observations, the next step explores whether targeted modifications to the shared encoder can mitigate the representational trade-offs. Specifically, the use of Task-specific Adapter (TSA) is investigated as a means to reduce interference between attribute recognition and pose estimation while preserving the benefits of multitask learning. In parallel, it is hypothesized that increasing the capacity of the shared encoder through a larger backbone could similarly alleviate task interference, potentially improving generalization across both attribute recognition and pose estimation.

Table 6.1: Specialization results on UPAR-Pose – Attribute recognition benefits from shared representations with consistent gains in mA and F1. In contrast, 2D-HPE losses accuracy compared to its baseline. Adding TSA reduces this drop, while still retaining most of the PAR improvements. **Red** highlights the best score per column, while **blue** indicates the second-best.

Approach	Market-1501				PA-24K			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE	**91.1**	**98.9**	–	–	**89.7**	**98.9**	–	–
Baseline PAR	–	–	69.8	80.8	–	–	80.8	78.8
Shared Encoder	89.4	**98.8**	**73.4**	**85.1**	85.6	**97.9**	**82.5**	**81.5**
Shared Encoder + TSA	**90.0**	**98.9**	**73.0**	**85.0**	**87.4**	**97.9**	**82.8**	**81.7**
Approach	PETA				UPAR-Pose			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE	**92.1**	**98.8**	–	–	**92.6**	**98.8**	–	–
Baseline PAR	–	–	83.9	87.5	–	–	82.6	84.2
Shared Encoder	**91.0**	**98.7**	84.2	88.9	**91.3**	**98.8**	**83.9**	**86.2**
Shared Encoder + TSA	90.9	**98.7**	**84.4**	**89.3**	**91.3**	**98.8**	**83.4**	**86.3**

Table 6.2: Generalization results on UPAR-Pose – Attribute recognition improves with shared representations, but adding TSA reduces mA and F1 across both protocols, with the drop less pronounced for F1. In contrast, 2D-HPE loses accuracy compared to its baseline, which is mitigated using TSA. **Red** highlights the best score per column, while **blue** indicates the second-best.

Approach	UPAR-Pose 3FCV			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE	**80.8** ± 4.1	**96.4** ± 1.0	–	–
Baseline PAR	–	–	66.4 ± 3.6	68.0 ± 5.4
Shared Encoder	77.1 ± 4.8	94.5 ± 0.6	**68.4** ± 4.2	**72.0** ± 5.8
Shared Encoder + TSA	**77.9** ± 4.9	**95.9** ± 1.5	**68.1** ± 4.2	**71.7** ± 5.6
Approach	UPAR-Pose LOOCV			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE	**86.2** ± 2.6	**97.7** ± 0.8	–	–
Baseline PAR	–	–	68.2 ± 1.2	**71.6** ± 5.0
Shared Encoder	81.4 ± 2.9	96.8 ± 1.4	**70.5** ± 1.3	**74.6** ± 4.7
Shared Encoder + TSA	**83.8** ± 3.0	**96.9** ± 1.3	**69.4** ± 1.7	**74.6** ± 4.6

6.1.1 Task-specific Adapters

The first hypothesis states that a small task-specific adaptation can reduce the representational conflict between tasks in a shared encoder. Introducing a single fully-connected layer at the start of each task head, as illustrated in Figure 6.2, is intended to reduce the capacity pull from the shared encoder towards PAR, while adding only a small number of parameters. As shown in Tables 6.1 and 6.2, this halves the HPE performance loss compared to the simple shared encoder, with a minor cost in PAR performance—which remains above the baseline. The results indicate that TSA successfully reduce representational conflict between tasks.

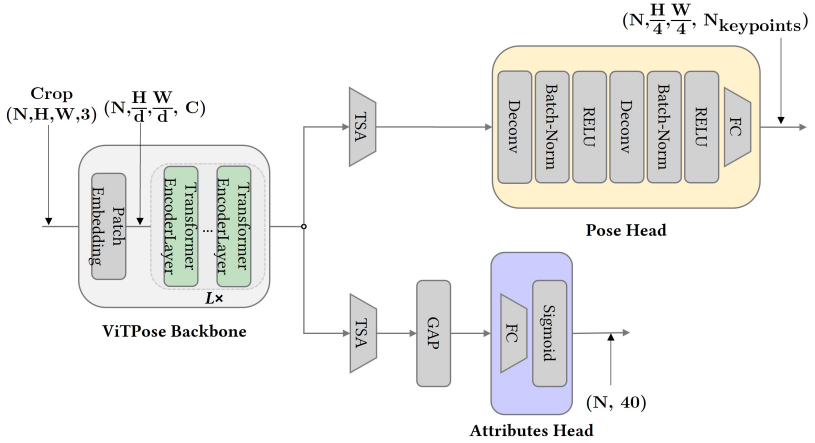


Figure 6.2: Shared Encoder with TSA – Multitask learning configuration with early fusion. A shared encoder extracts a common representation, which is first processed by lightweight fully connected adapters before being passed to the task-specific heads: a heatmap regression head for 2D-HPE and a global average pooling plus classification head for PAR. The adapters introduce task-dependent transformations while retaining the efficiency of a shared backbone.

6.1.2 Larger Backbone

The second hypothesis states that the VitPose-small backbone used as baseline may lack sufficient capacity to handle both tasks simultaneously. To test this, experiments with the largest VitPose backbone variant (Huge) were conducted.

As shown in Table 6.3, scaling up the backbone markedly improves both 2D-HPE and PAR baselines, indicating that model capacity constitutes a major limiting factor. In the multitask setting, the shared encoder with Huge backbone retains nearly all of the 2D-HPE accuracy of the single-task baseline, while attribute recognition benefits substantially from shared representations. Introducing TSA does not further improve 2D-HPE in this configuration and leads to a small reduction in PAR performance compared to the plain shared encoder.

These results suggest that increasing capacity alleviates representational conflict more effectively than architectural modifications, while the utility of TSA diminishes as the backbone size grows.

Table 6.3: Generalization results on UPAR-Pose – Scaling the backbone from Small to Huge improves both 2D-HPE and PAR baselines. Shared encoder models retain most of the HPE accuracy while providing clear gains in attribute recognition. Task-Specific Adapters (TSA) do not further benefit HPE in the Huge setting and slightly reduce PAR performance.**Red** highlights the best score per column, while **blue** indicates the second-best.

Approach	UPAR-Pose 3FCV			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE Small	**80.8** ± 4.1	**96.4** ± 1.0	–	–
Baseline PAR Small	–	–	66.4 ± 3.6	68.0 ± 5.4
Shared Encoder Small	77.1 ± 4.8	94.5 ± 0.6	**68.4** ± 4.2	**72.0** ± 5.8
Shared Encoder + TSA Small	**77.9** ± 4.9	**95.9** ± 1.5	**68.1** ± 4.2	**71.7** ± 5.6
Baseline HPE Huge	**89.6** ± 1.5	**98.4** ± 0.6	–	–
Baseline PAR Huge	–	–	71,4 ± 4.8	75.4 ± 5.2
Shared Encoder Huge	**88.5** ± 2.0	**98.4** ± 0.5	**72.6** ± 4.0	**77.2** ± 5.2
Shared Encoder + TSA Huge	**88.5** ± 2.0	**98.4** ± 0.6	**71.0** ± 4.9	**76.2** ± 6.2

6.2 Data Augmentation

In this section, the aim is to reduce overfitting using the common technique of DA. In a previous work [Spe23], the authors demonstrated that AugMix [Hen20] improves attribute recognition on the UPAR dataset. In contrast, in this work experiments are lead using the heavier augmentation procedures proposed in Sapiens [Khi24a], since they have already been validated for 2D-HPE and therefore provide a consistent setup for multitask training. Following the augmentation pipeline proposed in Sapiens, two complementary sets of transformations are applied: photometric distortions and structural perturbations.

6.2.1 Photometric Distortion

The photometric distortion pipeline from the MMpose toolbox is applied sequentially, with each transformation executed with a probability of 0.5. The random contrast adjustment is performed either as the second or as the last operation, ensuring variability in pixel intensity modifications. The complete sequence of operations is defined as follows:

- Random brightness adjustment
- Random contrast adjustment (mode 0)
- Conversion from BGR to HSV color space
- Random saturation adjustment
- Random hue adjustment
- Conversion from HSV to BGR color space
- Random contrast adjustment (mode 1)
- Random channel swapping

6.2.2 Structural Perturbations

Additional augmentations are performed using Albumentations [Bus20] operators. These transformations introduce blur and occlusion effects to simulate challenging visual conditions. The configuration includes:

- Gaussian blur with probability $p = 0.1$
- Median blur with probability $p = 0.1$
- Coarse dropout with one occlusion hole of variable size, covering up to 40% of image height and width ($p = 1.0$)

6.2.3 Evaluation

In the following, the DA pipeline is evaluated. The training setup mirrors the baseline, using equivalent parameters, optimizers, and training schemes to ensure a controlled comparison. Both specialization and generalization protocols are considered across multiple datasets, including UPAR-Pose and UPPET, with results summarized in Tables 6.4–6.8.

Under the specialization protocol, the introduction of DA generally results in slight performance losses for both PAR and 2D-HPE. For instance, on UPAR-Pose, baseline models augmented with DA show small reductions in mA and F1 for PAR, while 2D-HPE performance remains largely stable (Table 6.4). The shared encoder variant shows a notable exception: on the PA-24K dataset, DA slightly improves PAR performance, suggesting that the added variability can reduce overfitting in certain configurations. Similar trends are observed on UPPET, where small models experience marginal gains for 2D-HPE on CAMEL-P, whereas other variants generally experience minor performance reductions (Table 6.7).

In contrast, under generalization protocols, DA consistently enhances robustness to domain shifts. Across both 3FCV and LOOCV protocols on UPAR-Pose, 2D-HPE benefits are observed primarily in AP_{50}, while PAR shows modest gains in F1, despite occasional slight declines in mA (Table 6.5). Similar effects are observed on UPPET, where both small and huge models demonstrate improved generalization performance, particularly for 2D-HPE (AP and AP_{50}) (Table 6.8). These results indicate that DA effectively increases model robustness to unseen domains, although it may slightly disturb fine-grained specialization.

The introduction of TSAs shows complementary effects. When combined with the shared encoder, TSA mitigates some of the reductions in 2D-HPE observed under specialization while preserving most of the PAR improvements. Under generalization, TSA maintains competitive performance, suggesting that targeted adaptations can reduce the representational conflict between tasks in multitask settings. Overall, the evaluation highlights a consistent trade-off: DA tends to reduce within-domain specialization while enhancing

cross-domain generalization, and TSA can help balance task-specific accuracy with shared representation benefits.

Table 6.4: DA specialization results on UPAR-Pose – Variants with DA incur consequent losses, except for the shared encoder on PA-24K, which shows a slight improvement. **Red** highlights the best score per column, while **blue** indicates the second-best.

Approach	Market-1501				PA-24K			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE	**91.1**	**98.9**	–	–	**89.7**	**98.9**	–	–
Baseline HPE + DA	**90.8**	**98.9**	–	–	**89.1**	**98.9**	–	–
Baseline PAR	–	–	69.8	80.8	–	–	80.8	78.8
Baseline PAR + DA	–	–	70.1	82.4	–	–	79.4	79.2
Shared Encoder	89.4	**98.8**	**73.4**	**85.1**	85.6	**97.9**	**82.5**	**81.5**
Shared Encoder + DA	89.3	**98.8**	72.4	84.9	87.1	**98.9**	80.4	81.0
Shared Encoder + TSA	90.0	**98.9**	**73.0**	**85.0**	87.4	**97.9**	**82.8**	**81.7**
Shared Encoder + TSA + DA	90.1	**98.9**	72.6	84.9	87.1	**98.9**	80.4	81.0
Approach	PETA				UPAR-Pose			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE	**92.1**	**98.8**	–	–	**92.6**	**98.8**	–	–
Baseline HPE + DA	**91.0**	**98.7**	–	–	**92.0**	**98.8**	–	–
Baseline PAR	–	–	83.9	87.5	–	–	82.6	84.2
Baseline PAR + DA	–	–	81.1	86.7	–	–	81.5	84.5
Shared Encoder	**91.0**	**98.8**	**84.2**	**88.9**	91.3	**98.8**	**83.9**	**86.2**
Shared Encoder + DA	90.0	**98.7**	80.6	86.8	91.0	**98.8**	82.0	85.9
Shared Encoder + TSA	90.9	**98.7**	**84.4**	**89.3**	91.3	**98.8**	**83.4**	**86.3**
Shared Encoder + TSA + DA	90.0	**98.7**	80.6	86.8	91.0	**98.8**	82.3	86.0

Table 6.5: DA generalization results on UPAR-Pose – Across both 3FCV and LOOCV protocols, DA yields minor gains for 2D-HPE (AP_{50}) and PAR (F1), while mA may slightly decrease. Similar trends are observed for shared encoder and TSA variants, with occasional slight losses in AP. **Red** highlights the best score per column, while **blue** indicates the second-best.

Approach	UPAR-Pose 3FCV			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE	**80.8** ± 4.1	**96.4** ± 1.0	–	–
Baseline HPE + DA	**80.8** ± 3.9	**96.8** ± 0.6	–	–
Baseline PAR	–	–	66.4 ± 3.6	68.0 ± 5.4
Baseline PAR + DA	–	–	65.0 ± 4.0	70.0 ± 5.1
Shared Encoder	77.1 ± 4.8	94.5 ± 0.6	**68.4** ± 4.2	**72.0** ± 5.8
Shared Encoder + DA	78.0 ± 4.2	**96.4** ± 1.0	66.9 ± 4.2	**72.1** ± 4.8
Shared Encoder + TSA	77.9 ± 4.9	95.9 ± 1.5	**68.1** ± 4.2	71.7 ± 5.6
Shared Encoder + TSA + DA	**78.9** ± 4.1	**96.4** ± 1.0	66.3 ± 3.9	**72.1** ± 4.9
Approach	UPAR-Pose LOOCV			
	AP ↑	AP_{50} ↑	mA ↑	F1 ↑
Baseline HPE	**86.2** ± 2.6	**97.7** ± 0.8	–	–
Baseline HPE + DA	**86.0** ± 2.5	**97.7** ± 0.8	–	–
Baseline PAR	–	–	68.2 ± 1.2	71.6 ± 5.0
Baseline PAR + DA	–	–	67.9 ± 2.1	72.6 ± 5.3
Shared Encoder	81.4 ± 2.9	96.8 ± 1.4	**70.5** ± 1.3	**74.6** ± 4.7
Shared Encoder + DA	83.0 ± 1.8	**97.6** ± 0.9	69.3 ± 1.8	**74.9** ± 4.5
Shared Encoder + TSA	83.8 ± 3.0	96.9 ± 1.3	**69.4** ± 1.7	**74.6** ± 4.6
Shared Encoder + TSA + DA	84.5 ± 2.5	97.3 ± 1.1	68.7 ± 2.0	**74.9** ± 4.5

As shown in Table 6.6, under 3FCV on UPAR-Pose, DA leaves the AP of the Small model unchanged while reducing variability and improving AP_{50}, whereas for the Huge model it increases AP with AP_{50} remaining stable and slightly reduced variability. Across scales, the Huge model consistently outperforms the Small model on both AP and AP_{50}, and augmentation systematically contracts cross-fold variance, supporting enhanced generalization.

Table 6.6: DA generalization results on UPAR-Pose for different model scales – Data DAmentation consistently improves generalization under the 3FCV protocol for both small and huge models. **Bold** indicates the best.

Approach	UPAR-Pose 3FCV	
	AP ↑	AP_{50} ↑
Baseline HPE Small	**80.8** ± 4.1	96.4 ± 1.0
Baseline HPE Small + DA	**80.8** ± 3.9	**96.8** ± 0.6
Baseline HPE Huge	89.6 ± 1.5	**98.4** ± 0.6
Baseline HPE Huge + DA	**90.1** ± 1.3	**98.4** ± 0.5

A similar pattern is observed on the thermal dataset UPPET. Under specialization, DA improves HPE on CAMEL-P for both small and huge backbones, and yields marginal gains for the small model on LLVIP-P. However, for other datasets (OTP, TPE, and LLVIP-P with the huge model), DA generally leads to reduced performance. By contrast, generalization experiments show consistent improvements across both model sizes and evaluation protocols, with the sole exception of a slight drop in AP_{50} under the LOOCV protocol. This confirms that the regularization effect of DA is particularly beneficial when models are exposed to domain shifts, while it may hinder optimization when training and testing remain within the same domain.

Table 6.7: DA specialization results on UPPET – On CAMEL-P, data DAmentation improves HPE for both small and huge models. For LLVIP-P, small model see only slight gains, while other variants generally experience performance loss. **Bold** indicates the best.

Method	LLVIP-P		OTP		CAMEL-P		TPE		UPPET	
	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑
Baseline HPE Small	84.6	**97.9**	**79.4**	**94.4**	74.9	94.8	**75.5**	**91.7**	**79.8**	**94.7**
Baseline HPE Small + DA	**84.9**	**97.9**	78.0	93.1	**75.4**	**94.9**	74.9	**91.7**	79.6	**94.7**
Baseline HPE Huge	**92.0**	**99.0**	**90.8**	**97.9**	80.3	96.9	**81.4**	**94.7**	**85.3**	**96.9**
Baseline HPE Huge + DA	91.9	99.0	90.3	97.6	**80.9**	**97.0**	81.1	93.7	84.6	**96.9**

Table 6.8: DA generalization results on UPPET – Data DAmentation consistently improves generalization for both small and huge models across protocols, with the exception of a slight drop in AP_{50} on the LOOCV protocol. **Bold** indicates the best.

Approach	UPPET 4FCV		UPPET LOOCV	
	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑
Baseline HPE Small	39.1 ± 5.4	65.7 ± 5.3	49.3 ± 14.1	**75.9** ± 13.1
Baseline HPE Small + DA	**42.5** ± 3.4	**70.5** ± 2.0	**50.9** ± 16.3	75.5 ± 13.4
Baseline HPE Huge	57.8 ± 4.5	79.5 ± 4.3	68.3 ± 13.4	85.4 ± 12.1
Baseline HPE Huge + DA	**61.9** ± 3.4	**82.3** ± 3.1	**71.1** ± 11.5	**87.6** ± 9.3

Overall, these findings underline a systematic trade-off: DA sacrifices a degree of within-domain specialization performance but improves cross-domain robustness. This trade-off is observed consistently across both RGB (UPAR-Pose) and thermal (UPPET) datasets, highlighting DA as a practical strategy for enhancing generalization in multi-domain human analysis tasks.

6.3 Summary

The ablation studies presented above reveal complementary insights into how different strategies affect the balance between PAR and HPE in multitask learning.

The first hypothesis addressed representational conflict between tasks. TSAs effectively reduce the pull of the shared encoder towards PAR in the small backbone setting, halving the HPE performance loss while maintaining improvements in PAR. In contrast, for the Huge backbone, TSAs provide no additional benefit for HPE and slightly reduce PAR accuracy, suggesting that their utility diminishes as capacity increases. Lightweight adapters therefore represent a practical means of mitigating interference in low-capacity regimes, but appear redundant once sufficient representational space is available.

The second hypothesis concerned model capacity. Scaling the backbone from Small to Huge substantially improves HPE performance, narrowing the gap to the single-task baseline while retaining benefits for PAR. This indicates

that limited capacity is a central bottleneck in multitask settings. However, the computational overhead associated with large backbones is considerable, highlighting a trade-off between efficiency and accuracy.

Finally, DA was introduced as a regularization strategy, following prior work on UPAR and recent augmentation pipelines for HPE. The results show that DA consistently improves cross-domain generalization on both UPAR-Pose and UPPET, with gains observed for both small and large backbones. At the same time, DA tends to reduce specialization accuracy, particularly in datasets with strong domain-specific biases such as UPPET-specialization. This confirms DA as an effective tool for robustness under domain shift, albeit at the cost of reduced peak performance on in-domain data.

In conclusion, a shared representation can indeed support both fine-grained spatial localization and global semantic classification, however, only under certain conditions. When model capacity is limited, representational conflict leads to a drop in pose estimation performance, which may be alleviated either by lightweight adapters or by expanding backbone capacity. Attribute recognition consistently benefits from shared features, confirming that structural pose cues inform semantic prediction. Conversely, improvements in pose estimation from attribute information remain modest, suggesting that the flow of complementary information is asymmetrical.

These insights highlight three complementary levers in multitask optimization: (i) TSAs reduce task conflict in low-capacity models, (ii) larger backbones expand representational capacity, and (iii) DA enhances robustness under domain shifts. The results suggest that an effective multitask strategy requires balancing these dimensions depending on whether specialization or generalization is prioritized.

While this chapter focused on single-image analysis of humans for PAR and 2D-HPE, many applications require reasoning over temporal sequences and lifting into 3D. The next chapter therefore extends this investigation beyond static images, examining how sequence-based modeling and 2D–3D uplift address challenges of motion, temporal consistency, and 3D structure.

7 3D Human Pose Estimation

This thesis focuses on root-relative 3D-HPE, which offers practical advantages for real-world scenarios by removing the dependency on absolute 3D information, such as camera parameters. A human-centric approach ensures applicability in unconstrained environments, where camera calibration or depth data is often unavailable. Root-relative coordinates ensure robustness and generalizability across diverse camera setups and viewpoints. Challenges associated with 3D pose estimation are addressed by introducing domain-specific constraints and extending the prediction framework to include joint orientations. Root-relative 3D pose prediction relies on the quality of 2D keypoints detected in the camera coordinate system. However, without constraints, uplifted 3D poses may lack anatomical plausibility or kinematic consistency. To address this, a framework is proposed that enforces anatomical and kinematic constraints using additional losses while incorporating a module for precise joint orientation prediction [Cor26b]. Anatomical constraints, discussed in Section 7.1.1, ensure biologically plausible outputs by maintaining consistent bone lengths and limiting joint positions to realistic ranges. Kinematic constraints, examined in Section 7.1.2, utilize structural relationships between joints to improve prediction stability and reduce ambiguity, particularly during dynamic motion sequences. Joint orientation prediction, described in Section 7.2, employs quaternions to represent rotations ensuring computational efficiency.

7.1 Constraints for 3D Joint Prediction

This section formalizes the constraints used to guide root-relative 3D joint predictions toward anatomically plausible and temporally consistent poses. The

constraints are expressed as differentiable penalties that are model-agnostic, enabling combination with any regressor that uplifts 2D keypoints to 3D.

Two complementary families of constraints are introduced. Anatomical constraints (Section 7.1.1) act on the skeletal structure within a frame and promote realistic body proportions by encouraging consistent bone lengths. In practice, to accommodate measurement and annotation noise, these constraints are implemented as soft regularizers that favor temporal smoothness and low variance of bone lengths over hard equality. Kinematic constraints (Section 7.1.2) act across time and penalize implausible motion patterns. By leveraging first- and second-order temporal differences of joint positions (flow, velocity, and acceleration), they suppress frame-to-frame jitter, improve motion smoothness, and help resolve monocular depth ambiguities. The following subsections detail each family of constraints, and their impact is quantified in the subsequent evaluation and ablation study.

7.1.1 Anatomical Constraints

The intuition in this section is that bone lengths in a video should remain consistent over time [Che21a, Hsu24]. However, in this work, it is shown that GT bone lengths from motion capture are noisy in Figure 7.1. The figure show the oscillation of bone length over time for the right thigh bone GT. While the bone length is expected to remain constant, the Ground Truth (GT) data shows oscillations between 506 mm and 518 mm. This highlights inconsistencies in the GT annotations, which could impact downstream tasks that rely on stable anatomical measurements.

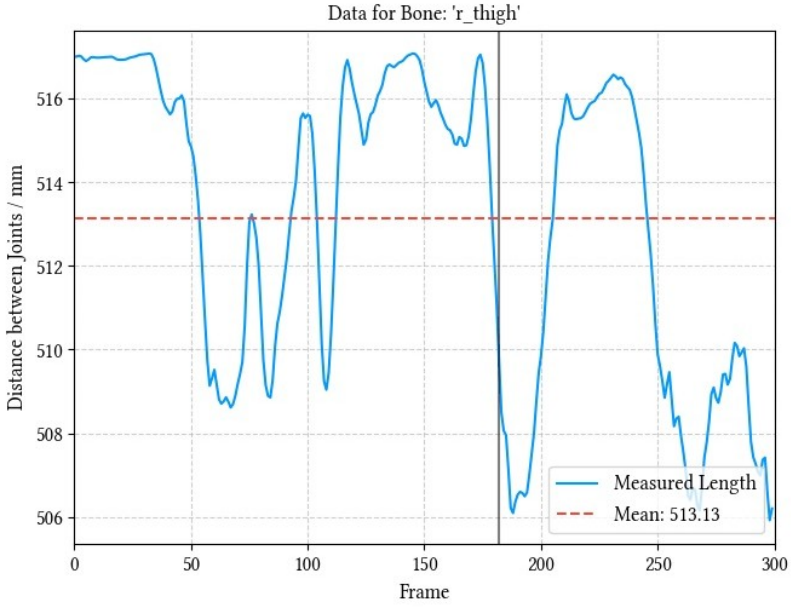


Figure 7.1: Bone length oscillation over frames – The figure shows the GT distance between joints for the right thigh bone over time. Although this bone length is expected to remain constant, it oscillates between 506 mm and 518 mm, indicating inconsistencies in the ground truth data of [Yeu25].

Thus, two losses are proposed and combined into a bone consistency loss $\mathcal{L}_{\text{B}}$, to account for this noise: (1) the $\mathcal{L}_{\text{smooth}}$ minimizes abrupt bone length changes across frames and (2) the $\mathcal{L}_{\text{var}}$ ensures bone lengths remain constant throughout the sequence. The losses are defined as follows:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{T-1} \sum_{t=2}^{T} \sum_{b=1}^{\text{N}_{\text{bones}}} \left| \text{length}(b,t) - \text{length}(b,t-1) \right|, \tag{7.1}$$

$$\mathcal{L}_{\text{var}} = \frac{1}{\text{N}_{\text{bones}}} \sum_{b=1}^{\text{N}_{\text{bones}}} \text{std}_{t=1}^{T}\big(\text{length}(b,t)\big), \tag{7.2}$$

$$\mathcal{L}_{\text{B}} = \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{var}}, \tag{7.3}$$

where: N_{bones} is the number of bones, f the number of frames, and $\text{length}(b,t)$, the length of bone b in the frame t, respectively.

7.1.2 Kinematic Constraints

Kinematic constraints regularize temporal evolution of root-relative 3D joints to reduce jitter and depth ambiguity while preserving dynamic motion. Joint kinematic features are computed within a short temporal context, following [Jin23], but extended to 3D joints rather than 2D keypoints. Let $\hat{\mathbf{J}}_{t,j}, \mathbf{J}_{t,j} \in \mathbb{R}^3$ denote the predicted and GT root-relative positions of joint j at frame t, and let T be the sequence length. The losses below are supervised and compare temporal derivatives of predictions and GT using Mean Squared Error (MSE).

7.1.2.1 Flow

Using flow as a kinematic loss enforces that predicted 3D joint displacements remain consistent with the underlying 2D motion observed in the image sequence. This ties the uplifted pose dynamics to actual pixel-level movement, reducing temporal drift and producing anatomically more plausible motion.

Δ-frame joint displacement is matched between predictions and GT:

$$\mathcal{L}_{\text{flow}} = \frac{1}{(T-\Delta)\,\mathrm{N}_{\text{joints}}} \sum_{t=\Delta+1}^{T} \sum_{j=1}^{\mathrm{N}_{\text{joints}}} \left\| \left(\hat{\mathbf{J}}_{t,j} - \hat{\mathbf{J}}_{t-\Delta,j}\right) - \left(\mathbf{J}_{t,j} - \mathbf{J}_{t-\Delta,j}\right) \right\|_2^2, \quad \Delta \in \mathbb{N},\, \Delta \geq 1, \tag{7.4}$$

with $\Delta = 2$ in implementation. See Figure 7.2 (right) for a visualization of joint-wise Δ-frame displacements. The corresponding RGB frame is shown on the left. Using $\Delta > 1$ aggregates motion over a wider temporal baseline, improving robustness to single-frame outliers and missing detections, and acting as a coarse prior for rapid movements.

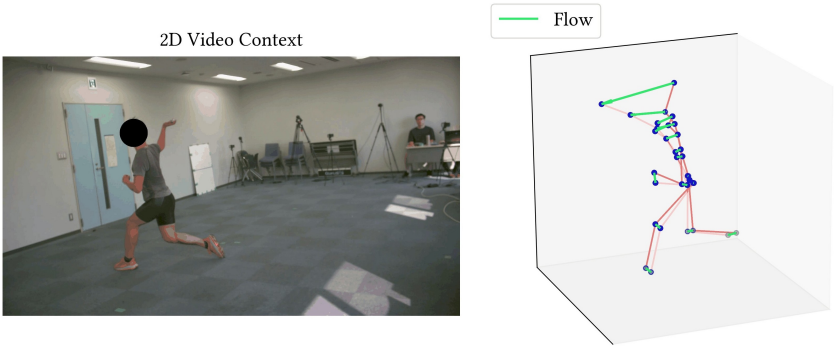


Figure 7.2: Flow (Δ-frame displacement) visualization — Left: input RGB frame. Right: joint-wise $\Delta = 2$ displacements on the root-relative 3D skeleton.

7.1.2.2 Velocity

Incorporating velocity as a kinematic loss ensures that predicted 3D joints evolve smoothly over time, discouraging abrupt or implausible changes. By aligning temporal differences in joint positions with realistic motion patterns, the model learns more stable and physically consistent trajectories.

First-order temporal differences are matched to suppress frame-to-frame jitter:

$$\mathcal{L}_{\text{vel}} = \frac{1}{(T-1)\,\mathrm{N}_{\text{joints}}} \sum_{t=2}^{T} \sum_{j=1}^{\mathrm{N}_{\text{joints}}} \left\| \left(\hat{\mathbf{J}}_{t,j} - \hat{\mathbf{J}}_{t-1,j}\right) - \left(\mathbf{J}_{t,j} - \mathbf{J}_{t-1,j}\right) \right\|_2^2. \quad (7.5)$$

Computed in the root-relative frame, this term targets high-frequency noise while factoring out global camera motion, and matches both speed and direction of motion. See Figure 7.3 (right) for joint-wise velocities. The RGB input is shown on the left.

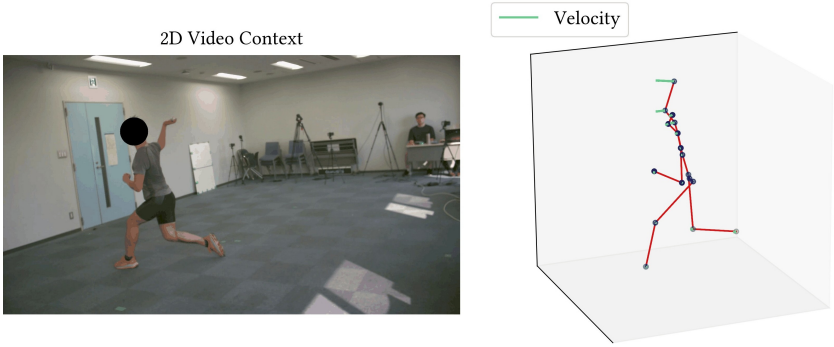


Figure 7.3: Velocity (first-order difference) visualization — Left: input RGB frame. Right: joint-wise velocities of root-relative 3D joints between consecutive frames.

7.1.2.3 Acceleration

Second-order temporal differences are matched to encourage approximately constant-velocity segments and reduce jerk, as shown in Figure 7.4:

$$\mathcal{L}_{\text{accel}} = \frac{1}{(T-2)\,\mathrm{N}_{\text{joints}}} \sum_{t=2}^{T-1} \sum_{j=1}^{\mathrm{N}_{\text{joints}}} \Big\| \left(\hat{\mathbf{J}}_{t+1,j} - 2\hat{\mathbf{J}}_{t,j} + \hat{\mathbf{J}}_{t-1,j}\right) - \left(\mathbf{J}_{t+1,j} - 2\mathbf{J}_{t,j} + \mathbf{J}_{t-1,j}\right) \Big\|_2^2. \quad (7.6)$$

This constraint complements the velocity term by discouraging sudden changes in speed while retaining sustained high-velocity segments. See Figure 7.4 for an example of per-joint acceleration magnitudes.

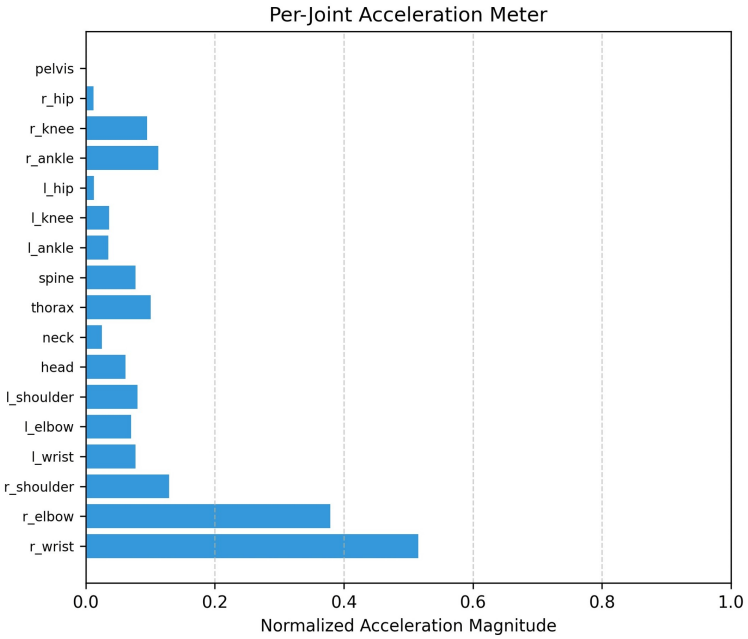


Figure 7.4: Acceleration (second-order difference) visualization — Per-joint acceleration magnitudes (second-order temporal differences) shown as a bar chart.

7.1.2.4 Combined Kinematics Loss

The proposed Losses are finally combined and weighted.

The overall regularizer is defined as

$$\mathcal{L}_{\mathrm{K}} = \lambda_{\mathrm{flow}}\,\mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{vel}}\,\mathcal{L}_{\mathrm{vel}} + \lambda_{\mathrm{accel}}\,\mathcal{L}_{\mathrm{accel}}. \tag{7.7}$$

Two variants are considered for evaluation: $\mathcal{L}_{K_{v1}}$, which uses flow and velocity only, similar to [Pen24], with $\lambda_{\text{flow}} = 5$, $\lambda_{\text{vel}} = 20$, $\lambda_{\text{accel}} = 0$, and $\mathcal{L}_{K_{v2}}$, which additionally incorporates acceleration with $\lambda_{\text{flow}} = 5$, $\lambda_{\text{vel}} = 20$, $\lambda_{\text{accel}} = 2.5$. In implementation, MSE with reduction over batch, time, joints, and coordinates is used. Normalization differences can be absorbed into the weights λ.

7.1.3 Evaluation

The impact of anatomical and kinematic constraints is evaluated on four datasets (Fit3D, H36M, AP3D, H4D) using MPJPE, P-MPJPE, MPJVE, and MPBLE. The results of the ablation study are summarized in Table 7.1, with best (in red) and second-best (in blue) values highlighted. The main focus is on reducing temporal jitter (MPJVE), which is crucial for producing stable skeleton sequences suitable for SBAR, while also maintaining or improving pose accuracy and anatomical consistency.

Table 7.1: Ablation of anatomical and kinematic constraints on four datasets (Fit3D, H36M, AP3D, H4D). B denotes the bone-length consistency term from Section 7.1.1; $\mathcal{L}_{K_{v1}}$ and $\mathcal{L}_{K_{v2}}$ denote the kinematics losses from Section 7.1.2. Best and second best are highlighted in red and blue respectively.

Method	Fit3D				H36M			
	MPJPE	P-MPJPE	MPJVE	MPBLE	MPJPE	P-MPJPE	MPJVE	MPBLE
Baseline	25.2	19.2	1.7	2.5	47.9	38.8	2.9	11.2
Baseline + $\mathcal{L}_{\text{B}}$	25.3	19.2	1.7	**2.2**	47.9	38.8	**2.8**	11.2
Baseline + $\mathcal{L}_{K_{v1}}$	**24.7**	**18.8**	**1.5**	2.4	**47.5**	**38.7**	**2.1**	**11.1**
Baseline + $\mathcal{L}_{K_{v2}}$	**24.7**	**18.8**	**1.4**	**2.3**	**47.5**	**38.6**	**2.1**	**11.0**
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v1}}$	**24.8**	**18.7**	**1.5**	2.4	**47.3**	**38.6**	**2.1**	**11.0**
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v2}}$	25.1	**18.8**	**1.4**	2.5	**47.3**	**38.6**	**2.1**	**11.0**
Method	AP3D				H4D			
	MPJPE	P-MPJPE	MPJVE	MPBLE	MPJPE	P-MPJPE	MPJVE	MPBLE
Baseline	**16.3**	12.6	**2.4**	**3.8**	64.8	34.6	10.7	3.5
Baseline + $\mathcal{L}_{\text{B}}$	16.4	12.6	**2.4**	3.9	64.3	34.7	10.7	3.9
Baseline + $\mathcal{L}_{K_{v1}}$	**15.4**	**11.8**	**1.8**	**3.6**	**58.4**	**30.5**	**7.8**	3.7
Baseline + $\mathcal{L}_{K_{v2}}$	**15.4**	**11.8**	**1.8**	**3.6**	**58.6**	**30.5**	**7.8**	3.6
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v1}}$	**15.4**	**11.8**	**1.8**	**3.6**	59.0	**30.5**	**7.8**	**3.1**
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v2}}$	**15.4**	**11.8**	**1.8**	**3.6**	58.9	**30.5**	**7.8**	**3.5**

Kinematic Constraints: Adding kinematic losses ($\mathcal{L}_{K_{v1}}$ and $\mathcal{L}_{K_{v2}}$) consistently lowers MPJVE across all datasets. On Fit3D, MPJVE decreases from 1.7 in the baseline to 1.5–1.4 with kinematic regularization. On H36M, it drops from 2.9 to 2.1, on AP3D from 2.4 to 1.8, and on H4D from 10.7 to 7.8. These improvements in temporal stability are accompanied by better pose accuracy: MPJPE decreases from 25.2 to 24.7 on Fit3D, from 47.9 to 47.5 on H36M, from 16.3 to 15.4 on AP3D, and from 64.8 to 58.4–58.6 on H4D. Similarly, P-MPJPE shows consistent reductions across datasets, reflecting improved alignment and structural coherence of predicted poses over time. The most pronounced gains appear on H4D, highlighting the effectiveness of kinematic regularization in challenging, fast-motion, and occlusion-heavy scenarios typical of surveillance contexts.

Bone-length Consistency: The anatomical bone-length term $\mathcal{L}_{\mathrm{B}}$ primarily improves MPBLE, enforcing realistic limb proportions. Its effect on overall accuracy is more dataset-dependent. On Fit3D, it achieves the lowest bone-length error (2.2 vs. 2.5 baseline), while combining it with kinematic losses does not provide further improvements. On H36M, bone-length consistency slightly enhances the benefits of kinematic regularization. For AP3D, kinematics alone yields the best bone-length results, while the addition of $\mathcal{L}_{\mathrm{B}}$ slightly increases the error. On H4D, however, the combination of bone-length and kinematic losses reduces MPBLE from 3.5 to 3.1, showing that anatomical constraints can compensate for minor side effects of kinematic smoothing under extreme motions.

Velocity vs. Acceleration: Comparing $\mathcal{L}_{K_{v1}}$ (flow + velocity) with $\mathcal{L}_{K_{v2}}$ (flow + velocity + acceleration) shows that acceleration provides only marginal additional benefits. MPJVE is essentially unchanged between the two variants, while minor fluctuations are observed in MPJPE and P-MPJPE. Given these small differences and the increased sensitivity of acceleration to annotation noise, $\mathcal{L}_{K_{v1}}$ is adopted as the preferred configuration for subsequent experiments. It achieves consistent temporal stabilization across all datasets while maintaining strong pose accuracy and structural fidelity.

Summary: Overall, kinematic regularization is the primary driver of improvements, ensuring smooth and realistic motion. Bone-length consistency

acts as a complementary term, mainly preserving anatomical plausibility. The combination is particularly beneficial in surveillance-oriented benchmarks such as H4D, where dynamic and occluded actions require stable and anatomically coherent skeletons for downstream action recognition. These results validate the proposed design choices as effective for producing high-quality 3D skeleton sequences suitable for SBAR.

7.2 3D Joint Orientation Prediction

In sports analytics and related fields, the precise estimation of 3D joint orientations plays a crucial role in understanding biomechanics, calculating forces, and analyzing motion. While traditional 3D-HPE models are able to estimate joint positions, they fail to capture joint rotations, which are essential for applications such as action recognition (which should translate for surveillance) and detailed biomechanical analysis. Human Mesh Recovery (HMR) models, such as those based on SMPL [Lop23], address this limitation by estimating full 3D poses, including joint rotations, through parametric body representations that separate shape and pose [Pav19a]. This enables the generation of detailed human meshes. However, while these parametric models offer plausible surface representations, their skeleton design is not anatomically accurate. For instance, the kinematic tree used in models such as SMPL-X does not align with the actual skeletal structure of the human body [Kel23]. Furthermore, HMR models are limited by their inability to utilize temporal information from video sequences, leading to reduced accuracy in scenarios involving high-speed or extreme movements, as frequently observed in sports [Lud25a] or in surveillance scenarios.

To overcome these challenges, hybrid methods have been proposed that combine 3D-HPE models with Inverse Kinematics (IK) to infer joint rotations [Lud25a]. While effective, this approach has significant drawbacks, as IK is computationally expensive and must be applied independently for each frame. Furthermore, such methods introduce additional complexity in the processing pipeline.

A similar concurrent work by Ludwig et al. [Lud25b] addresses these limitations by jointly predicting 3D joint positions and orientations to eliminate the need for IK. In line with this idea, however with a different architecture, the method proposed in this thesis also predicts both 3D joint positions and orientations simultaneously within a unified framework. This eliminates the need for IK, simplifying the workflow and reducing computational overhead. The simultaneous prediction of positions and orientations not only benefits sports analytics but also enhances applications such as skeleton-based action recognition, where accurate joint orientation is critical for distinguishing fine-grained motion patterns. Moreover, the explicit estimation of joint orientations improves the explainability of both action recognition [Qin22, Hua24, Sch24] and sports analytics by providing interpretable features that align with biomechanical principles, enabling deeper insights into movement dynamics. This integrated approach thus offers a more efficient, accurate, and interpretable solution for modeling human motion.

7.2.1 Rotation Representations

Rotations in three-dimensional space are mathematically represented through various parameterizations, each with distinct properties in terms of compactness, continuity, and computational constraints. The three most commonly used representations are rotation matrices, axis-angle form, and quaternions, which are defined in the mathematical spaces $SO(3)$, $\mathbb{R}^3$, and $\mathbb{H}$, respectively.

Rotation matrices. Rotation matrices belong to the special orthogonal group $SO(3)$, the set of all valid 3D rotations. A rotation matrix $\mathbf{R} \in SO(3)$ satisfies two key properties:

$$\mathbf{R}^\top \mathbf{R} = \mathbf{I}, \quad \det(\mathbf{R}) = 1, \tag{7.8}$$

where $\mathbf{I}$ is the 3×3 identity matrix, and $\mathbf{R}^\top$ is the transpose of $\mathbf{R}$. These properties ensure that rotation matrices preserve vector magnitudes and orientation in $\mathbb{R}^3$. A rotation is applied to a vector $\mathbf{v}$ as:

$$\mathbf{v}' = \mathbf{R}\mathbf{v}. \tag{7.9}$$

Although rotation matrices are explicitly used in practical implementations such as [Lud25b], they require nine parameters to encode a rotation, despite having only three degrees of freedom due to the constraints above. This redundancy complicates their use in optimization tasks, particularly in neural networks, where unconstrained outputs may result in invalid rotation matrices. To address this, projections to the closest valid rotation matrix are often employed using Singular Value Decomposition (SVD) [Lev20]. Alternatively, Zhou et al. [Zho19] propose using Gram-Schmidt orthogonalization as a computationally simpler method for enforcing orthonormality. However, this method is numerically unstable when the input vectors are nearly linearly dependent. These instabilities, coupled with the computational complexity of SVD, motivate the search for alternative representations, as explored by Zhou et al. through their proposed 6D continuous representation.

Axis-angle representation. The axis-angle representation describes a rotation using a unit vector $\mathbf{u} \in \mathbb{R}^3$, where $\|\mathbf{u}\| = 1$, and an angle $\theta \in [0, \pi]$ that specifies the magnitude of the rotation about $\mathbf{u}$. The compact representation is:

$$\boldsymbol{\omega} = \theta \mathbf{u}, \quad \boldsymbol{\omega} \in \mathbb{R}^3. \tag{7.10}$$

While axis-angle representation is compact, it suffers from discontinuities due to its double-cover property: a rotation by θ around $\mathbf{u}$ is equivalent to $-\theta$ around $-\mathbf{u}$. This ambiguity complicates optimization tasks. Degeneracies also occur when $\theta = 0$, where the axis becomes undefined, or when $\theta = \pi$, where small perturbations in the axis can lead to large changes in the rotation [Huy09].

Quaternions. Quaternions are unit elements of the four-dimensional Hamiltonian space $\mathbb{H}$ and provide a compact and continuous representation of 3D rotations. A quaternion is expressed as:

$$q = q_w + q_x \mathbf{i} + q_y \mathbf{j} + q_z \mathbf{k}, \tag{7.11}$$

where q_w is the scalar part, and (q_x, q_y, q_z) are the vector (imaginary) components. A unit quaternion satisfies:

$$\|q\| = 1. \tag{7.12}$$

Quaternions may be parameterized in terms of axis-angle representation. Given a unit rotation axis $\mathbf{u}$ and angle θ, the quaternion is written as:

$$q = \cos\left(\frac{\theta}{2}\right) + \sin\left(\frac{\theta}{2}\right)(u_x\mathbf{i} + u_y\mathbf{j} + u_z\mathbf{k}). \tag{7.13}$$

Quaternions require only four parameters with a single constraint ($\|q\| = 1$), thus those are more efficient than rotation matrices. They avoid singularities such as gimbal lock and allow smooth interpolation through spherical linear interpolation (SLERP) [Huy09, Fis21]. However, quaternions exhibit a double-cover property, where q and $-q$ represent the same rotation, introducing discontinuities. Neural networks may output non-unit quaternions, necessitating normalization, which adds some complexity but remains computationally efficient compared to other representations [Zho19].

Rotation matrices, axis-angle representations, and quaternions each present unique advantages and challenges. Rotation matrices are explicit, however, redundant, and computationally expensive to constrain. Axis-angle representations are compact, however, suffer from discontinuities and degeneracies. Quaternions, despite their double-cover property, offer an effective balance between compactness, robustness, and continuity. Thus, for this thesis, quaternions are the most suitable representation for tasks that require efficient and stable computation.

7.2.2 Ground Truth Computation

The GT orientations are generated for the H36M skeleton type by leveraging positional relationships between joints to define a complete rotational basis for each joint in 3D space, which is subsequently represented as quaternions. For each joint j, a primary child (the first child joint, or the joint itself if no child exists) and an auxiliary neighbor (a sibling or parent joint, with the

root joint defaulting to itself) are identified. For joints without a parent (*e.g.*, the root joint, shown in Figure 7.5), the rotation is initialized as the 3×3 identity matrix.

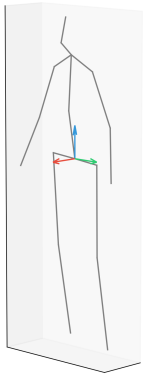


Figure 7.5: Root joint orientation – Visualization of the pelvis (root joint) orientation, aligned with the global coordinate axes, illustrating the reference frame used for 3D-HPE and skeletal transformations.

For all other joints, the X-axis is derived as the normalized vector pointing from the parent joint position to the primary child joint position:

$$\mathbf{x} = \text{normalize}(C_1 - P_J), \tag{7.14}$$

where C_1 is the position of the primary child joint and P_J is the position of the parent joint. A helper vector is then computed as the difference between the auxiliary neighbor position and the current joint position:

$$\mathbf{v} = C_2 - J, \tag{7.15}$$

where C_2 is the position of the auxiliary neighbor joint and J is the position of the current joint. Using this helper vector, the Z-axis is defined as the normalized cross product of the X-axis and the helper vector:

$$\mathbf{z} = \text{normalize}(\mathbf{x} \times \mathbf{v}), \tag{7.16}$$

and the Y-axis is computed as the cross product of the Z-axis and the X-axis:

$$\mathbf{y} = \mathbf{z} \times \mathbf{x}. \tag{7.17}$$

These three orthogonal axes form a right-handed coordinate system, which is assembled into a 3×3 rotation matrix:

$$\mathbf{R}_j = \begin{bmatrix} \mathbf{x} & \mathbf{y} & \mathbf{z} \end{bmatrix}_{3\times3}, \tag{7.18}$$

for each joint. The process is illustrated in Figure 7.6 and Figure 7.7.

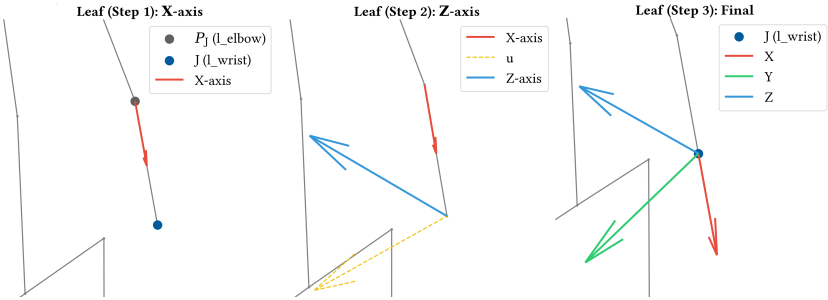


Figure 7.6: Leaf joint local coordinate system – Step-by-step visualization of how the local coordinate system is defined for a leaf joint, illustrated here for the left wrist, showing the alignment relative to its parent bone and the global skeleton structure.

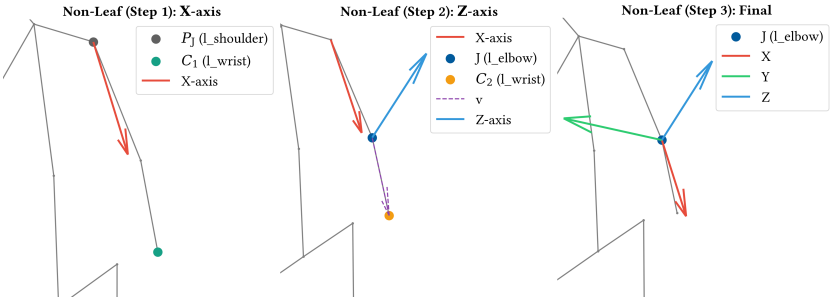


Figure 7.7: Non-leaf joint local coordinate system – Step-by-step illustration of how the local coordinate system is defined for a non-leaf joint, exemplified here with the left elbow, highlighting its orientation relative to parent and child bones within the skeleton hierarchy.

The rotation matrices are reshaped into a single array of dimensions $(N \cdot T \cdot J, 3, 3)$, where N is the number of sequences, T is the number of time steps, and J is the total number of joints in the H36M skeleton. These matrices are then converted into quaternions (w, x, y, z) using a standard matrix-to-quaternion conversion technique. Finally, the quaternion data is reshaped into the form $(N, T, J, 4)$, offering a compact and computationally efficient representation of joint orientations. This representation encodes the complete 3 Degree of Freedom (DoF) for each joint, encompassing both flexion/extension and twist/rotation, and serves as the GT. Furthermore, this design aligns with the hierarchical structure of the skeleton, capturing the internal dynamics and dependencies that propagate from parent nodes to child nodes, as established in prior work [Hu21].

7.2.3 Loss Function

To minimize errors in the predicted joint quaternions, the Average Angular Distance (AAD) [Son24] is employed, here denoted as $\mathcal{L}_{\text{angular}}$ and adapted for joint orientation instead of bone orientation in this contribution for the first time. This loss function minimizes the angular distance between the predicted and GT joint quaternions and is defined as:

$$\mathcal{L}_{\text{angular}} = \frac{1}{T\text{N}_{\text{joints}}} \sum_{t=1}^{T} \sum_{j=1}^{\text{N}_{\text{joints}}} 2 \arccos\left(\text{Re}\left(\bar{q}_{t,j} \times \text{conj}(q_{t,j})\right)\right), \tag{7.19}$$

where $\bar{q}_{t,j}$ and $q_{t,j}$ denote the GT and predicted quaternions of joint j at frame t, respectively. The functions $\text{Re}(\cdot)$ and $\text{conj}(\cdot)$ return the real part and the conjugate of a quaternion, respectively.

The $\mathcal{L}_{\text{angular}}$ operates directly on unit quaternions in $\mathbb{S}^3 \subset \mathbb{H}$, but it effectively minimizes the angular difference between rotations in $\mathbb{RP}^3$ by accounting for the double-cover property of quaternions.

Quaternions provide a double cover of the rotation group SO(3), meaning that both q and $-q$ represent the same rotation. The $\mathcal{L}_{\text{angular}}$ handles this naturally, as the inner product between quaternions (used in arccos) is invariant to the sign of the quaternion. This ensures that the loss function operates in the equivalence class $\mathbb{S}^3/\{\pm 1\} \cong \mathbb{RP}^3 \cong \text{SO}(3)$, allowing for consistent optimization of rotational accuracy.

The $\mathcal{L}_{\text{angular}}$ differs from the geodesic loss $\mathcal{L}_{\text{geodesic}}$, which measures the angular distance between two rotations on the manifold of the rotation group SO(3). The $\mathcal{L}_{\text{geodesic}}$ is defined as:

$$\mathcal{L}_{\text{geodesic}} = \frac{1}{T\text{N}_{\text{joints}}} \sum_{t=1}^{T} \sum_{j=1}^{\text{N}_{\text{joints}}} \arccos\left(\frac{\text{trace}(\mathbf{R}_{\text{gt}}^{\top}\mathbf{R}_{\text{pred}}) - 1}{2}\right), \tag{7.20}$$

where $\mathbf{R}_{\text{gt}}$ and $\mathbf{R}_{\text{pred}}$ are the GT and predicted rotation matrices for joint j at frame t, respectively, and $\text{trace}(\cdot)$ computes the sum of the diagonal elements of a matrix. This formulation measures the angle of the minimal rotation required to align $\mathbf{R}_{\text{pred}}$ with $\mathbf{R}_{\text{gt}}$ on the SO(3) manifold.

When the $\mathcal{L}_{\text{geodesic}}$ is reformulated for quaternions, it becomes equivalent to the $\mathcal{L}_{\text{angular}}$. Both losses ultimately compute the angular distance between

two rotations, but the $\mathcal{L}_{\text{angular}}$ operates directly in $\mathbb{S}^3$, the space of unit quaternions, while the $\mathcal{L}_{\text{geodesic}}$ is traditionally defined in SO(3), the space of rotation matrices. For the remaining of this work, $\mathcal{L}_{\text{angular}}$ is used.

7.2.4 Model Design

This thesis introduces two model variants designed to simultaneously predict joint positions and orientations.

Naive Approach. The first model adopts a straightforward multitask learning strategy by introducing a second branch, or "head,"for rotation estimation, alongside the existing position estimation head. Both heads share a common encoder derived from the backbone network, enabling joint feature extraction for both tasks. To predict the joint orientations, a linear layer is appended to the output of the backbone, producing 4D quaternions representing the rotations for the entire sequence. The newly introduced rotation head is trained using the $\mathcal{L}_{\text{angular}}$, while the position estimation head is optimized with the $\mathcal{L}_{3d}$. This approach directly extends the architecture to handle orientation prediction without modifying the core backbone structure, as illustrated in Figure 7.8, leveraging the shared encoder for multitask learning.

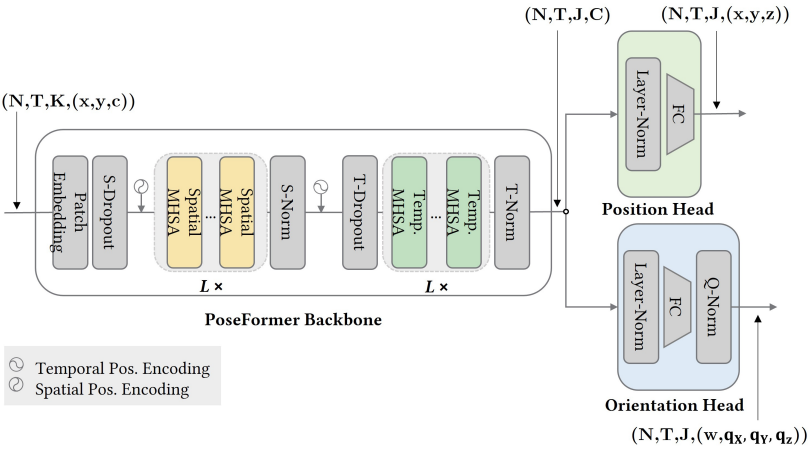


Figure 7.8: Naive approach to joint orientation prediction – The PoseFormer backbone outputs shared spatiotemporal features that feed both a position head for 3D joint coordinates and an orientation head for joint orientation quaternions in parallel, with no extra fusion layers or gating. Both heads backpropagate into the shared encoder.

Combining Pose and Orientation. The second approach draws inspiration from the work of Song et al. [Son24], where a model called QuaterGCN is introduced to predict joint positions and bone orientations simultaneously. In their design, separate encoders are used for each task, with each encoder followed by a Squeeze-and-Excitation block to recalibrate the channel-wise features for both positions and rotations. Bone orientations are predicted by first concatenating the rotation features with the joint position predictions. This concatenated feature vector is then processed through a series of layers: a linear layer, a batch normalization layer, a RELU activation layer, and a final linear layer that outputs the bone orientation quaternions. This connection between the position and orientation heads enables a direct influence of the $\mathcal{L}_{\text{angular}}$ on the position prediction head.

In this thesis, the idea is adapted to predict joint orientations instead of bone orientations. As in the single-task baseline, the joint position head consists of a linear layer that processes the output of the shared backbone to predict 3D joint positions. These predicted joint positions are then concatenated with

a copy of the backbone features. The concatenated feature vector is passed through a series of layers, including a linear layer, a batch normalization layer, a RELU activation layer, and a final linear layer, which outputs the joint orientation quaternions. In constrast to the setup in [Son24], the backbone encoder is shared between both tasks, avoiding the need for parallel streams. This shared backbone allows gradients from the $\mathcal{L}_{\text{angular}}$ loss to influence the shared features and, consequently, the position predictions both directly and indirectly.

To mitigate the excessive influence of the joint orientation head on position predictions via the shared backbone, the gradient flow from the joint orientation head to the backbone is actively blocked, as illustrated in Figure 7.9. This modification ensures that, in contrast to the naive approach, the joint position head benefits from additional constraints provided by the joint orientation head without introducing unintended coupling through shared gradients.

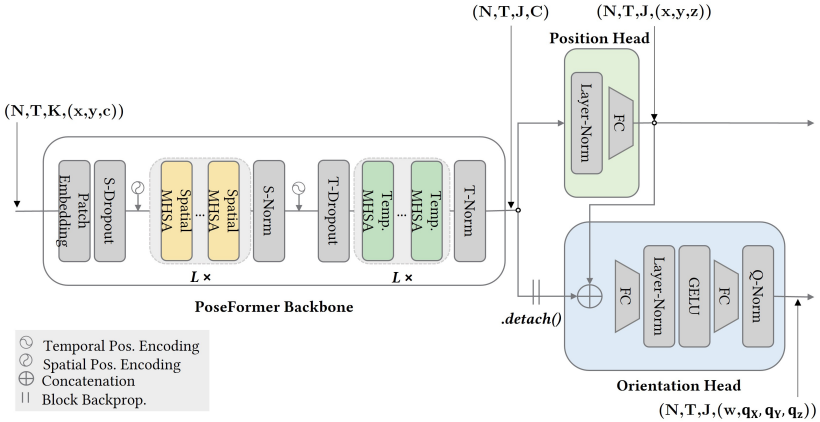


Figure 7.9: O_2 approach to joint orientation prediction – The PoseFormer backbone outputs shared spatiotemporal features. The position head produces 3D joint coordinates whose representation is concatenated with the backbone features and fed to the orientation head joint orientation quaternions. Gradients from the orientation head are blocked (detached) and do not backpropagate into the shared encoder, while the position head continues to update the backbone. The O_1 version is identical, except that its omit the detachment, thus allowing the gradient to flow back to the backbone.

7.2.5 Evaluation

The performance of joint orientation prediction in the multitask setting is evaluated, considering two main factors: the choice of feature coupling and the supervision loss for quaternion predictions. Performance is primarily assessed using MPJAE, which measures the mean per-joint angular error, while coordinate accuracy and temporal stability were monitored via MPJPE, P-MPJPE, and MPJVE. Two coupling strategies are investigated: concatenated features with full gradient flow (O_1) and concatenated features with gradient detachment (O_2). The results are summarized in Table 7.2.

Table 7.2: Comparison of design choices for joint orientation prediction - O_1 consistently attains the lowest MPJAE across datasets, while O_2 better preserves MPJPE/P-MPJPE and MPJVE. Best and second-best entries are highlighted in red and blue, respectively.

Method	Fit3D					H36M				
	MPJPE	P-MPJPE	MPJAE	MPJVE	MPBLE	MPJPE	P-MPJPE	MPJAE	MPJVE	MPBLE
Baseline	**25.2**	**19.2**	–	**1.7**	2.5	**47.9**	**38.8**	–	**2.9**	**11.2**
Baseline + naive	25.7	19.6	**10.0**	**1.7**	**2.3**	48.3	39.2	**17.3**	**2.9**	**11.2**
Baseline + concatenated (O_1)	25.6	**19.3**	**9.9**	**1.8**	**1.9**	48.1	**39.1**	**16.9**	3.0	**11.2**
Baseline + concatenated + detach (O_2)	**25.3**	**19.2**	**9.9**	**1.7**	**2.3**	**48.0**	**39.1**	18.0	**2.8**	**11.3**
Method	AP3D					H4D				
	MPJPE	P-MPJPE	MPJAE	MPJVE	MPBLE	MPJPE	P-MPJPE	MPJAE	MPJVE	MPBLE
Baseline	**16.3**	**12.6**	–	**2.4**	**3.8**	**64.8**	**34.6**	–	**10.7**	**3.5**
Baseline + naive	17.5	13.5	**8.8**	**2.8**	4.1	68.8	38.0	12.7	12.2	5.5
Baseline + concatenated (O_1)	17.5	13.5	**8.5**	**2.8**	4.1	66.9	37.3	**11.9**	**11.8**	**3.7**
Baseline + concatenated + detach (O_2)	**16.6**	**12.9**	9.9	**2.4**	**3.9**	**65.4**	**34.5**	**12.5**	**10.7**	3.8

Across datasets, O_1 consistently achieved the lowest angular errors, indicating superior orientation estimation. These improvements are accompanied by minor increases in coordinate and temporal errors, reflecting a trade-off between precise orientation and overall pose stability. In comparison, O_2 preserved positional and temporal metrics more effectively while exhibiting slightly higher angular errors, suggesting that gradient detachment provides a conservative alternative that prioritizes the stability of skeleton sequences.

Two supervision losses for quaternion orientations are compared: the geodesic $\mathcal{L}_{\text{angular}}$ and the standard $\mathcal{L}_{\text{MSE}}$. The results are presented in Table 7.3.

Table 7.3: Comparison of orientation losses – for MPJAE $\mathcal{L}_{\text{angular}}$ significantly outperforms $\mathcal{L}_{\text{MSE}}$ under both O_1 and O_2; O_1+$\mathcal{L}_{\text{angular}}$ yields the lowest MPJAE, while O_2+$\mathcal{L}_{\text{angular}}$ offers the best balance with coordinate and temporal metrics. Best and second-best entries are highlighted in red and blue, respectively.

Method	Fit3D					H36M				
	MPJPE	P-MPJPE	MPJAE	MPJVE	MPBLE	MPJPE	P-MPJPE	MPJAE	MPJVE	MPBLE
Baseline	**25.2**	**19.2**	–	**1.7**	2.5	**47.9**	**38.8**	–	**2.9**	**11.2**
Baseline + O_1 ($\mathcal{L}_{\text{MSE}}$)	27.1	20.5	**11.8**	2.0	3.0	48.7	39.7	20.4	3.0	11.8
Baseline + O_1 ($\mathcal{L}_{\text{angular}}$)	25.6	19.3	**9.9**	**1.8**	**1.9**	48.1	**39.1**	**16.9**	3.0	**11.2**
Baseline + O_2 ($\mathcal{L}_{\text{MSE}}$)	**25.0**	**18.9**	12.2	**1.7**	2.4	48.1	**39.1**	20.5	**2.9**	**11.1**
Baseline + O_2 ($\mathcal{L}_{\text{angular}}$)	25.3	**19.2**	**9.9**	**1.7**	**2.3**	**48.0**	**39.1**	**18.0**	**2.8**	11.3
Method	AP3D					H4D				
	MPJPE	P-MPJPE	MPJAE	MPJVE	MPBLE	MPJPE	P-MPJPE	MPJAE	MPJVE	MPBLE
Baseline	**16.3**	**12.6**	–	**2.4**	**3.8**	**64.8**	**34.6**	–	**10.7**	**3.5**
Baseline + O_1 ($\mathcal{L}_{\text{MSE}}$)	17.7	13.7	11.6	2.9	4.2	72.6	41.7	14.8	15.1	5.7
Baseline + O_1 ($\mathcal{L}_{\text{angular}}$)	17.5	13.5	**8.5**	**2.8**	4.1	66.9	37.3	**11.9**	11.8	**3.7**
Baseline + O_2 ($\mathcal{L}_{\text{MSE}}$)	19.2	15.0	16.0	**2.8**	4.5	65.8	34.8	15.1	**10.8**	4.1
Baseline + O_2 ($\mathcal{L}_{\text{angular}}$)	**16.6**	**12.9**	**9.9**	**2.4**	**3.9**	**65.4**	**34.5**	**12.5**	**10.7**	3.8

The $\mathcal{L}_{\text{angular}}$ consistently outperforms $\mathcal{L}_{\text{MSE}}$ across all datasets and coupling strategies, reducing angular errors while maintaining or improving coordinate and temporal metrics. The advantage of the angular loss is particularly evident in datasets characterized by rapid motions and frequent occlusions, where stable orientation predictions are critical. In contrast, the us of $\mathcal{L}_{\text{MSE}}$ results in notable degradations in both angular and positional accuracy under the same conditions.

Overall, the configuration O_1 with $\mathcal{L}_{\text{angular}}$ is preferred when the primary objective is minimizing orientation error. When a balance between orientation accuracy, temporal stability, and coordinate precision is required, O_2 with angular loss provides a more conservative option. These observations demonstrate that both the choice of feature coupling and the supervision loss significantly influence the trade-off between orientation accuracy and overall skeletal stability, particularly in challenging scenarios with fast or occluded movements.

7.3 Summary

This chapter explored the integration of anatomical constraints and joint orientation prediction in 3D human pose estimation. Anatomical priors, such as bone-length consistency, were found to enhance skeletal plausibility and contribute to the stabilization of motion sequences, particularly in datasets with rapid movements, occlusions, or complex interactions. Joint orientation estimation was examined in a multitask framework, comparing different feature coupling strategies and supervision losses. The use of concatenated features with full gradient flow (O_1) in combination with the $\mathcal{L}_{\text{angular}}$ consistently yielded the most accurate orientation predictions, whereas a detached coupling (O_2) provided a more conservative alternative that maintained higher temporal and positional stability. Together, these results demonstrate that anatomical constraints and orientation modeling serve complementary roles: the former ensures physically plausible poses, while the latter refines rotational accuracy. Their combined effects on pose fidelity and sequence stability are further investigated in Section 9.2. These improvements are particularly relevant for downstream applications such as surveillance and activity recognition, where both stable and anatomically consistent 3D poses are critical for reliable performance.

8 Skeleton Based Action Recognition

In this chapter, the focus lies on SBAR under surveillance-specific constraints. SBAR is performed using skeletal inputs derived from 2D-HPE or 3D-HPE and may exploit joint orientations in addition to joint- and bone-based modalities. A top-down, human-centric pipeline is assumed, with per-person detections and tracks provided. In contrast to much of the literature—where SBAR is evaluated on laboratory-captured 3D-HPE with known camera parameters and near-perfect skeletons [Wan14, Sha16, Liu19, Das19]—real-world deployments are uncalibrated and unconstrained. Assisted-living contexts may still offer favorable conditions (*e.g.* occasional depth sensing, few subjects, high resolution), but industrial monitoring typically involves frequent occlusions and clutter (see Figure 4.4(d)), while urban surveillance is marked by uncooperative behavior and high crowd density. Under such conditions, 2D keypoints are often noisy, incomplete, or fragmented, and absolute 3D-HPE becomes unreliable.

To realistically capture these challenges, we introduce a set of dedicated data augmentation strategies (Section 8.1) that simulate degradations characteristic of surveillance-based 2D-HPE. These augmentations, designed as part of this work, include occlusion, truncation, missing keypoints, temporal jitter, scale and viewpoint variation, and identity fragmentation. They provide a controlled means of stress-testing recognition models against the distortions most likely to appear in deployment.

In Section 8.2, skeletal representations are then systematically evaluated for SBAR. Comparisons are made between 2D keypoint sequences, 3D joint sequences, and 3D joints enriched with joint orientations (expressed as quaternions; *c.f.* Section 7.2). Root-relative coordinates are adopted throughout to remove dependence on camera calibration.

8.1 Data Augmentation for Real-world Limitations

In urban surveillance, either static or UAV-based, inputs to SBAR typically consist of per-person sequences of 2D skeletons, based on 2D keypoints, produced by a detector, a tracker, and a top-down pose estimator. Prior analyses by the author of this thesis have documented important failure modes in pose estimation [Cor22b], which are likewise observed in realistic action recognition datasets (*c.f.* Figure 8.1). Since pose estimation is computationally demanding, temporal upsampling is often applied, through interpolation or learned models, to increase temporal resolution or fill missing frames [Zen22a, Jin23]. However, such upsampling may introduce artifacts or amplify existing errors.

Skeleton inputs are often affected by multiple error types. Person detections may truncate individuals, causing missing body parts, detections may be missed entirely, or tracks may be fragmented, leading to identity switches. Partial or full occlusions by objects or other people are frequent in the wild. Under these conditions, temporal upsamplers can produce unnatural motion or interpolation artifacts. Furthermore, individual keypoints can be swapped, and left–right ambiguity may be incorrectly resolved by the pose estimator.

Figure 8.1: Wrong skeleton annotation in UAVHuman [Li21b] - From left to right: a hand keypoint is placed in the air while the hand is on the body of the person; the knee and feet of the person are placed between his legs while the right hand is placed on the left; the predictor seems to detect the head of the person on the backpack; two persons are merged into a single skeleton.

Despite these inaccuracies, such sequences are routinely forwarded to action recognition models, which are rarely designed to handle them. In this section, a set of augmentation techniques specific to skeleton sequences is introduced. The aim is to inject errors into the training data that are representative of real-world artifacts. These augmentations are intended to increase robustness to

such errors and to mitigate overfitting, thereby improving generalization. The techniques described below are intuitive and should be combined.

Those are briefly introduced here, while detailed algorithms are provided in a publication by the author in [Cor24a]. An overview of the realistic data augmentation techniques is shown in Figure 8.2.

The following subsections present the individual strategies in turn: occlusions (Section 8.1.1), interpolation (Section 8.1.2), keypoint swapping (Section 8.1.3), and the combined framework of Skelbumentations (Section 8.1.4). Their impact is then evaluated and discussed in Section 8.1.5.

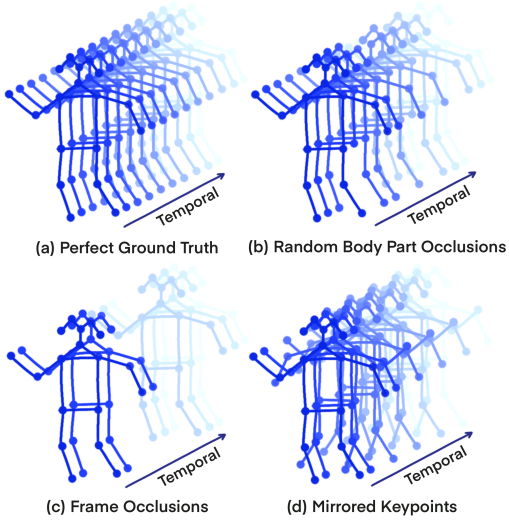


Figure 8.2: Realistic data augmentation – Illustration of augmented skeleton sequences. A skeleton sequence is shown as ground truth in a) where all keypoints of the skeletons are perfectly placed for each frame. Missing keypoints due to low-confidence are represented in b) where body parts such as a leg or an arm are missing on single frames. The case of a person totally occluded for a few seconds is shown in c), where a block of skeletons is missing for several frames. The case of mirror swapping for which keypoints detector fails to correctly differentiate between left and right is shown in d).

8.1.1 Occlusions

Artificial occlusions have been previously used to increase training data [Ang20]. Rather than randomly occluding the same keypoints throughout an entire sequence, an approach is adopted that better reflects real-world occlusion patterns. Inspired by Song et al. [Son21], multiple occlusion cases are employed. In contrast to mutually exclusive schemes, these cases can be applied simultaneously. Although a pose-aware augmentation comprising global and part-based jitter has been proposed [Sha22], different invariances are here targeted.

Instead of occluding selected keypoints across the entire sequence, keypoints are occluded only on randomly selected subsequences, reflecting the intermittent nature of real-world occlusions. A minimum occlusion duration is enforced so that trivial interpolation alone is insufficient. Following common practice, occluded joints are set to the origin [Son21]. The following occlusion augmentation cases are applied on randomly selected subsequences:

1 Frame Occlusions: All keypoints in the subsequence are set to zero, simulating the loss of keyframes.

2 Random body part occlusions: Four groups of keypoints are defined: left leg, right leg, left arm, and right arm. One body part is randomly selected and its keypoints are set to zero within the subsequence, simulating localized occlusions.

3 Random Keypoints Occlusion: Random keypoints are set to zero within the subsequence.

Each case is assigned a probability that governs its application to a given sample. When multiple cases are applied simultaneously, each operates on an independently selected subsequence.

8.1.2 Interpolation

Short-term occlusions may be reasonably reconstructed using simple interpolation [Ang20]. However, the reconstructed signal generally deviates from the original. Accordingly, the interpolation of short occlusions is employed as an additional data augmentation strategy, both to increase the variety of the data and to improve the robustness to interpolation artifacts.

For each sample, two subsequences are randomly selected. The first and last frames serve as anchors for interpolation, and the inner frames are treated as occluded. In one subsequence, the entire skeleton is occluded; in the other, only a random body part is occluded. The body parts considered are left leg, right leg, left arm, right arm, and back. The occluded keypoints are then reconstructed by linear interpolation between the two boundary frames.

8.1.3 Keypoint swapping

Since many pose estimation models operate on single frames, punctual failure in keypoint assignments may occur. To increase robustness to such errors, a simple augmentation is introduced that randomly swaps two keypoints within a frame. Building on this, an additional augmentation is considered in which all keypoints of a pair of limbs (legs or arms) are swapped, effectively mirroring the body parts.

8.1.4 Skelbumentations

A Python library, Skelbumentations, is introduced for skeleton sequence augmentation. Its design follows the principles of Albumentations [Bus20]. Although Albumentations supports keypoint augmentation, available operations are image-centric (*e.g.*, cropping, blurring), and sequence-level augmentation with keypoints is not supported. Concepts such as select and compose are, therefore, adopted, and the augmentation cases described above are implemented accordingly. Select operations allow for restricting transformations to parts of a sequence.

An internal invalid-map keeps track of which keypoints are occluded. If occlusion information is already present in the input data, this map can be passed directly into the pipeline. The map serves two purposes: it ensures that occluded joints are not altered by further perturbations, and it allows them to be excluded when computing motion-based features such as velocities. By default, occluded keypoints are set to the origin at the end of the pipeline, but this step can be disabled. In that case, the invalid-map is returned instead, leaving the decision on how to handle missing joints to downstream components. An example augmentation pipeline is shown in Figure 8.3: random body parts or entire frames are first occluded and interpolated, after which swap and mirror augmentations are applied.

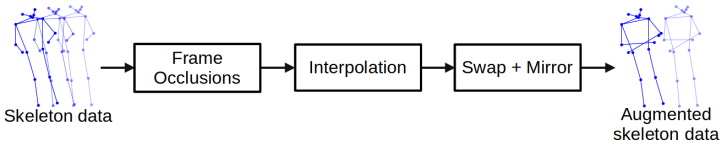


Figure 8.3: Augmentation pipeline overview – Schematic illustration of the proposed augmentation pipeline. First a random block of 5 to 21 Frames is set as occluded. Then the interpolation augmentation is applied. First, a block of whole frames is deleted and reconstructed by interpolation and second, the same procedure is repeated on an other random block, this time only for a single body part. Then random keypoints are swapped for random frames in the sequence. Finally, mirror augmentation is applied for small random blocks of frames.

8.1.5 Results and Discussion

The ablation results for the proposed augmentations in UAVHuman are summarized in Table 8.1. In general, applying Skelbumentations to the baseline increases the accuracy of Top-1 under both CSv1 and CSv2, where CSv1 follows the official cross-subject split of UAVHuman and CSv2 adopts a stricter re-split with non-overlapping subject identities to enforce stronger subject independence.

Table 8.1: Skeleton DA ablations on UAVHuman— The Top-1 accuracy is reported for two streams and their combination. Under CSv1, improvements are observed across all modalities; the largest gains occur for Joints and 2-Streams with Swap & Mirror, and for Bones with Interpolation. Under CSv2, Joints are most improved by Frame Augmentation, Bones are degraded, and 2-Streams show modest gains with Frame Augmentation or Interpolation but not with Swap & Mirror. **Red** indicates the best score per column, while **blue** marks the second-best.

Method	CSv1 ACC(%) ↑			CSv2 ACC(%) ↑		
	Joints	Bones	2-Streams	Joints	Bones	2-Streams
Baseline	40.0	39.8	42.8	64.7	**67.8**	**70.4**
+ Frame Aug	**41.5**	**42.4**	**44.6**	**69.0**	**65.2**	**71.6**
+ Interpolation	41.2	**42.5**	44.1	**68.0**	64.8	**71.6**
+ Swap & Mirror	**43.4**	42.2	**45.9**	**68.0**	64.5	**70.4**

Under CSv1, consistent gains are observed across all modalities: the joint stream (using raw joint coordinates), the bone stream (using relative joint differences), and their 2-stream ensemble. The greatest improvements occur for joints and 2-streams when applying swap & mirror, whereas bones benefit most from interpolation. This pattern suggests that invariances to left–right ambiguities and local keypoint perturbations are particularly valuable in this protocol.

Under CSv2, the strongest improvement for joints is obtained with frame augmentation. Bones degrade with all augmentations, indicating the sensitivity of this representation to the induced perturbations in this protocol. This is probably due to the rather poor quality of the skeletons provided by the dataset. However, such variable skeleton quality is a property of real-world action recognition and thus needs to be accounted for. For 2-Streams, modest gains arise from frame augmentation or interpolation, while swap & mirror yields no benefit. Across both protocols, 2-Streams consistently outperform single-modality input, indicating complementary cues and improved robustness to augmentation-induced variability.

These results suggest protocol- and modality-specific augmentation choices. Swap & mirror is most effective for CSv1 with joints and 2-Streams, Interpolation is most beneficial for CSv1 bones, and frame augmentation is preferable

for CSv2 Joints; for CSv2, bones-only pipelines appear less compatible with the tested perturbations. For deployment, the streams can be optimized separately to maximize performance, for instance by selecting stream-specific augmentation policies and hyper-parameters for joints, bones, and their fusion.

8.2 Comparative Evaluation of Skeleton Representations

This section evaluates alternative skeletal input representations for action recognition under a unified protocol (CS) with ACC and MAPCA as metrics. Representations at a given frame include 2D keypoint sequences mapped to pseudo-3D $(x_{2D}, y_{2D}, 1)$, ground-truth 3D joints (x_{3D}, y_{3D}, z_{3D}), and 3D joints augmented with joint orientations, encoded either as an axis–angle (Euler) vector $\mathbf{v}$ or as a unit quaternion (w, q_x, q_y, q_z). Two settings are considered: oracle inputs using ground-truth 3D, and realistic inputs using fixed 2D positions with depth and orientations estimated by learned models. The study investigates whether 3D ground-truth provides more informative inputs than 2D, whether adding joint orientations improves performance, how quaternion versus axis–angle encodings compare, and whether conclusions remain consistent when replacing ground-truth with predicted depth and orientations. Unless stated otherwise, skeletons are root-relative and scale normalized.

2D Keypoint Sequences: 2D inputs consist of per-frame keypoints (x_{2D}, y_{2D}) assigned to pseudo-3D by appending a depth channel, which produces $(x_{2D}, y_{2D}, 1)$ or $(x_{2D}, y_{2D}, \hat{z}_{3D})$ when depth is supplied by an uplift model. The 2D coordinates are normalized identically to the 3D setting—root-centered and scale-normalized applied before any depth substitution. Temporal alignment and sequence sampling follow the same protocol as for 3D inputs to ensure comparability across modalities.

3D Joint Sequences: 3D inputs comprise root-relative joint positions (x_{3D}, y_{3D}, z_{3D}) obtained from ground-truth or by uplifting predicted 2D keypoints. This representation assesses the contribution of depth disambiguation

and the sensitivity to uplift errors under realistic conditions. Coordinate normalization and temporal alignment are identical to the 2D setting.

3D Joint Sequences with Orientations: Joint orientations augment 3D positions to capture part-level rotational dynamics. Two encodings are considered: the axis–angle (Euler) vector $\mathbf{v}$ and the unit quaternion (w, q_x, q_y, q_z). Orientations are defined in local joint frames relative to parent bones. Quaternions are ℓ_2-normalized, and axis–angle magnitudes are bounded to avoid wrap-around. When predicted from lifted 3D, orientations are produced jointly with $\hat{z}_{3D}$.

Results and Discussion:

All models are trained and evaluated on Smarthome using the CS protocol with ACC and MAPCA. The same SBAR backbone, training schedule, sequence length, sampling strategy, and augmentation settings are used across inputs.

A four-stream ensemble is employed with joints, joints-motion, bones, and bones-motion. When orientations are included, a fifth stream (joints-orientation) is added, using local joint rotations encoded either as quaternions (w,q_x,q_y,q_z) or axis–angle vectors $\mathbf{v}$. All streams share identical backbones, normalization, and training schedules; only input channels differ. At inference, per-class logits are averaged across streams. Reported scores correspond to the four-stream ensemble for position-only inputs, while rows with orientations use the five-stream ensemble.

Depth $\hat{z}_{3D}$ is obtained from the lifting model with kinematic constraints (*c.f.* Section 7.1.2), trained with $T = 64$ frames on AMASS and finetuned on Smarthome following the same protocol as in Chapter 7. Predicted orientations $\hat{\mathbf{v}}$ and $(\hat{w},\hat{q}_x,\hat{q}_y,\hat{q}_z)$ are produced by a variant with a supplementary orientation head (*c.f.* Section 9.2).

The ablation in Table 8.2 is structured into five blocks, each probing a different regime of positional quality and supervision. This design allows disentangling the contributions of true versus predicted depth and evaluating

whether additional orientation cues help or harm recognition under varying input reliability.

Table 8.2: Ablation study on Smarthome – CS protocol. Comparison of GT 3D, mixed/predicted 3D, and 2D-based pipelines (pseudo or uplifted depth), with/without orientations. Hat notation denotes predicted quantities. **Red** indicates the best score per column, while **blue** marks the second-best.

Method	Smarthome (%) ↑	
	ACC	MAPCA
Block A: Ground-truth 3D		
$(x_{3D},y_{3D},1)$	78.2	59.1
(x_{3D},y_{3D},z_{3D})	78.1	**59.2**
(x_{3D},y_{3D},z_{3D}) + $\mathbf{v}$	**78.5**	58.8
(x_{3D},y_{3D},z_{3D}) + (w,q_x,q_y,q_z)	**78.4**	58.8
Block B: Fixed 3D inputs (x_{3D},y_{3D}) with uplifted depth		
$(x_{3D},y_{3D},\hat{z}_{3D})$	77.7	58.9
$(x_{3D},y_{3D},\hat{z}_{3D})$ + $\hat{\mathbf{v}}$	76.9	55.2
$(x_{3D},y_{3D},\hat{z}_{3D})$ + $(\hat{w},\hat{q}_x,\hat{q}_y,\hat{q}_z)$	77.0	56.0
Block C: Fully predicted 3D from (x_{3D},y_{3D}) inputs		
$(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$	**78.5**	58.5
$(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$ + $\hat{\mathbf{v}}$	77.8	56.5
$(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$ + $(\hat{w},\hat{q}_x,\hat{q}_y,\hat{q}_z)$	77.6	56.9
Block D: Fixed 2D input (x_{2D},y_{2D}) with uplifted depth		
$(x_{2D},y_{2D},1)$	76.8	57.2
$(x_{2D},y_{2D},\hat{z}_{3D})$	77.3	**59.3**
$(x_{2D},y_{2D},\hat{z}_{3D})$ + $\hat{\mathbf{v}}$	76.4	54.8
$(x_{2D},y_{2D},\hat{z}_{3D})$ + $(\hat{w},\hat{q}_x,\hat{q}_y,\hat{q}_z)$	76.2	55.2
Block E: Fully predicted 3D from 2D inputs		
$(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$	76.6	59.0
$(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$ + $\hat{\mathbf{v}}$	75.7	55.3
$(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$ + $(\hat{w},\hat{q}_x,\hat{q}_y,\hat{q}_z)$	75.8	55.3

Block A (GT 3D): Using accurate 3D skeletons provides the strongest overall performance. Adding orientations brings no clear benefit, while the highest ACC (78.5%) is achieved with positional-only inputs. The best MAPCA

(59.2%) comes from pure 3D positions. Both axis–angle $\mathbf{v}$ and quaternion (w,q_x,q_y,q_z) encodings behave similarly under near-perfect 3D inputs.

Block B (Mixed GT and uplifted depth): Replacing true depth with uplifted $\hat{z}_{3D}$ leads to a slight drop in both metrics. Adding predicted orientations further reduces performance, highlighting their sensitivity to noise when depth is uncertain.

Block C (Fully predicted 3D): Fully predicted $(\hat{x},\hat{y},\hat{z})$ achieves top ACC (78.5%) comparable to GT, though MAPCA drops to 58.5%, indicating weaker class balance. Predicted orientations again degrade performance for both encodings.

Block D (2D with uplifted depth): Depth cues are critical: appending $\hat{z}_{3D}$ to (x_{2D},y_{2D}) yields the best MAPCA (59.3%), outperforming GT 3D in class balance, though raw ACC is slightly lower (77.3%). Predicted orientations harm both metrics.

Block E (Predicted 3D from 2D): ACC (76.6%) and MAPCA (59.0%) remain competitive. Predicted orientations reduce performance, confirming their sensitivity to upstream 2D noise.

Summary: Across all regimes, positional depth is consistently beneficial: true depth for GT 3D and uplifted depth for 2D pipelines. Orientation streams help only when derived from reliable 3D and degrade otherwise, regardless of encoding. For deployment, position-only pipelines with uplifted depth are most robust when operating from 2D keypoints, while orientation streams should be enabled only with stable, high-quality 3D estimates. Ensembling across joints, bones, and motion streams provides additional robustness.

8.3 Summary

This chapter presents a comprehensive study of SBAR under surveillance-oriented constraints, highlighting both methodological contributions and empirical insights. First, dedicated data augmentation strategies, collectively implemented in the Skelbumentations library, are introduced to simulate real-world artifacts such as occlusions, missing keypoints, temporal jitter, and identity fragmentation. These augmentations not only improve model robustness and generalization under noisy and incomplete 2D skeleton sequences, but also reveal which types of perturbations most strongly impact each skeletal representation. Second, alternative skeletal representations are systematically evaluated, comparing 2D keypoints, 3D joint sequences, and 3D joints enriched with joint orientations, encoded as either axis–angle vectors $\mathbf{v}$ or quaternions (w,q_x,q_y,q_z). These experiments demonstrate that while uplifted 3D depth consistently enhances recognition, orientation cues are beneficial only when derived from reliable 3D inputs, highlighting the sensitivity of fine-grained motion features to upstream noise. Third, the interplay between input quality, depth lifting, and orientation estimation is analyzed through ablation studies, quantifying how errors propagate through the recognition pipeline and affect class-wise performance. Finally, a multi-stream ensemble combining joints, bones, motion, and orientation cues is proposed, showing that complementary streams provide robustness across different input regimes and partially compensate for noisy or incomplete data. Collectively, these findings advance understanding of which skeletal features and augmentation strategies are most effective for realistic surveillance scenarios, providing a modular, deployment-ready framework that balances accuracy, and robustness.

9 Evaluation

This chapter provides a unified evaluation of image- and sequence-based human analysis under surveillance-oriented constraints, assuming a top-down setting (*i.e.*, person detections and tracks are available). Three components are considered: 2D-HPE on RGB and thermal imagery, root-relative 3D-HPE with anatomical/kinematic regularization and joint-orientation prediction, and SBAR with controlled stream multiplicity. Accuracy, cross-domain generalization, and accuracy–efficiency trade-offs are emphasized. Complexity and latency are reported where relevant.

Section 9.1 evaluates 2D-HPE on UPAR-Pose (RGB) and UPPET (thermal) under specialization and cross-domain protocols (*e.g.*, 3FCV, 4FCV, and LOOCV). State of the art methods span coordinate regression [Tos14], CNN heatmap regression [Wan21], coordinate classification [Li22b], and transformer heatmap regression [Xu22]. A multitask variant with PAR and stronger data augmentation is analyzed for robustness and efficiency. Section 9.2 investigates 3D HPE on Fit3D, H36M, AP3D, and H4D, combining kinematic losses and soft anatomical consistency with a quaternion-based orientation head (*c.f.* Sections 7.1.1, 7.1.2 and 7.2). Section 9.3 compares SBAR methods on UAVHuman (CSv1) and Smarthome (CS) under a unified protocol using GT 2D pseudo-depth. A 3D-lifted variant is additionally evaluated on Smarthome. Per-stream reporting is used to normalize across architectures with differing stream multiplicities.

9.1 2D Human Pose Estimation

2D-HPE is evaluated on visible-spectrum (UPAR-Pose) and thermal (UPPET) benchmarks under specialization (in-domain) and cross-domain protocols. Approaches considered comprise coordinate regression (DeepPose), CNN heatmap regression (HRNet), coordinate classification (SimCC), and transformer-based heatmap regression (ViTPose). In addition, single-task baselines with DA and a multitask shared-encoder configuration with TSA are included at Small and Huge backbone scales. Unless stated otherwise, metrics are AP and AP_{50}. The analysis emphasizes accuracy, generalization, and computational efficiency, *e.g.*, FLOPs and latency.

9.1.1 Specialization

This subsection reports in-domain specialization results for visible-spectrum and thermal benchmarks. The objective is to quantify per-keypoints localization accuracy when training and testing occur within the same dataset partitions, *i.e.*, absent cross-dataset domain shift.

9.1.1.1 UPAR-Pose

The specialization results on UPAR-Pose and its sub-datasets are summarized in Table 9.1. Across all splits, ViTPose-Huge achieves the strongest performance in AP, underscoring the benefits of high-capacity transformer backbones for heatmap regression. Qualitative results are visualized in Figure 9.2.

Figure 9.1: Qualitative results on UPAR-Pose – Predicted keypoints from the ViTPose-Huge model trained under the UPAR-Pose specialization protocol. The model achieves highly accurate pose predictions across diverse viewpoints, illuminations and resolutions.

The multitask shared-encoder with TSA and DA contributed in this thesis remains highly competitive: the Small variant lags moderately behind the strongest single-task models, while the Huge variant nearly matches ViTPose-Huge, highlighting the effectiveness of specialization within a unified framework.

Two additional observations arise. First, AP_{50} is saturated on UPAR-Pose (98–99%). Thus, AP is the more informative metric for distinguishing model performance as it is more sensitive to small localization errors. Second, CNN-based baselines (*e.g.*, HRNet) continue to perform strongly under specialization, though transformer models gain a larger advantage as backbone scale increases, consistent with capacity-driven improvements reported in prior work.

Table 9.1: UPAR-Pose specialization results on the full benchmark and its three sub-datasets (Market-1501, PA-24K, and PETA). The compared methods span different paradigms: coordinate regression (DeepPose), CNN-based heatmap regression (HRNet), coordinate classification (SimCC), and transformer-based heatmap regression (ViTPose). ViTPose-Huge achieves the strongest overall performance in AP, consistently outperforming other models across all splits. Despite reduced capacity, the multitask shared-encoder with TSA and DAmentation remains competitive, demonstrating the effectiveness of specialization within a unified architecture. **Red** highlights the best score per column, while **blue** indicates the second-best.

Method	Market-1501		PA-24K		PETA		UPAR-Pose	
	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑
DeepPose-r50 [Tos14]	90.6	**98.8**	87.9	97.9	90.6	97.6	91.7	**98.8**
HRNetw48-udp [Wan21]	93.1	**98.8**	93.6	**98.9**	94.5	**98.8**	93.9	**98.8**
SimCC [Li22b]	91.7	**98.9**	90.5	**99.0**	93.0	98.5	93.0	**98.9**
ViTPose-Small [Xu22]	91.1	**98.9**	87.7	**98.9**	92.1	**98.8**	92.6	**98.8**
ViTPose-Huge [Xu22]	**93.6**	**98.8**	**94.2**	**98.9**	**95.4**	**98.8**	**94.2**	**98.8**
Shared Encoder Small + TSA + DA (ours)	90.1	**98.9**	87.1	**98.9**	90.0	**98.7**	91.0	**98.8**
Shared Encoder Huge + TSA + DA (ours)	**93.3**	**98.9**	**94.1**	**99.0**	**95.3**	**98.8**	**94.1**	**98.8**

9.1.1.2 UPPET

Specialization results on UPPET and its four sub-datasets are presented in Table 9.2. ViTPose-Huge provides the highest overall accuracy, with consistent gains over CNN-based baselines. Qualitative results are visualized in Figure 9.2.

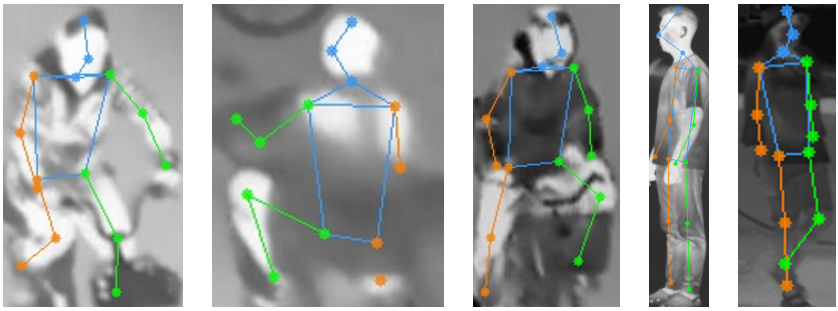


Figure 9.2: Qualitative results on UPPET – Predicted keypoints from the ViTPose-Huge model trained under the UPPET specialization protocol. The model achieves highly accurate pose predictions across diverse viewpoints and crop size for diverse sensors.

The single-task baseline with DA and a Huge backbone achieves competitive performance and attains the top AP/AP_{50} on CAMEL-P, indicating that stronger regularization is beneficial in thermal imagery (*c.f.* Section 6.2). On the full UPPET benchmark, the augmented Huge baseline approaches ViTPose-Huge within a narrow margin while matching or nearly matching AP_{50}, suggesting limited headroom for per-joint localization improvements under in-domain conditions.

Table 9.2: UPPET specialization results on the full benchmark and its four sub-datasets (LLVIP-P, OTP, CAMEL-P, TPE). Methods include coordinate regression (DeepPose), CNN heatmap regression (HRNet), coordinate classification (SimCC), and transformer-based heatmap regression (ViTPose). ViTPose-Huge achieves the highest overall AP/AP_{50}, while the DAmented Huge baseline is competitive and reaches the top scores on CAMEL-P. **Red** indicates the best score per column, while **blue** marks the second-best.

Method	LLVIP-P		OTP		CAMEL-P		TPE		UPPET	
	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑
DeepPose-r50 [Tos14]	80.7	96.9	76.9	93.8	66.6	92.7	69.9	87.4	76.2	92.6
HRNetw48-udp [Wan21]	89.8	**98.0**	87.0	96.8	75.4	94.8	77.6	91.7	81.0	**94.8**
SimCC [Li22b]	85.6	**98.0**	81.8	95.7	72.0	93.6	73.6	90.5	78.7	93.7
ViTPose-Small [Xu22]	84.6	97.9	79.4	94.4	74.9	94.8	75.5	91.7	79.8	94.7
ViTPose-Huge [Xu22]	**92.0**	**99.0**	**90.8**	**97.9**	**80.3**	**96.9**	**81.4**	**94.7**	**85.3**	**96.9**
Baseline HPE Small + DA (ours)	84.9	97.9	78.0	93.1	75.4	94.9	74.9	91.7	79.6	94.7
Baseline HPE Huge + DA (ours)	**91.9**	**99.0**	**90.3**	**97.6**	**80.9**	**97.0**	**81.1**	**93.7**	**84.6**	**96.9**

9.1.2 Generalization

This subsection assesses cross-domain robustness under 3FCV/LOOCV for UPAR-Pose and 4FCV/LOOCV for UPPET. The analysis targets performance stability under dataset shift *e.g.*, changes in scene, viewpoint, clothing, or sensor characteristics, and the effect of capacity and DA on mean AP and variance.

9.1.2.1 UPAR-Pose

Cross-domain generalization under 3FCV and LOOCV is reported in Table 9.3. Model capacity emerges as a primary factor: ViTPose-Huge substantially outperforms smaller backbones. The multitask shared-encoder (Huge) with

TSA and DA closely matches ViTPose-Huge on AP and equals or slightly exceeds it on AP_{50}, indicating that increased capacity largely mitigates representational competition between tasks. In contrast, the Small-capacity shared-encoder exhibits a moderate reduction in AP relative to single-task ViTPose-Small, consistent with the capacity-limited regime observed in ablations *c.f.* Section 6.1. Variability across folds decreases with larger backbones and with augmentation, reinforcing the conclusion that capacity and regularization jointly enhance cross-domain robustness, *e.g.*, lower standard deviations under 3FCV.

Table 9.3: Cross-domain generalization on UPAR-Pose under 3FCV and LOOCV. Mean±std AP and AP_{50} over folds are reported. The multitask shared-encoder with a Huge backbone matches ViTPose-Huge on AP_{50} and trails marginally on AP, indicating that increased capacity largely mitigates multitask interference. **Red** indicates the best score per column, while **blue** marks the second-best.

Approach	3FCV		LOOCV	
	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑
DeepPose-r50 [Tos14]	73.1 ± 8.0	93.4 ± 4.2	83.3 ± 2.8	96.9 ± 1.3
HRNetw48-udp [Wan21]	82.4 ± 6.0	95.6 ± 2.1	89.1 ± 1.4	97.7 ± 1.4
SimCC [Li22b]	77.5 ± 7.5	94.5 ± 3.0	85.7 ± 2.2	97.4 ± 1.0
ViTPose-Small [Xu22]	80.8 ± 4.1	**96.4** ± 1.0	86.2 ± 2.6	97.7 ± 0.8
ViTPose-Huge [Xu22]	**89.6** ± 1.5	**98.4** ± 0.6	**91.3** ± 1.0	**98.1** ± 0.5
Shared Encoder Small + TSA + DA (ours)	78.9 ± 4.1	**96.4** ± 1.0	84.5 ± 2.5	97.3 ± 1.1
Shared Encoder Huge + TSA + DA (ours)	**89.2** ± 1.6	**98.4** ± 0.5	**91.4** ± 0.6	**98.5** ± 0.6

9.1.2.2 UPPET

Results for 4FCV and LOOCV on UPPET are shown in Table 9.4. The single-task baseline with DA and a Huge backbone achieves the strongest AP/AP_{50} across both protocols, surpassing ViTPose-Huge. Gains are observed in both mean AP and reduced variance across folds, confirming that heavy DA is particularly effective for cross-domain robustness in thermal imagery, *i.e.*, domain-shift robustness benefits more from regularization than from additional capacity alone.

Table 9.4: Cross-domain generalization on UPPET under 4FCV and LOOCV. Mean±std AP and AP_{50} over folds are reported. Data DAmentation consistently improves robustness; with the Huge backbone, the DAmented baseline attains the strongest accuracy across both protocols. **Red** indicates the best score per column, while **blue** marks the second-best.

Approach	4FCV		LOOCV	
	AP ↑	AP_{50} ↑	AP ↑	AP_{50} ↑
DeepPose-r50 [Tos14]	27.6 ± 10.6	50.5 ± 14.3	40.1 ± 14.4	67.3 ± 15.1
HRNetw48-udp [Wan21]	37.7 ± 13.3	58.7 ± 17.0	54.9 ± 12.8	78.0 ± 11.6
SimCC [Li22b]	32.0 ± 12.0	54.7 ± 17.0	46.3 ± 13.5	73.6 ± 14.5
ViTPose-Small [Xu22]	39.1 ± 5.4	65.7 ± 5.3	49.3 ± 14.1	75.9 ± 13.1
ViTPose-Huge [Xu22]	**57.8** ± 4.5	**79.5** ± 4.3	**68.3** ± 13.4	**85.4** ± 12.1
Baseline HPE Small + DA (ours)	42.5 ± 3.4	70.5 ± 2.0	50.9 ± 16.3	75.5 ± 13.4
Baseline HPE Huge + DA (ours)	**61.9** ± 3.4	**82.3** ± 3.1	**71.1** ± 11.5	**87.6** ± 9.3

9.1.3 Benchmarking 2D Human Pose Estimation in Nighttime RGB vs Thermal Images

To systematically investigate the challenges of 2D-HPE under low-light conditions, this work evaluates state-of-the-art models on both nighttime RGB and thermal images. The experiments leverage the LLVIP-P dataset, which was introduced as part of the contributions of this thesis to enable quantitative and qualitative comparisons across modalities. The goal is to assess the relative effectiveness of thermal imagery in overcoming the limitations of low-illumination RGB inputs.

All models were trained under identical configurations, with a batch size of 16. Pretrained weights from ImageNet were employed whenever available. Otherwise, model parameters were initialized randomly. To ensure a fair comparison, each model was trained independently on thermal and RGB images, resulting in separate training instances for the two modalities.

The evaluation results on the test set of LLVIP-P are summarized in Table 9.5 and Table 9.6. Overall, thermal-based models consistently outperform RGB-based models across all *AP* metrics. Among the evaluated approaches,

ViTPose-Huge achieves the highest performance for both modalities, considering both GT and predicted bounding boxes.

Table 9.5: Benchmarking results for LLVIP-Pose – Average precision (AP) is reported for different bounding box sizes (AP_M, AP_L). **Red** highlights the best score per column, while **blue** indicates the second-best.

Thermal								
Models	**GT BBoxes**				**Predicted BBoxes** (Thermal, *AP*: 66.4)			
	AP	AP_{50}	AP_M	AP_L	AP	AP_{50}	AP_M	AP_L
HRNetw48-udp [Wan21]	**90.0**	**98.0**	**60.5**	**90.2**	**88.6**	**96.5**	**24.8**	**88.7**
ViTPose-Huge [Xu22]	**91.6**	**99.0**	**63.4**	**91.7**	**89.9**	**96.5**	**27.0**	**90.2**
DeepPose-r50 [Tos14]	85.2	97.9	35.0	85.4	84.1	**96.4**	17.6	84.2
SimCC [Li22b]	87.7	97.9	45.7	87.9	86.7	**96.4**	21.7	86.9
DEKR [Gen21]					84.5	95.3	6.5	85.0
YOLOX-Pose-l [Maj22]					84.8	96.2	22.3	85.0
RTMO-l [Lu24]					85.5	96.1	24.3	85.8
RGB								
Models	**GT BBoxes**				**Predicted BBoxes** (RGB, *AP*: 53.6)			
	AP	AP_{50}	AP_M	AP_L	AP	AP_{50}	AP_M	AP_L
HRNetw48-udp [Wan21]	**64.3**	**91.4**	17.6	**64.7**	**59.1**	86.6	8.7	**59.4**
ViTPose-Huge [Xu22]	**68.1**	**93.7**	**25.3**	**68.4**	**63.2**	**87.9**	**14.2**	**63.4**
DeepPose-r50 [Tos14]	59.6	90.6	15.3	59.9	54.7	85.4	6.4	54.9
SimCC [Li22b]	61.9	90.6	**19.9**	62.1	57.2	85.6	7.4	57.4
DEKR [Gen21]					56.2	85.4	1.0	56.6
YOLOX-Pose-l [Maj22]					56.9	**86.7**	**12.8**	57.1
RTMO-l [Lu24]					56.1	85.3	9.6	56.3

Focusing on top-down approaches, a notable difference emerges between results obtained using GT versus predicted bounding boxes. Transitioning from GT to predicted bounding boxes leads to an approximate AP drop of 0.015 for thermal models and 0.05 for RGB models. The smaller decrease in thermal-based models highlights their higher detection reliability, reflecting the inherent difficulty of person detection in low-illumination RGB images. Thermal models also demonstrate higher AP_M values, indicating improved performance for medium-sized bounding boxes. Nevertheless, this metric may not fully capture the advantage of thermal imaging for medium-scale human instances, as medium and large bounding boxes frequently correspond to similar person scales, as illustrated in Figure 9.3.

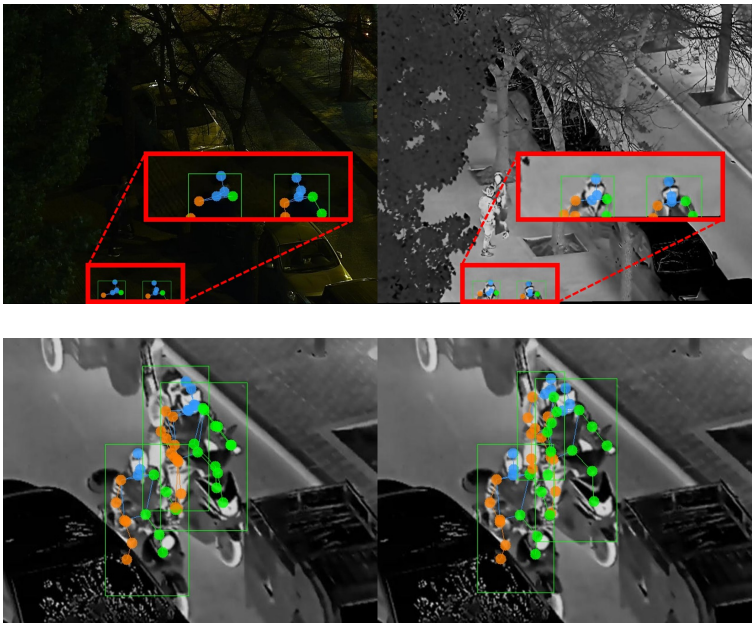


Figure 9.3: Qualitative results on H4D – Visualization of pose predictions under different conditions. The left column shows medium bounding boxes, while the right column corresponds to a hard crowding level. Medium bounding boxes represent human instances of similar scale to large bounding boxes but are cropped due to image boundaries. Comparison of SimCC predictions with ground truth shows that poses with ground truth bounding boxes (left) achieve lower AP_H^C than those with predicted bounding boxes (right).

Bottom-up and single-stage approaches show similar trends, with thermal-based models achieving higher AP values comparable to top-down approaches.

The analysis of crowding levels, reported in Table 9.6, further confirms that thermal-based models achieve higher AP^C across both GT and predicted bounding boxes. Single-stage models—except for ViTPose-Huge—demonstrate higher AP_H^C than top-down approaches in both modalities. In top-down models, overlapping bounding boxes may result in sub-optimal pose predictions: the background bounding box may inherit the pose of

the foreground bounding box, leaving the foreground bounding box with duplicate poses. This phenomenon is particularly evident in thermal images for the SimCC model, where higher AP_H^C is observed for predicted bounding boxes, as illustrated in Figure 9.3.

Table 9.6: Benchmarking results by crowding level – Average precision (AP) is reported for different crowding conditions (AP_E^C, AP_M^C, AP_H^C). **Red** highlights the best score per column, while **blue** indicates the second-best.

Thermal										
Models	**Groundtruth BBoxes**					**Predicted BBoxes** (Thermal, AP: 66.4)				
	AP	AP_{50}	AP_E^C	AP_M^C	AP_H^C	AP	AP_{50}	AP_E^C	AP_M^C	AP_H^C
HRNetw48-udp [Wan21]	**90.0**	**98.0**	**90.6**	**88.6**	**83.3**	**88.6**	**96.5**	**89.3**	**86.6**	76.0
ViTPose-h [Xu22]	**91.6**	**99.0**	**92.1**	**90.1**	**83.7**	**89.9**	**96.5**	**90.7**	**87.8**	**79.7**
DeepPose-r50 [Tos14]	85.2	97.9	85.9	82.6	75.1	84.1	**96.4**	84.9	80.9	72.8
SimCC [Li22b]	87.7	97.9	88.4	85.3	72.1	86.7	**96.4**	87.6	83.9	72.4
DEKR [Gen21]						84.5	95.3	85.2	81.8	70.7
YOLOX-Pose-l [Maj22]						84.8	96.2	85.1	83.8	**77.5**
RTMO-l [Lu24]						85.5	96.1	85.9	84.5	**79.7**
RGB										
Models	**Groundtruth BBoxes**					**Predicted BBoxes** (RGB, AP: 53.6)				
	AP	AP_{50}	AP_E^C	AP_M^C	AP_H^C	AP	AP_{50}	AP_E^C	AP_M^C	AP_H^C
HRNetw48-udp [Wan21]	**64.3**	**91.4**	**65.1**	**62.6**	54.6	**59.1**	86.6	**59.8**	**57.6**	49.7
ViTPose-h [Xu22]	**68.1**	**93.7**	**68.3**	**67.8**	**71.4**	**63.2**	**87.9**	**63.5**	**62.6**	**56.7**
DeepPose-r50 [Tos14]	59.6	90.6	60.1	58.4	56.0	54.7	85.4	55.1	52.9	48.0
SimCC [Li22b]	61.9	90.6	62.6	60.1	**57.5**	57.2	85.6	57.8	55.6	50.8
DEKR [Gen21]						56.2	85.4	57.0	53.6	51.9
YOLOX-Pose-l [Maj22]						56.9	**86.7**	57.1	56.5	50.4
RTMO-l [Lu24]						56.1	85.3	56.1	56.5	**52.7**

Qualitative results are presented in Figure 9.4, showing pose predictions from ViTPose-Huge for both RGB and thermal images. While both modalities produce generally accurate poses, thermal-based predictions exhibit superior accuracy under challenging illumination conditions, such as low-light regions (top-right) and areas with strong illumination changes (top-left, bottom-middle, and bottom-right). For instance, in the top-right image pair, thermal-based predictions correctly capture the body orientations of pedestrians in dark regions, whereas RGB-based predictions fail to do so.

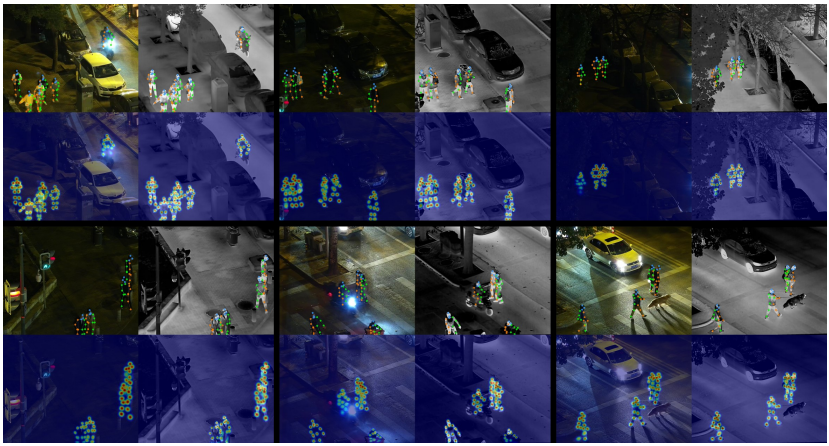


Figure 9.4: Qualitative results of ViTPose for RGB and thermal image pairs. In low-light scenarios (e.g., top right) and scenes with strong illumination changes (e.g., top-left, bottom-middle, bottom-right), thermal-based predictions achieve higher accuracy, as reflected by stronger pose heatmap activations.

The conducted experiments demonstrate that thermal-based models consistently outperform RGB-based models for 2D human pose estimation under low-light conditions. ViTPose-Huge achieves the highest performance across both modalities for GT and predicted bounding boxes. Top-down models show a smaller drop in AP when moving from GT to predicted bounding boxes in thermal images compared to RGB, highlighting the higher detection reliability of thermal imagery. Thermal-based models also perform better for medium-sized instances (AP_M) and under crowded conditions (AP^C), while single-stage approaches generally handle highly crowded scenarios more effectively. Qualitative results further confirm that thermal images provide more accurate pose predictions in challenging illumination conditions, particularly in dark or strongly illuminated regions.

9.1.4 Computational Efficiency

Model complexity and runtime on UPAR-Pose are summarized in Table 9.7. Parameters and FLOPs are reported at the evaluation input size, and runtime corresponds to per-person-crop latency (batch size 1) on an NVIDIA L40 GPU. The shared-encoder variants maintain parameter counts and FLOPs similar to their corresponding ViTPose backbones, with runtime slightly higher in the Huge setting due to multitask computations. In the Huge configuration, the shared-encoder achieves near-identical 3FCV accuracy to ViTPose-Huge while slightly increasing latency, illustrating that high-capacity multitask models can retain competitive performance with minimal overhead. In the Small configuration, the shared-encoder provides the lowest latency among high-performing models, though with a modest reduction in cross-domain AP, highlighting an attractive option when runtime or memory constraints are critical, such as for edge deployments. Overall, the results demonstrate that the shared-encoder design balances multitask capability, accuracy, and efficiency effectively.

Table 9.7: Model complexity and runtime on UPAR-Pose. Parameters (M) and FLOPs (G) are computed at the evaluation input size. Runtime is per-person-crop latency (batch size 1) measured on an NVIDIA L40 GPU. The shared-encoder Huge variant achieves near-identical accuracy to ViTPose-Huge under 3FCV with slightly lower FLOPs and latency. **Red** indicates the best score per column, while **blue** marks the second-best.

Method	Parameters (M) ↓	FLOPs (G) ↓	Runtime (ms) ↓	3FCV (%)	
				AP ↑	AP_{50} ↑
DeepPose-r50 [Tos14]	**23.7**	**4.0**	20.0	73.1 ± 8.0	93.4 ± 4.2
HRNetw48-udp [Wan21]	63.6	15.7	38.6	82.4 ± 6.0	95.6 ± 2.1
SimCC [Li22b]	36.8	5.5	**17.7**	77.5 ± 7.5	94.5 ± 3.0
ViTPose-Small [Xu22]	**24.3**	**5.3**	**20.7**	80.8 ± 4.1	**96.4** ± 1.0
ViTPose-Huge [Xu22]	637.2	122.9	37.0	**89.6** ± 1.5	**98.4** ± 0.6
Shared Encoder Small + TSA + DA (ours)	24.6	**5.3**	22.0	78.9 ± 4.1	**96.4** ± 1.0
Shared Encoder Huge + TSA + DA (ours)	640.5	123.5	51.4	**89.2** ± 1.6	**98.4** ± 0.5

9.1.5 Results and Discussion

Under specialization, ViTPose-Huge achieves the highest accuracy across both visible and thermal benchmarks, while augmented single-task baselines and the Small-capacity multitask shared-encoder remain competitive.

Cross-domain evaluation highlights the impact of backbone capacity and augmentation strategies: on UPAR-Pose, the Huge shared-encoder attains near-identical AP_{50} to ViTPose-Huge and trails only slightly in AP. On UPPET, the augmented Huge single-task baseline achieves the top performance across most sub-datasets and metrics. Computational efficiency analysis shows that shared-encoder designs in the Huge regime maintain near-maximum accuracy with slightly higher runtime, whereas the Small variant provides substantial latency reductions at a modest cost in cross-domain accuracy.

These results are particularly relevant for real-world surveillance and human-centric applications where a joint solution for 2D-HPE and PAR is desirable to support tasks such as suspect description. The experiments demonstrate that a multitask backbone successfully learns both HPE and soft biometric attributes, achieving competitive performance across tasks. In scenarios where PAR is not required, it is preferable to omit soft biometric estimation, which allows higher processing speed and slightly improved 2DHPE accuracy by focusing the model capacity on a single task. Nevertheless, the shown multitask results are significant: as observed in Section 6.1, PAR benefits greatly from HPE features, confirming that shared representations offer synergistic gains when multiple tasks are jointly trained. The findings highlight the practical trade-offs between accuracy, efficiency, and the collection of sensitive information, demonstrating the feasibility of a multitask approach while informing when task-specific specialization may be advantageous.

9.2 3D Human Pose Estimation

This section evaluates sequence-based, root-relative 3D-HPE using the constraints and orientation prediction framework introduced in Chapter 7. Four datasets are considered: Fit3D, H36M, AP3D, and H4D. Metrics include MPJPE, P-MPJPE, MPJAE, MPJVE, and MPBLE. The ablations examine combinations of (i) kinematic constraints $\mathcal{L}_{K_{v1}}$ and $\mathcal{L}_{K_{v2}}$, (ii) bone-length consistency $\mathcal{L}_{\mathrm{B}}$, and (iii) joint-orientation coupling variants O_1 and O_2 (*c.f.* Section 7.2). The analysis targets the trade-off between positional

accuracy and kinematic plausibility *e.g.*, temporal smoothness and stable bone lengths.

9.2.1 Combination of Approaches

This subsection quantifies the effect of combining improved kinematics ($\mathcal{L}_{K_{v1}}$, $\mathcal{L}_{K_{v2}}$), anatomical consistency ($\mathcal{L}_{\mathrm{B}}$), and orientation coupling (O_1, O_2) across datasets and metrics. The goal is to characterize consistent patterns, *i.e.*, whether improvements in MPJPE orP-MPJPE correlate with gains in MPJAE,MPJVE, orMPBLE.

As shown in Table 9.8, upgrading kinematics from the baseline to $\mathcal{L}_{K_{v1}}/\mathcal{L}_{K_{v2}}$ reduces MPJPE and P-MPJPE while consistently lowering MPJVE, indicating improved temporal stability. The orientation couplings exhibit a clear asymmetry: O_1 (full coupling) typically yields the strongest MPJAE across datasets but can slightly increase MPJVE, whereas O_2 (concatenate+detach) preserves or improves MPJVE with only a minor compromise in MPJAE. The anatomical term $\mathcal{L}_{\mathrm{B}}$ primarily benefits MPBLE and shows dataset-dependent interactions with kinematics and orientation, with the clearest improvements on H4D where motion and occlusion are most challenging. Qualitative results on H4D are shown in Figure 9.5. In line with Section 7.2.5, a conservative operating point is obtained with Baseline+$\mathcal{L}_{K_{v1}}$ (and optionally $\mathcal{L}_{\mathrm{B}}$) when temporal stability and positional accuracy are prioritized, while $\mathcal{L}_{K_{v1}} + O_1$ is preferable when minimizing MPJAE is paramount.

Table 9.8: Combinations of kinematics, anatomy, and orientation – Ablation on Fit3D, H36M, AP3D, and H4D. Metrics: MPJPE, P-MPJPE, MPJAE, MPJVE, MPBLE (all ↓). $\mathcal{L}_{K_{v1}}/\mathcal{L}_{K_{v2}}$: kinematic losses (flow+velocity; flow+velocity+acceleration) *c.f.* Section 7.1.2; O_1/O_2: orientation coupling (concat+full gradient; concat+detach) *c.f.* Section 7.2; $\mathcal{L}_{\text{B}}$: soft bone-length consistency *c.f.* Section 7.1.1. **Red** indicates the best score per column, while **blue** marks the second-best. Stronger kinematics reduce position and velocity errors; O_1 yields the lowest MPJAE, O_2 better preserves MPJVE; $\mathcal{L}_{\text{B}}$ improves MPBLE.

Method	Fit3D					H36M				
	MPJPE ↓	P-MPJPE ↓	MPJAE ↓	MPJVE ↓	MPBLE ↓	MPJPE ↓	P-MPJPE ↓	MPJAE ↓	MPJVE ↓	MPBLE ↓
Baseline	25.2	19.2	–	1.7	2.5	47.9	38.8	–	2.9	11.2
Baseline + O_1	25.6	19.3	9.9	1.8	1.9	48.1	39.1	16.9	3.0	11.2
Baseline + $\mathcal{L}_{K_{v1}}$ + O_1	24.6	18.7	9.7	1.5	1.9	47.5	38.5	17.3	2.2	11.2
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v1}}$ + O_1	25.0	18.9	9.8	1.5	1.9	47.7	38.6	17.1	2.2	11.1
Baseline + $\mathcal{L}_{K_{v2}}$ + O_1	24.7	18.7	9.7	1.5	2.0	47.7	38.7	17.3	2.2	11.3
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v2}}$ + O_1	25.0	19.0	9.8	1.5	2.2	47.5	38.5	17.1	2.2	11.0
Baseline + O_2	25.3	19.2	9.9	1.7	2.3	48.0	39.1	18.0	2.8	11.3
Baseline + $\mathcal{L}_{K_{v1}}$ + O_2	24.7	18.8	10.0	1.5	2.7	47.4	38.5	17.9	2.1	11.2
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v1}}$ + O_2	24.5	18.8	10.0	1.5	2.4	47.4	38.6	17.8	2.1	11.1
Baseline + $\mathcal{L}_{K_{v2}}$ + O_2	26.0	19.8	10.6	1.5	2.7	47.5	38.3	17.7	2.1	11.1
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v2}}$ + O_2	25.4	19.5	10.2	1.5	3.2	47.4	38.5	17.7	2.1	11.1
Method	AP3D					H4D				
	MPJPE ↓	P-MPJPE ↓	MPJAE ↓	MPJVE ↓	MPBLE ↓	MPJPE ↓	P-MPJPE ↓	MPJAE ↓	MPJVE ↓	MPBLE ↓
Baseline	16.3	12.6	–	2.4	3.8	64.8	34.6	–	10.7	3.5
Baseline + O_1	17.5	13.5	8.5	2.8	4.1	66.9	37.3	11.9	11.8	3.7
Baseline + $\mathcal{L}_{K_{v1}}$ + O_1	15.6	12.0	8.3	1.9	3.6	59.6	30.8	10.6	8.0	3.7
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v1}}$ + O_1	15.7	12.1	8.4	1.9	3.7	59.9	31.2	10.7	8.0	4.3
Baseline + $\mathcal{L}_{K_{v2}}$ + O_1	15.9	12.2	8.3	1.9	3.8	59.6	30.8	10.6	7.9	2.3
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v2}}$ + O_1	15.7	12.0	8.2	1.9	3.7	59.9	31.1	10.7	7.9	4.3
Baseline + O_2	16.6	12.9	9.9	2.4	3.9	65.4	34.5	12.5	10.7	3.8
Baseline + $\mathcal{L}_{K_{v1}}$ + O_2	15.4	11.8	9.5	1.8	3.5	58.9	30.9	11.7	7.8	3.7
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v1}}$ + O_2	15.4	11.9	9.6	1.8	3.5	60.3	32.0	11.8	8.2	3.6
Baseline + $\mathcal{L}_{K_{v2}}$ + O_2	15.4	11.9	9.5	1.8	3.6	58.7	30.8	11.7	7.8	2.4
Baseline + $\mathcal{L}_{\text{B}}$ + $\mathcal{L}_{K_{v2}}$ + O_2	15.4	11.9	9.6	1.8	3.5	59.1	30.5	11.7	7.8	3.0

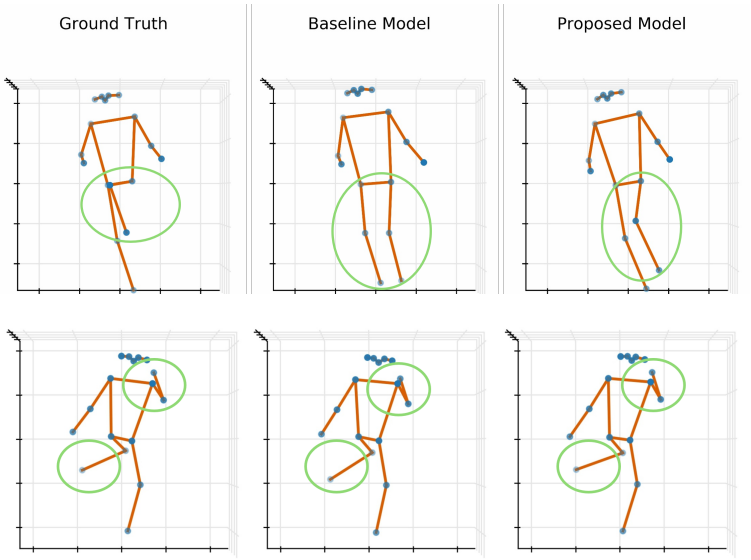


Figure 9.5: Qualitative results on H4D – Left: ground truth. Middle: predictions from the 3D-HPE baseline. Right: predictions from the proposed model. In the first row, the baseline fails to capture the knee motion, whereas the proposed model provides a closer approximation, though inaccuracies remain. In the second row, the baseline underestimates the kick movement and places the hand near the shoulder. The proposed model better resolves the leg orientation and achieves an almost accurate hand placement, although the leg is predicted slightly too long.

Guided by these ablations and the objectives in Chapter 7, two configurations are selected for subsequent comparisons and deployment: (i) Baseline+$\mathcal{L}_{K_{v1}}$ as the default model without orientations, due to its consistent reductions in MPJPE/P-MPJPE and MPJVE at negligible complexity; and (ii) Baseline+$\mathcal{L}_{K_{v1}}$+O_1 as the orientation-enabled model. As it achieves the strongest MPJAE across datasets with acceptable impact on positional and temporal errors. The variant with $\mathcal{L}_{\mathrm{B}}$ is retained as an optional add-on where bone-length stability is critical, but is not adopted as default due to dataset-dependent gains.

9.2.2 Comparison with the State-of-the-Art

This subsection benchmarks the proposed combinations against recent temporal transformer architectures on Fit3D, H36M, AP3D, and H4D. The focus lies on the balance between positional accuracy and kinematics, and on accuracy–efficiency trade-offs, *i.e.*, parameters, FLOPs, and runtime.

The comparison in Table 9.9 confirms that large-capacity temporal transformers (*e.g.*, TCPFormer, KTPFormer, MotionAGFormer, DSTFormer) obtain the lowest MPJPE/P-MPJPE. The combinations with $\mathcal{L}_{K_{v1}}$+O_1 achieve competitive MPJAE/MPJVE at substantially lower complexity, reflecting the complementary role of kinematic regularization and lightweight orientation coupling. This suggests that temporal refinement disproportionately benefits orientation and motion smoothness relative to absolute positional error, which is increasingly saturated by large models.

Table 9.9: State-of-the-art comparison on Fit3D/H36M and AP3D/H4D – Metrics: MPJPE, P-MPJPE, MPJAE, MPJVE, MPBLE (all ↓). Missing entries indicate unavailable reports in the original works. **Red** indicates the best score per column, while **blue** marks the second-best. Large temporal transformers attain the lowest positional errors; configurations with $\mathcal{L}_{K_{v1}}$+O_1 are competitive on kinematic metrics at substantially lower complexity.

Method	Fit3D					H36M				
	MPJPE ↓	P-MPJPE ↓	MPJAE ↓	MPJVE ↓	MPBLE ↓	MPJPE ↓	P-MPJPE ↓	MPJAE ↓	MPJVE ↓	MPBLE ↓
PoseFormer [Zhe21]	25.2	19.2	–	1.7	2.5	47.9	38.8	–	2.9	11.2
PoseFormerv2 [Zha23b]	25.3	19.0	–	1.7	2.8	48.2	39.5	–	2.7	11.6
MixSTE [Zha22]	25.2	19.2	–	1.6	3.2	45.4	37.4	–	2.4	11.9
MHFormer [Li22a]	33.2	25.6	–	2.2	3.1	52.1	42.1	–	3.4	14.6
UU [Ein23]	25.9	19.6	–	2.3	3.4	48.0	39.4	–	3.8	11.9
DSTFormer [Zhu23]	24.1	18.1	–	1.5	2.5	**43.4**	**36.0**	–	2.3	11.1
KTPFormer [Pen24]	24.5	18.3	–	**1.3**	**2.1**	**43.0**	**35.5**	–	**1.8**	**10.6**
TCPFormer [Liu25b]	**23.5**	**17.4**	–	1.7	2.7	44.5	36.6	–	2.6	10.9
MotionAGFormer [Meh24]	**23.6**	**17.6**	–	1.5	2.5	44.5	36.7	–	2.3	**10.4**
Baseline + $\mathcal{L}_{K_{v1}}$ (ours)	24.7	18.8	–	**1.5**	2.4	47.5	38.7	–	**2.1**	11.1
Baseline + $\mathcal{L}_{K_{v1}}$ + O_1 (ours)	24.6	18.7	**9.7**	1.5	**1.9**	47.5	38.5	**17.3**	2.2	11.2
Method	AP3D					H4D				
	MPJPE ↓	P-MPJPE ↓	MPJAE ↓	MPJVE ↓	MPBLE ↓	MPJPE ↓	P-MPJPE ↓	MPJAE ↓	MPJVE ↓	MPBLE ↓
PoseFormer [Zhe21]	16.3	12.6	–	2.4	3.8	64.8	34.6	–	10.7	3.5
PoseFormerv2 [Zha23b]	17.8	14.0	–	2.7	4.2	64.1	34.0	–	10.3	3.0
MixSTE [Zha22]	14.6	11.0	–	2.1	3.6	50.4	25.0	–	7.6	2.8
MHFormer [Li22a]	33.5	25.4	–	4.7	9.5	98.9	58.7	–	16.1	6.6
UU [Ein23]	17.5	13.7	–	2.8	4.3	64.3	37.1	–	11.7	6.6
DSTFormer [Zhu23]	12.6	9.2	–	1.6	2.8	47.8	22.4	–	6.8	2.5
KTPFormer [Pen24]	12.6	9.1	–	**1.4**	2.9	45.6	**20.6**	–	**5.8**	**2.5**
TCPFormer [Liu25b]	**11.3**	**8.2**	–	**1.5**	**2.5**	**41.5**	**19.1**	–	**6.2**	2.6
MotionAGFormer [Meh24]	**12.0**	**8.8**	–	1.6	**2.6**	**44.9**	20.8	–	6.5	**1.9**
Baseline + $\mathcal{L}_{K_{v1}}$ (ours)	15.4	11.8	–	1.8	3.6	58.4	30.5	–	7.8	3.7
Baseline + $\mathcal{L}_{K_{v1}}$ + O_1 (ours)	15.6	12.0	**8.3**	1.9	3.6	59.6	30.8	**10.6**	8.0	3.7

9.2.2.1 Runtime and Complexity

This subsection reports model complexity and per-sequence latency on H4D and examines sequence-length scaling. All runtime measurements are conducted on an NVIDIA L40 GPU with batch size 1, *i.e.*, single video per forward pass.

As summarized in Table 9.10, compact baselines (*e.g.*, PoseFormer, UU) provide favorable latency but fall short in accuracy, whereas high-capacity transformers (*e.g.*, TCPFormer, KTPFormer) achieve lower MPJPE at the expense of a substantial increase in parameters, FLOPs, and runtime. By contrast, the proposed models consistently yield the lowest per-sequence latency across all settings while improving accuracy over compact baselines. The Baseline+$\mathcal{L}_{K_{v1}}$ matches the smallest models in terms of parameters and FLOPs, yet reduces runtime to 4.0 ms per sequence—the best result overall—while improving MPJPE. Adding O_1 preserves this runtime advantage with negligible FLOP overhead and provides strong gains in MPJAE.

Table 9.10: Runtime and complexity on H4D (default $T = 27$) – Parameters (M) and FLOPs (G) are computed at the evaluation resolution; runtime is per-sequence latency (batch size 1) measured on an NVIDIA L40 GPU. **Red** indicates the best score per column, while **blue** marks the second-best. Compact baselines provide low latency but lower accuracy; larger transformers minimize MPJPE at significantly higher cost; the selected configurations maintain near-baseline latency while improving MPJAE.

Method	Parameters (M)↓	FLOPs (G)↓	Runtime (ms)↓	MPJPE (mm)↓	MPJAE (∘)↓
PoseFormer [Zhe21]	**9.6**	**4.4**	4.2	64.8	–
PoseFormerv2 [Zha23b]	14.4	8.7	–	64.1	–
MixSTE [Zha22]	16.9	124.7	11.7	50.4	–
MHFormer [Li22a]	19.2	8.4	9.9	98.9	–
UU [Ein23]	12.9	**2.4**	4.3	64.3	–
DSTFormer [Zhu23]	15.9	118.3	15.6	47.8	–
KTPFormer [Pen24]	34.8	257.6	25.3	45.6	–
TCPFormer [Liu25b]	35.0	–	79.5	**41.5**	–
MotionAGFormer [Meh24]	11.7	88.2	36.3	**44.9**	–
Baseline + $\mathcal{L}_{K_{v1}}$(ours)	**9.6**	**4.4**	**4.0**	58.4	–
Baseline + $\mathcal{L}_{K_{v1}}$ + O_1)(ours)	**9.9**	4.6	**4.1**	59.6	**10.6**

A sequence of 27 frames corresponds to a short temporal window, approximately one second of video. In surveillance scenarios, however, input horizons for prediction typically span between one and ten seconds, depending on camera placement and scene dynamics. The scaling experiment in Table 9.11 shows that extending the sequence length T (*e.g.*, from 27 to 64 frames) generally enhances the accuracy of temporal models but incurs a substantial increase in computational cost. In contrast, the proposed configurations retain their status as the fastest models in all comparisons: both Baseline+$\mathcal{L}_{K_{v1}}$ and Baseline+$\mathcal{L}_{K_{v1}}$+O_1 introduce only marginal overhead while maintaining competitive MPJPE and achieving improved MPJAE. This combination of accuracy and consistently superior runtime defines an efficient operating point for applications subject to strict latency constraints.

Table 9.11: Runtime and complexity on H4D for $T = 64$ – Parameters (M), FLOPs (G), and per-sequence latency (batch size 1) measured on an NVIDIA L40 GPU. Rows indicate the evaluated sequence length (*e.g.*, $T = 27$ vs. $T = 64$). **Red** indicates the best score per column, while **blue** marks the second-best. Increasing T generally improves accuracy but increases cost markedly; the selected configurations show modest overhead with competitive MPJPE and improved MPJAE.

Method	Parameters (M)↓	FLOPs (G)↓	Runtime (ms)↓	MPJPE (mm)↓	MPJAE (∘)↓
UU [Ein23] ($T = 27$)	12.9	**2.4**	4.3	64.3	–
UU [Ein23] + O_1 ($T = 27$)	56.3	22.3	5.8	67.4	11.7
DSTFormer [Zhu23] ($T = 27$)	15.9	118.3	15.6	**47.8**	–
DSTFormer [Zhu23] + O_1 ($T = 27$)	92.8	151.5	17.3	**49.6**	**8.0**
Ours ($T = 27$) (Baseline + $\mathcal{L}_{K_{v1}}$)	**9.6**	**4.4**	**4.0**	58.4	–
Ours ($T = 27$) (Baseline + $\mathcal{L}_{K_{v1}}$ + O_1)	**9.9**	4.6	**4.1**	59.6	**10.6**
UU [Ein23] ($T = 64$)	12.9	**5.9**	4.6	61.9	–
UU [Ein23] + O_1 ($T = 64$)	56.3	52.9	6.9	65.2	11.6
DSTFormer [Zhu23] ($T = 64$)	16.0	283.8	38.5	**46.4**	–
DSTFormer [Zhu23] + O_1 ($T = 64$)	92.8	362.4	43.5	**47.4**	**8.3**
Ours ($T = 64$) (Baseline + $\mathcal{L}_{K_{v1}}$)	**9.6**	**10.7**	**4.0**	56.5	–
Ours ($T = 64$) (Baseline + $\mathcal{L}_{K_{v1}}$ + O_1)	**10.0**	11.1	**4.5**	56.3	**10.7**

9.2.3 Results and Discussion

The evaluations demonstrate that model-agnostic kinematic constraints and soft anatomical consistency jointly improve root-relative 3D-HPE across diverse datasets. Kinematic supervision ($\mathcal{L}_{K_{v1}}$, $\mathcal{L}_{K_{v2}}$) consistently reduces

MPJVE and improves MPJPE/P-MPJPE; the anatomical term $\mathcal{L}_{\mathrm{B}}$ enhances MPBLE where needed. A unified orientation head with quaternion supervision enables simultaneous prediction of joint positions and orientations. The coupling O_1 achieves the best MPJAE with minor trade-offs in temporal smoothness, whereas O_2 preserves MPJVE at slightly lower orientation accuracy. Based on these findings, Baseline+$\mathcal{L}_{K_{v1}}$ is adopted as the default model without orientations, and Baseline+$\mathcal{L}_{K_{v1}}$+O_1 as the orientation-enabled counterpart.

The thesis introduces (i) a principled, lightweight set of temporal kinematic losses tailored to root-relative 3D-HPE that are effective across datasets, (ii) a soft bone-length regularizer robust to annotation noise, and (iii) a multitask formulation for joint orientation prediction with quaternion geodesic supervision and controlled coupling to coordinates. Comprehensive ablations and state-of-the-art comparisons highlight favorable accuracy–efficiency trade-offs, particularly on H4D where fast, occlusion-heavy motion is prevalent. In all runtime evaluations, the proposed models consistently achieve the lowest per-sequence latency, outperforming compact baselines while avoiding the computational burden of high-capacity transformers. The selected configurations therefore combine robust kinematics, competitive positional accuracy, and consistently superior runtime, addressing the core requirements of surveillance-oriented human analysis under strict latency constraints.

9.3 Skeleton-based Action Recognition

This section presents a controlled comparison of SBAR architectures on UAVHuman (CSv1) and Smarthome (CS) under a unified protocol. Baselines operate on ground-truth 2D keypoints with pseudo-depth $(x_{2D}, y_{2D}, 1)$ to isolate recognition capacity under surveillance-style inputs. The proposed framework first uplifts inputs to predicted 3D $(\hat{x}_{3D}, \hat{y}_{3D}, \hat{z}_{3D})$ and subsequently applies a 4-stream CTR-GCN. Results are reported only on Smarthome due to the absence of 3D ground truth on UAVHuman for in-domain uplifter training. All inputs are root-centered, scale-normalized, and temporally aligned. No data augmentation is used. Metrics are ACC and

MAPCA. Parameter counts, FLOPs, and runtime are reported per stream to account for differing stream multiplicities across architectures. Runtimes are measured on an NVIDIA L40 GPU, and effective cost scales with the number of active streams.

9.3.1 Experimental Setup

Baselines consume GT 2D keypoints with pseudo-depth $(x_{2D}, y_{2D}, 1)$ to factor out variability from 2D–3D lifting and to probe recognition capacity directly on surveillance-style skeletons. In contrast, the proposed framework uplifts inputs to root-relative $(\hat{x}_{3D}, \hat{y}_{3D}, \hat{z}_{3D})$ using the in-domain lifting setup from Chapter 7, with kinematic regularization as described in Section 7.1.2. The same CTR-GCN backbone is then employed with four streams (Joints, Bones, Joint-Motion, Bone-Motion). This 3D variant is evaluated only on Smarthome, since UAVHuman lacks 3D ground truth for in-domain uplifter training. Cross-dataset lifting is not considered to avoid confounding effects (*c.f.* Chapter 7).

Architectural designs differ mainly in the number of streams at inference: CTR-GCN and the proposed framework use four streams, HD-GCN employs six hierarchical streams, HybridFormer processes two streams, Hyperformer maintains four, and MotionBert relies on a single stream. To ensure comparability, parameters, FLOPs, and runtime are reported per stream, with the total cost approximated by scaling to the active stream count and, for the proposed framework, adding uplifter latency. Each method is evaluated with its official topology and joint ordering. Training schedules—including optimizer, batch size, learning-rate policy, and number of epochs are harmonized across methods, as are regularization strategies such as weight decay and dropout. Sequence sampling follows the conventions of public implementations, and no data augmentation is applied, in order to avoid introducing method-specific tuning effects.

9.3.2 Comparison with the State-of-the-Art

Results are summarized in Table 9.12. On UAVHuman (CSv1), compact graph-convolutional baselines (CTR-GCN, HD-GCN) deliver the strongest accuracy under GT 2D pseudo-depth inputs: HD-GCN attains the highest ACC/MAPCA (46.2%/46.1%), while CTR-GCN is second (45.2%/45.2%). Higher nominal capacity or heavier per-stream compute (HybridFormer, MotionBert) does not translate into superior recognition accuracy in this regime, indicating that relational reasoning on the kinematic graph with modest temporal capacity is advantageous when inputs are noisy, occluded, or truncated. On Smarthome (CS), HD-GCN achieves the best ACC (78.8%) and shares the best MAPCA (59.1%) with CTR-GCN. The proposed framework—which uplifts to $(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$ and applies the same 4-stream CTR-GCN—obtains comparable performance (78.5% ACC, 58.5% MAPCA). Overall, reliable depth cues can be exploited within the same backbone to match the performance of strong graph-based methods.

Table 9.12: State of the art on SBAR—UAVHuman (CSv1) and Smarthome (CS) with GT 2D pseudo-depth $(x_{2D},y_{2D},1)$. The Proposed Framework uplifts to $(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$ and applies a 4-stream CTR-GCN (reported only on Smarthome). Params/FLOPs/runtime are per stream (runtime on NVIDIA L40). For the Proposed Framework, the computational complexity (Params/FLOPs/runtime) is reported as the combined cost of the uplifter and the CTR-GCN. Metrics: ACC, MAPCA.

Method	Params (M)↓	FLOPs (G)↓	Runtime (ms)↓	UAVHuman (%) ↑		Smarthome (%) ↑	
				ACC	MAPCA	ACC	MAPCA
CTR-GCN [Che21b]	1.4	1.2	10.8	45.2	45.2	78.2	59.1
HD-GCN [Lee23]	1.5	1.2	49.4	46.2	46.1	78.8	59.1
HybridFormer [Zho25]	2.6	2.0	9.9	39.8	39.7	73.9	49.1
Hyperformer [Zho22b]	2.6	3.3	12.0	42.6	42.6	76.9	55.3
MotionBert [Zhu23]	60.3	92.1	10.4	36.1	36.3	70.9	49.9
Proposed Framework	16.0 + 1.4	17.3 + 1.2	5.0 + 11.0	–	–	78.5	58.5

Several trends are consistent across datasets. First, under surveillance-style GT 2D inputs with pseudo-depth, localized spatial priors paired with moderate temporal modeling (*e.g.*, CTR-GCN/HD-GCN) remain robust to occlusion and viewpoint changes. Second, greater model capacity or higher per-stream compute is not sufficient for gains when the input signal is degraded.

Accuracy–efficiency trade-offs depend more on stream design and graph priors than on raw FLOPs. Third, the 3D uplifted pipeline demonstrates that stable root-relative sequences—regularized by kinematics—can be consumed by off-the-shelf graph backbones without architectural changes, yielding competitive accuracy with an additional uplifter stage and multi-stream inference.

Per-stream latency and FLOPs should be interpreted jointly with stream count. HD-GCN's higher runtime per stream yields only a marginal accuracy improvement over CTR-GCN. HybridFormer is comparatively fast per stream but incurs a notable accuracy cost. Because parameters/FLOPs/runtime are reported per stream, deployment can adjust the number of active streams to fit resource budgets. Total FLOPs and energy scale approximately linearly with stream multiplicity, and peak memory scales with the number of concurrent streams. Latency, however, is not necessarily multiplicative: with parallel execution on a single accelerator, the graph stage's end-to-end latency is close to the maximum per-stream latency (*e.g.*, 4 CTR-GCN streams at $\approx$ 11 ms remain $\approx$ 11 ms given sufficient concurrency), rather than the sum across streams. If streams are scheduled sequentially (due to memory or scheduling limits), latency is additive. The uplifter contributes $\approx$ 5 ms. Depending on overlap with the graph stage, the end-to-end latency approximates the slowest stage or the sum of non-overlapping stages.

9.3.3 Results and Discussion

The study indicates that (i) graph-based SBAR backbones set a strong baseline under surveillance-style 2D skeletons; (ii) uplifted 3D can be integrated seamlessly into the same backbone to obtain comparable class-wise performance when an in-domain uplifter is available; and (iii) efficiency is governed more by stream multiplicity than by per-stream FLOPs. These findings align with the 3D-HPE conclusions in Chapter 7: temporally-aware, root-relative lifting with kinematic regularization produces stable 3D sequences that can be consumed by standard graph-based recognizers without task-specific modifications, while maintaining a controllable complexity profile measured per

stream. Reporting both per-stream and end-to-end latency (uplifter + streams) provides a complete accounting for deployment.

9.4 Summary

This chapter provided a unified evaluation of image- and sequence-based human analysis in surveillance-oriented settings, covering 2D human pose estimation (2D HPE), root-relative 3D human pose estimation (3D HPE) with anatomical/kinematic regularization and joint orientations, and skeleton-based action recognition (SBAR). The focus lay on accuracy, cross-domain robustness, and accuracy–efficiency trade-offs under a top-down assumption.

Contributions are threefold. i) A surveillance-oriented cross-domain evaluation protocol for 2D-HPE and PAR, with controlled ablations on capacity sharing, task adapters, and data augmentation to quantify robustness–efficiency trade-offs under domain shift. (ii) A lightweight kinematic supervision K_{v1} and a soft anatomical term B for 3D-HPE that consistently reduce MPJVE and improve MPBLE, together with an orientation head O_1 trained with a quaternion loss that lowers MPJAE while maintaining competitive positional/temporal errors relative to larger temporal transformers. (iii) A 3D-lifted SBAR pipeline that uplifts $(x_{2D}, y_{2D}, 1)$ to $(\hat{x}_{3D}, \hat{y}_{3D}, \hat{z}_{3D})$ and feeds a 4-stream CTR-GCN.

For 2D HPE (Section 9.1), four observations emerged. First, multitask learning with a shared encoder consistently benefited PAR, whereas 2D HPE experienced a moderate performance drop under cross-domain evaluation. Task-specific adapters mitigated this in low-capacity settings, while larger backbones alleviated interference more effectively. Second, data augmentation improved cross-protocol robustness on both UPAR-Pose and UPPET at the cost of a small within-domain accuracy reduction, a trade-off typical for regularization in surveillance domains. Third, ViTPose-Huge achieved the highest within-domain accuracy under specialization, while shared-encoder variants offered competitive accuracy with lower computational cost per

person crop. Fourth, runtime on NVIDIA L40 indicated that shared-encoder designs yielded favorable speed–accuracy trade-offs in multitask scenarios.

For 3D-HPE (Section 9.2), the introduction of kinematic constraints (*c.f.* Section 7.1.2) consistently reduced MPJVE across Fit3D, H36M, AP3D, and H4D, while also improving MPJPE and P-MPJPE by stabilizing temporal dynamics. The soft anatomical term B (*c.f.* Section 7.1.1) further enhanced MPBLE, with dataset-dependent effects. Acceleration constraints yielded only marginal additional gains and proved more sensitive to annotation noise. Based on these observations, Baseline+K_{v1} was adopted as the default model without orientations, whereas Baseline+K_{v1}+O_1 was selected as the orientation-enabled variant, offering the best MPJAE with only minor trade-offs in positional and temporal metrics. In comparison to recent temporal transformers, the proposed configurations remain competitive in kinematic accuracy while operating at substantially lower complexity. Moreover, sequence-length scaling confirmed that accuracy remains stable under modest overhead, underscoring their suitability for efficient deployment in real-world scenarios such as H4D.

For SBAR (Section 9.3), a controlled protocol on UAVHuman (CSv1) and Smarthome (CS) showed that graph-based backbones (CTR-GCN, HD-GCN) remained highly competitive under surveillance-style 2D skeletons with pseudo-depth. The 3D-lifted variant—uplifting to root-relative $(\hat{x}_{3D},\hat{y}_{3D},\hat{z}_{3D})$ and applying the same 4-stream CTR-GCN—achieved comparable accuracy and class-wise retrieval on Smarthome (78.5% ACC, 58.5% MAPCA), demonstrating that stable 3D sequences produced by the kinematic setup can be consumed by standard recognizers without architectural changes. Per-stream reporting clarified that FLOPs and energy scale approximately linearly with stream multiplicity, while end-to-end latency depends on concurrency: with parallel execution on a single accelerator, the graph stage's latency approaches the maximum per-stream latency (*e.g.*, 4 CTR-GCN streams at $\approx$ 11 ms remain $\approx$ 11 ms given sufficient concurrency), whereas sequential scheduling makes latency additive. The uplifter contributed $\approx$ 5 ms and dominated the additional overhead relative to 2D-only baselines.

Overall, these findings support three practical conclusions. (i) For 2D-HPE, robustness under domain shift is enhanced primarily by capacity and DA. Multitask learning with PAR is advantageous for attributes and efficient for joint inference, with adapters or larger backbones mitigating task interference. (ii) For 3D-HPE, lightweight kinematic supervision substantially improves temporal stability with favorable accuracy–efficiency trade-offs. Orientation prediction via a shared backbone and quaternion loss enhances interpretability and benefits tasks sensitive to rotational cues. (iii) For SBAR, graph-based backbones with controlled stream multiplicity provide strong performance on surveillance-style skeletons. Uplifted 3D can be integrated when in-domain lifting is available, without altering the relative ordering of competitive methods.

10 Human Action Recognition System

The objective of this chapter is to design a dataset that enables systematic evaluation of a full human action recognition pipeline under realistic surveillance conditions. A central motivation is to insert motion data captured from real people into the GTA V [Roc13] simulation environment, thereby bridging controlled experimental datasets and complex virtual worlds. This integration allows realistic yet fully controllable scenes that mimic surveillance perspectives while preserving precise ground-truth annotations.

To the best of the author's knowledge, prior to GTA-RWS no dataset enabled the evaluation of a complete surveillance-oriented pipeline—from 2D keypoint detection, through 3D skeleton uplifting, to skeleton-based action recognition—in an end-to-end manner. Existing resources have addressed individual components of the pipeline, but lacked the integration of realistic scene context, external motion sources, and systematic ground-truth supervision required for holistic benchmarking.

GTA-RWS closes this gap by providing a unified framework in which each stage of the pipeline can be trained, tested, and analyzed under surveillance-specific conditions. This chapter therefore goes beyond dataset construction to demonstrate how the resource supports end-to-end analysis of a human action recognition system. The pipeline progresses from 2D keypoint detection, through 3D skeleton uplift, to skeleton-based action recognition, allowing the propagation of errors to be quantified at each stage. Such a stepwise evaluation highlights both the robustness and the vulnerability of recognition under realistic surveillance conditions, providing insight into how early inaccuracies in detection or pose estimation impact the reliability of downstream recognition. In this way, the GTA-RWS dataset serves not only as a resource

for training but also as a testbed for rigorous, system-level benchmarking of human action recognition.

The chapter is organized as follows. Section 10.1 introduces the GTA-RWS dataset and its construction principles. Section 10.1.2 presents task-specific subsets tailored for 2D-HPE, 3D-HPE, and SBAR. Section 10.2 reports the end-to-end system evaluation, highlighting the propagation of errors across the pipeline. Finally, Section 10.3 concludes with a discussion of the key findings and implications for real-world human action recognition systems.

10.1 GTA-RWS Dataset

The GTA-RWS dataset operationalizes this idea by combining realism with controllability in a structured resource tailored for surveillance research. Beyond integrating external motion sources, GTA-RWS provides diverse environments, camera setups, and action classes, ensuring reproducibility and facilitating system-level evaluation across 2D-HPE, 3D pose uplifting, and skeleton-based action recognition.

10.1.1 Construction of the dataset

Scenes are authored to resemble surveillance footage, featuring elevated camera placements, downward pitch, varied fields of view, and diverse urban, industrial, and indoor environments that dynamically respond to time and weather. GTA V is chosen as the simulation environment, providing a rich virtual world with detailed assets, dynamic illumination and weather, extensive character models and clothing combinations, and a large built-in animation library with clipsets and blending. The engine further offers realistic pathfinding and physics (including ragdoll), as well as mature modding interfaces that allow insertion of custom motions, scripting of behaviors, and extraction of ground truth [Unk25a, Unk25b, Fab18, Fab21].

Motion diversity is extended beyond ambient behaviors by integrating curated one- and two-person sequences from Inter-X [Xu24], H4D[Khi24b] (*c.f.*

Section 4.2.1.5), and BlindWays [Kim24]. Inter-X provides high-quality, multi-view 3D human interactions, focusing on complex two-person behaviors such as hugging and handshakes. H4D contributes densely annotated sequences of fighting sports, capturing dynamic interactions, strikes, and grappling motions that resemble salient events in real-world surveillance, such as altercations or assaults. BlindWays supplies motion data specifically recorded for visually impaired individuals, representing everyday behaviors that should not be misclassified as salient. These external motions are complemented with selected in-game animations for classes not covered in external sources, such as calling for help or gesturing. Together, these sources allow the evaluation of highly dynamic and challenging actions, including fights, which are critical for safety- and security-oriented applications.

Actions are annotated using a coarse taxonomy of sixteen classes (*c.f.* Table 10.1) to emphasize surveillance-relevant events while supporting cross-dataset mapping and enabling robust evaluation. This combination of external motion capture and in-game animation ensures a diverse and realistic set of behaviors suitable for end-to-end benchmarking of surveillance-oriented human action recognition pipelines.

Table 10.1: GTA-RWS action taxonomy – The taxonomy defines 16 coarse-grained action classes. Each class is assigned a unique ID that is consistent across annotations and evaluations. Additionally, actions are categorized as *Normal* (do not trigger a notification) or *Salient* (trigger a notification).

ID	Action Class	Notification Category
0	Miscellaneous	Normal
1	Standing	Normal
2	Sitting	Normal
3	Walking	Normal
4	Carrying Object	Normal
5	Pushing Object	Normal
6	Bicycle	Normal
7	Greeting	Normal
8	Hugging	Normal
9	Dancing	Normal
10	Call for Help	Salient
11	Running	Normal
12	Stumbling	Normal
13	Fighting	Salient
14	Lying (+ Falling)	Salient
15	Grappling (+ Strangling)	Salient

Heterogeneous motion sources provide joint trajectories in different skeletal topologies and coordinate conventions, as shown in Figures 10.1 and 10.2. Poses are canonicalized to a right-handed, pelvis-centered frame with consistent scale and resampling, then mapped into a unified target topology compatible with the engine. Shared joints are transferred directly, and missing landmarks are estimated along anatomically plausible axes, such as placing intermediate spine joints between hip and shoulder centers. When richer extremity data is available—for example, SMPL-X finger and forefoot markers or X-Sens foot and cane markers, as shown in Figure 10.2—these additional landmarks are temporarily incorporated into the target topology to better guide realistic wrist and ankle orientations.

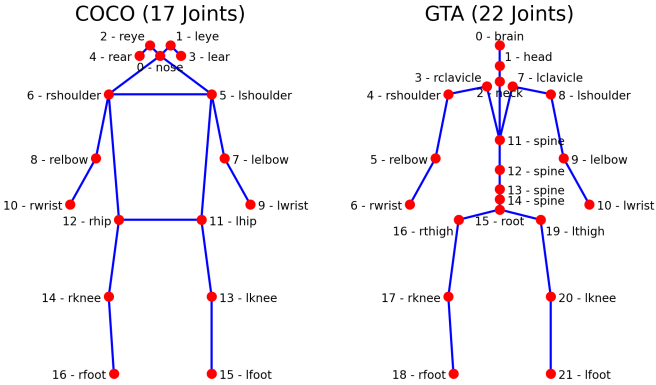


Figure 10.1: Comparison of the COCO (H4D) and GTA skeleton topologies – The COCO topology defines 17 keypoints covering major body joints. The GTA topology expands this to 22 keypoints by adding finer torso structure—multiple spine segments and explicit left/right clavicles—while keeping similar limb coverage.

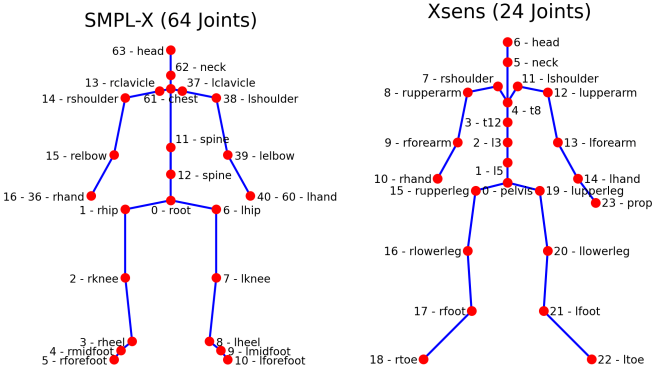


Figure 10.2: Comparison of the SMPL-X (Inter-X) and X-Sens (BlindWays) skeleton topologies – The SMPL-X model includes 64 joints in total, with 20 joints per hand and 3 per foot. In contrast, the X-Sens model contains only 24 joints, with a single joint per hand, one for an external prop (*e.g.*, a cane), and two per foot.

The engine consumes rotation-based animations with a global root translation. Retargeting converts the canonicalized joint trajectories into per-joint

quaternions in a top-down manner, aligning each bone with its parent in the parent's local frame. A robust root orientation is derived from the hip plane to maintain consistent global alignment. Temporal smoothing and careful quaternion sign handling reduce jitter and discontinuities. Bone lengths are inherited from the character models, allowing natural variation in proportions across subjects. The resulting animation streams are exported in the native YCD format [Unk25a] and visually verified against the source motions to ensure kinematic fidelity.

Scene composition is automated through screenplay blueprints executed via JTA-Mods [Fab18]. Thirty camera placements are selected across indoor, urban, and industrial locations, mounted at heights between approximately two and five meters with downward pitch angles typical of surveillance installations. Fields of view are sampled from a clipped normal distribution to reflect realistic optics without over-emphasizing extreme settings. Time of day and weather are randomized, eliciting environmental responses such as puddles, fog, and artificial lighting at night, and inducing locomotion changes in rain. An example of blueprint is provided in Figure 10.3. Each blueprint defines action points where targeted behaviors of a Pedestrian (PED) occur, group points that instantiate local crowds performing idle or scenario animations to create structured occlusions, and commute points that generate bidirectional flows traversing the field of view, resulting in edge truncations and partial visibility.

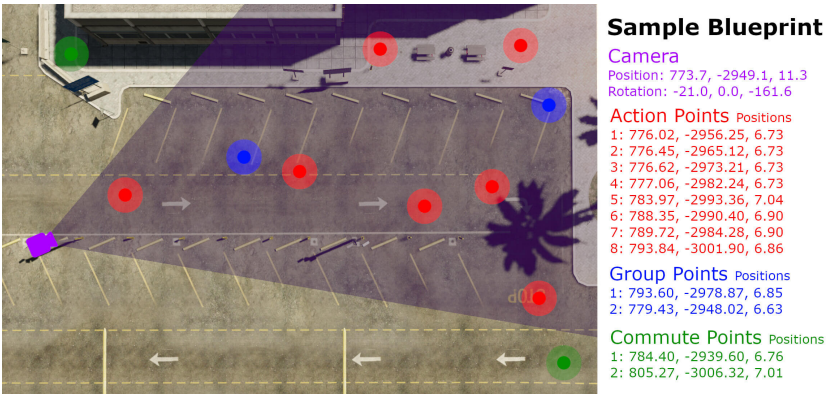


Figure 10.3: Sample blueprint of a screenplay – In addition to the position and orientation of the camera, a blueprint has coordinates for 12 PEDs spawn points. Eight of these points are referred to as *action points*, two are *group points* and the last two are *commute points*

The GTA V tasks (Stand, Wander, Phone, Scenario, Combat, Move) are reused and extended, and novel tasks are introduced to cover classes missing from built-ins, including single- and two-person curated animation playback (Figures 10.4 and 10.5) with randomized start times and spatial jitter, bicycling at variable speeds (Figure 10.6), stumbling via a drunk clipset (Figure 10.7), two-handed and one-handed carrying with props (Figures 10.8 and 10.9), pushing carts or bins with collisions enabled (Figure 10.10), and falls or lying via ragdoll with several stumble variants (Figure 10.11). In BlindWays scenes, a cane is attached to the hands so that orientation cues match the original motions (Figure 10.4 left).

Figure 10.4: One-person animations in GTA-RWS – Example images of PEDs performing single-person actions. The left panel shows a walking motion sourced from BlindWays [Kim24] (normal), while the right panel depicts a waving animation implemented in GTA V (salient).

Figure 10.5: Two-person animations in GTA-RWS – Example of PEDs engaging in paired interactions. The left panel shows a hugging animation from Inter-X [Xu24](normal) with a limited realism, while the right panel illustrates a fighting sequence from Harmony4D [Khi24b] (salient). These interactions highlight the capacity of GTA-RWS to reproduce complex multi-person behaviors for training and evaluation.

Figure 10.6: Bicycle animations in GTA-RWS – Example frames of PEDs riding bicycles within the simulation environment (normal). These clips extend coverage of everyday activities beyond pedestrian-only motions, enriching the diversity of the action recognition benchmark.

Figure 10.7: Drunk motion animations in GTA-RWS – Example of PEDs using the drunk (stumbling) clipset available in the simulation (normal). This animation illustrates atypical gait dynamics, which are valuable for assessing robustness of surveillance models to irregular or impaired motions, while still learning to not recognize these as salient.

Figure 10.8: Object-carrying animations in GTA-RWS – Example of PEDs wandering while holding objects other than boxes (normal). These animations contribute to the dataset's variety of human-object interactions, enabling better generalization in surveillance action recognition.

Figure 10.9: Box-carrying animations in GTA-RWS – Example of PEDs transporting boxes while wandering (normal). Object-carrying behaviors are central to realistic human activity datasets, allowing recognition systems to disambiguate between neutral locomotion and purposeful interactions with objects.

Figure 10.10: Object-pushing animations in GTA-RWS – Example of PEDs pushing objects within the environment (normal). This type of physical interaction is less common in standard action datasets, making it an important contribution of the GTA-RWS pipeline.

Figure 10.11: Ragdoll animations in GTA-RWS – Example of PEDs entering a ragdoll state after collapse (salient). These non-standard motions capture uncontrolled dynamics, simulating accidents or physical impacts, which are highly relevant in security- and safety-critical surveillance applications.

Scenes are recorded at a fixed frame rate of 20 frame per second. For every frame, synchronized outputs are saved: the RGB image, camera intrinsics and extrinsics, per-person 3D joints in world coordinates, 2D projections in image coordinates, and per-joint visibility determined by line-of-sight ray casting. Visibility flags distinguish self-occlusion from occlusion by other entities and encode the final state (visible, occluded, not visible).

Per-frame action labels are assigned by querying the active behaviors and animation clips. A hierarchical precedence scheme is applied to resolve conflicts: alert-relevant actions (*e.g.*, grappling, fighting) take highest priority, followed by curated animation classes. Next are vehicle states (e.g., bicycling), sitting (from curated animations or scenario context), lying or falling (*e.g.*, ragdoll, prone, or getting up), carrying and pushing, stumbling (from drunk clipsets), and running. Walking and standing are assigned if none of the higher-priority conditions apply. Frames that do not match any of these categories are labeled as Miscellaneous.

The dataset comprises 1,800 sequences of 12 seconds each, recorded at a fixed frame rate with a sequence length of 240 frames and authored from 30 blueprints/locations. The training split contains 1,260 sequences from 21 locations. The test split contains 540 sequences from 9 disjoint locations. Across all scenes, 102,030 pedestrians are included, drawn from 174 unique character models and 53,918 visually unique appearances (model plus component combinations), with 36,307 males and 65,723 females. Scene density ranges from 0 to 111 subjects, with an average of 56.68 per scene. In total, 25,074,933 unique poses are recorded. The dataset spans 1,347 unique combinations of location, time, and weather, and 1,254 unique combinations of camera position and field of view. Across all sequences, 3,811 unique animations are played. As illustrated in Figure 10.12, the scenes cover diverse environments and crowd levels that are important for realistic occlusions and viewpoint generalization.

Figure 10.12: Diverse environments in GTA-RWS – Sample images of PEDs in diverse environments and crowd level. The joint visibility is indicated. Green joints are visible in the frame, while red joints are occluded due to self-occlusion or scene geometry. These annotations demonstrate the detailed joint-level labeling provided by GTA-RWS, even under challenging viewing conditions.

10.1.2 Task-specific Subsets

The scale and diversity of GTA-RWS allow for the construction of task-specific subsets without exhausting scene variability. In this section, the focus lies on the three components studied previously: 2D-HPE, root-relative 3D-HPE, and SBAR. Additional subsets could be derived in future work, *e.g.*, for detection and tracking, global 3D-HPE, or open-set action recognition. All subsets preserve the train/test protocol of disjoint locations and camera placements to enforce generalization to unseen viewpoints.

10.1.2.1 2D Human Pose Estimation Subset

The 2D-HPE subset is constructed by downsampling each sequence to 30 frames and including every person with at least eight visible joints in the 17-keypoint layout. As shown in Figure 10.13, the topology is reduced to 17 keypoints since some keypoints are visually difficult to distinguish, *e.g.*, the spine keypoints. This procedure produces surveillance-style crowd scenes with high density, frequent occlusions, and truncations. As shown in Table 10.2, the subset contains 54k images and more than 3.1M person annotations, averaging over 58 persons per frame.

By contrast, the COCO Dataset [Lin14] contains about 200k person annotations across $118k$ images, or roughly 1.7 persons per image on average. Thus, GTA-RWS exceeds COCO by more than an order of magnitude in per-frame person density (58 vs.$\sim$ 1.7) and by more than 15$\times$ in the total number of person annotations ($3.1M$ vs.$\sim$ $200k$). With this scale and density, GTA-RWS is particularly well suited for evaluating robustness to occlusions, truncations, and dense multi-person surveillance scenarios.

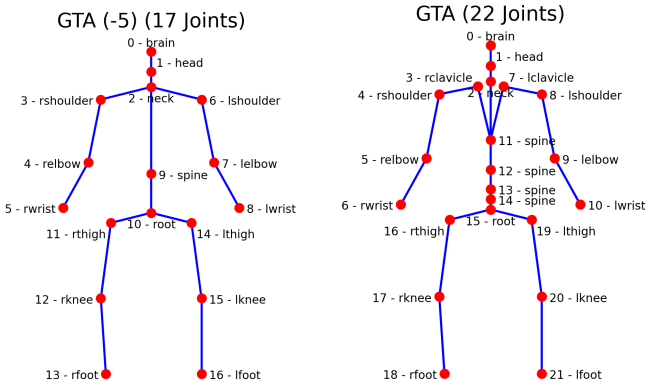


Figure 10.13: Comparison of the GTA-17 and GTA-22 skeleton topologies – The left image shows the GTA skeleton reduced to 17 keypoints, in which five keypoints are removed compared to the full 22-keypoint layout: the left and right clavicles, and three spine keypoints (11, 13, 14). The right image depicts the full 22-keypoint GTA topology, which provides additional spine and clavicle structure while preserving coverage of limbs and extremities. This expanded representation supports more detailed modeling of torso articulation and upper-body posture, is however visually difficult to locate.

Table 10.2: GTA-RWS 2D-HPE subset statistics. — The subset contains 54k images and 3.1M person annotations, averaging more than 58 persons per frame. Compared to COCO keypoints ($118k$ images, $200k$ persons, $\sim$ 1.7 persons per frame), GTA-RWS provides both an order of magnitude higher density and over 15× more total annotations, reflecting surveillance-style crowd scenes with heavy occlusion and truncation.

	Train	Test	Total	COCO
Images	37,770	16,230	54,000	118,000
Annotations (persons)	2,179,745	953,600	3,133,345	200,000
Avg. persons per frame	57.71	58.76	58.24	1.7

10.1.2.2 3D Human Pose Estimation Subset

The 3D-HPE uplifting subset targets single-person, root-relative lifting. A sliding-window protocol is applied to each person track across all sequences, using the same quality checks as in the SBAR extraction—*i.e.*, windows with

more than 50% visible joints and no empty skeletons are accepted. Accepted windows are root-centered and scale-normalized, and paired with their 2D projections and visibility flags to support temporal lifting models.

The joint layout follows the 22-keypoint GTA-RWS topology, which is reduced to the H36M [Ion13] joint set due to the spine keypoints, which cannot be reliably estimated (see Figure 10.14). In contrast to H36M, which contains about 3.6M frames recorded in constrained laboratory conditions with scripted motions, the GTA-RWS 3D-HPE subset provides more than 368k annotated samples (Table 10.3), each consisting of 64 consecutive frames. This corresponds to more than 23M annotated frames in total, spread across a wide range of unconstrained, realistic surveillance-style actions. These include everyday behaviors (standing, walking, sitting), object interactions (carrying, pushing, cycling), and salient activities (fighting, grappling, calling for help, lying).

The scale, behavioral diversity, and prevalence of naturalistic occlusions position GTA-RWS as a substantially more challenging benchmark than H36M and other laboratory datasets. Models trained on this subset must cope with crowding, truncated skeletons, and diverse motion patterns that reflect conditions closer to real-world deployment.

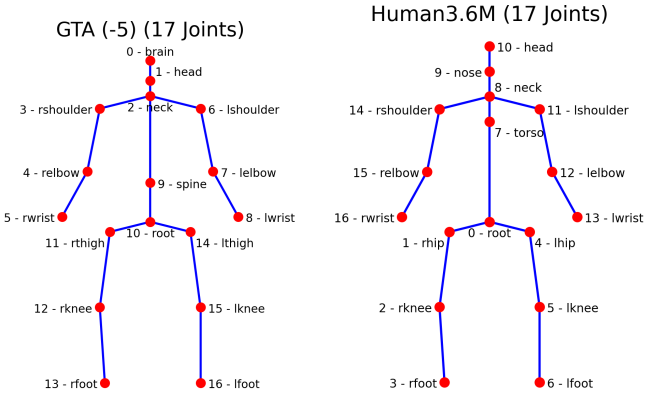


Figure 10.14: Comparison of the GTA-17 and H36M skeleton topologies – The left image shows the GTA-17 skeleton, which includes keypoints for the spine and head but omits clavicles and some torso keypoints. The right image depicts the H36M topology, which covers the torso and includes a nose keypoint, providing a different upper-body articulation representation. Limb coverage is broadly comparable, but the two topologies differ in the level of torso and head detail, reflecting their respective dataset design priorities.

Table 10.3: GTA-RWS 3D-HPE subset statistics – Per-class distribution of training and testing samples for the 3D-HPE subset. Each sample corresponds to a sliding window of 64 consecutive frames. The total scale exceeds 368k samples and 23.9M frames, offering substantially more diversity than lab-constrained datasets such asH36M

ID	Action Class	Train Samples	Test Samples	Train Frames	Test Frames
0	Miscellaneous	144	2,280	9,216	145,920
1	Standing	57,193	25,137	3,660,352	1,609, 0,768
2	Sitting	5,370	2,098	343,680	134,272
3	Walking	141,801	68,231	9,075,264	4,366,784
4	Carrying Object	8,110	4,161	519,040	266,304
5	Pushing Object	8,633	3,754	552,512	240,256
6	Bicycle	5,315	4,024	340,160	257,536
7	Greeting	3,495	1,466	223,680	93,824
8	Hugging	4,277	1,624	273,728	103,936
9	Dancing	6,263	2,463	400,832	157,632
10	Call for Help	2,325	652	148,800	41,728
11	Running	5,141	1,694	329,024	108,416
12	Stumbling	4,375	2,201	280,000	140,864
13	Fighting	5,618	2,547	359,552	163,008
14	Lying (+ Falling)	8,083	2,556	517,312	163,584
15	Grappling (+ Strangling)	3,806	1,575	243,584	100,800
Total		**256,409**	**111,672**	**16,009,712**	**7,147,008**

10.1.2.3 Skeleton-based Action Recognition Subset

The SBAR subset is constructed by applying a sliding window protocol independently to each person track across 1800 GTA-RWS sequences. A window is retained if at least 80 percent of its frames share the same action label, more than 50 percent of the joints are visible, and no frame contains an empty skeleton. For pairwise classes such as fighting or grappling, spatially nearest neighbors are checked for temporal overlap. If a valid partner is found, both individuals are extracted jointly, otherwise the individual is retained. The window advances by one frame when no sample is accepted and by the window length when a sample is extracted.

To mitigate class imbalance, sampling with replacement is applied under a hybrid weighting that blends the natural distribution with a uniform prior. This preserves the fact that common behaviors occur more often in surveillance style settings while preventing rare but salient actions from being underrepresented.

The resulting SBAR subset contains 24,100 training and 8,125 validation samples across 16 coarse grained action classes (Table 10.4). In terms of scale, this is comparable to UAV-Human which contains 67,000 action clips, and Toyota Smarthome with 16,000 clips. Beyond size, the subset provides greater behavioral diversity, naturalistic occlusions, and surveillance relevant salient actions such as fighting, grappling, or calling for help, which are largely absent in existing datasets. Thus, this a uniquely challenging benchmark for SBAR.

Table 10.4: GTA-RWS SBAR subset statistics – The table shows 16 coarse-grained action classes used for SBAR, along with notification category, and the number of training and validation samples extracted using the sliding-window protocol.

ID	Action Class	Notification	Training	Validation
0	Miscellaneous	Normal	546	230
1	Standing	Normal	4,903	1,170
2	Sitting	Normal	957	225
3	Walking	Normal	11,962	3,000
4	Carrying Object	Normal	1,235	337
5	Pushing Object	Normal	1,282	324
6	Bicycle	Normal	1,014	311
7	Greeting	Normal	752	189
8	Hugging	Normal	863	196
9	Dancing	Normal	1,006	254
10	Call for Help	Salient	739	137
11	Running	Normal	927	207
12	Stumbling	Normal	919	237
13	Fighting	Salient	943	248
14	Lying (+ Falling)	Salient	1,148	240
15	Grappling (+ Strangling)	Salient	804	195
Total			24,100	8,125

10.2 End-to-End System Evaluation

The evaluation of CTR-GCN on GTA-RWS examines performance across the full perception-to-recognition pipeline, highlighting how errors propagate from 2D keypoint detection to 3D skeleton reconstruction and ultimately to skeleton-based action recognition. This stepwise analysis captures realistic degradation patterns encountered in surveillance scenarios, including occlusions, truncations, and complex multi-person interactions.

VitPose Huge, trained on the 2D-HPE subset, achieves 45.76 AP and 46.58 AP_{50} when provided with ground-truth detections. Qualitative examples illustrating keypoint detection under diverse, crowded, and occluded scenarios are presented in Figure 10.15 and Figure 10.16.

MotionBERT with K_{v1}, trained on the 3D-HPE subset, demonstrates strong 3D reconstruction performance. When given ground-truth 2D keypoints, it achieves an MPJPE of 39.7 mm, P-MPJPE of 18.5 mm, MPBLE of 2.1 mm, and MPJVE of 10.6 mm, reflecting accurate recovery of joint positions and motion dynamics.

Figure 10.15: Qualitative results for 2D-HPE on GTA-RWS (Part 1) – Standing and walking people are estimated with excellent accuracy, whereas occluded individuals and people lying on the ground remain challenging.

Figure 10.16: Qualitative results for 2D-HPE on GTA-RWS (Part 2) – Performance remains high for standing and walking people across diverse weather conditions, while occlusions and individuals lying on the ground continue to pose challenges.

The evaluation proceeds through several structured blocks to isolate the contributions of different pipeline components. Block A establishes an upper-bound scenario using ground-truth 3D skeletons, yielding the highest overall performance. Introducing root-relative depth predictions while maintaining lateral coordinates (Block D) allows quantification of the fidelity of 3D reconstruction. Fully lifted 3D skeletons derived from ground-truth 2D keypoints (Block E) measure the combined effect of depth estimation and optional motion features such as joint velocities. Blocks involving predicted 2D detections capture the cumulative impact of upstream localization errors, simulating fully deployed pipelines.

CTR-GCN leverages 2D positions, lifted 3D coordinates, and joint velocities as input modalities. Estimated joint orientations were excluded in this evaluation since prior experiments showed negligible impact (Chapter 8). The inclusion of depth improves recognition for actions involving out-of-plane motion, self-contact, or multi-person interactions. Performance is measured using both overall ACC and MAPCA to account for class imbalance and ensure sensitivity to rare but salient behaviors.

Table 10.5 presents the results across evaluation blocks. Ground-truth 3D inputs in Block A achieve 72.7% ACC and 60.0% MAPCA, establishing the upper bound. While common actions such as *Walk*, *Run*, *Stand*, and *Lying* are recognized with high fidelity, interactive or occluded behaviors such as *Greet*, *Hug*, *Fight*, and *Strangle* show lower per-class accuracy, reflecting the intrinsic difficulty of these actions even under ideal 3D inputs.

Table 10.5: GTA-RWS CTR-GCN evaluation – Per-class and overall ACC (%) and MAPCA (%) across various input conditions. $\hat{x}$ denotes coordinates inferred using an uplifter, while un-hatted values correspond to ground-truth data. Blocks A and B rely on ground-truth detections and tracking; Blocks C and D extend the evaluation to fully predicted 2D inputs.

Method	Overvall ACC (%)↑		Per-Class ACC (%)↑														
	ACC	MAPCA	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
			Stand	Sit	Walk	Carry	Push	Bike	Greet	Hug	Dance	Help	Run	Stumble	Fight	Lying	Strangle
Block A: Ground-truth 3D																	
(x_{3D}, y_{3D}, z_{3D})	74.3	62.6	42.6	68.9	99.4	55.8	60.8	60.8	42.9	59.7	66.1	45.3	58.0	67.5	64.9	85.0	62.1
Block B: Fixed 2D GT input (x_{2D}, y_{2D}) with uplifted depth																	
$(x_{2D}, y_{2D}, 1)$	74.1	62.4	41.9	68.9	99.4	54.0	61.1	60.1	42.3	57.1	72.4	45.3	56.5	65.8	65.7	84.6	61.0
(x_{2D}, y_{2D}, z_{3D})	74.1	62.3	42.1	65.3	99.4	54.3	61.1	60.5	39.7	64.3	70.9	46.0	58.5	65.4	66.1	86.3	54.4
Block B: Fully predicted 3D from 2D GT inputs																	
$(\hat{x}_{3D}, \hat{y}_{3D}, \hat{z}_{3D})$	74.3	62.7	42.1	68.0	99.4	55.8	61.1	59.5	39.7	62.8	70.9	45.3	57.5	65.8	65.3	87.1	60.0
Block C: Fixed 2D detections (x_{2D}, y_{2D}) with uplifted depth																	
$(x_{2D}, y_{2D}, 1)$	62.1	42.6	32.1	48.0	97.5	8.0	47.8	42.8	14.3	41.3	44.5	35.0	36.7	35.0	43.2	81.7	30.8
(x_{2D}, y_{2D}, z_{3D})	61.8	42.4	30.9	55.6	97.4	6.2	47.8	41.8	10.6	39.8	40.6	35.0	35.3	35.0	49.6	80.4	30.3
Block D: Fully predicted 3D from 2D detections																	
$(\hat{x}_{3D}, \hat{y}_{3D}, \hat{z}_{3D})$	62.4	43.2	32.7	56.4	97.0	11.3	47.5	45.0	4.8	37.8	50.4	35.8	32.9	35.9	48.8	80.4	31.3

Introducing root-relative depth predictions with fixed lateral coordinates (Block D) results in a modest decrease to approximately 71.8% ACC and 58.5% MAPCA. Most spatial information is preserved, and actions with out-of-plane motion, such as *Carry*, *Push*, and *Bike*, retain relatively high accuracy, demonstrating that the uplifter captures structural cues effectively. Accuracy for complex interactions like *Hug* and *Fight* slightly decreases, highlighting the sensitivity of interactive actions to depth estimation errors.

Fully lifted 3D skeletons from ground-truth 2D keypoints (Block E) achieve 72.0% ACC and 58.8% MAPCA, showing that the integration of motion features partially compensates for depth estimation errors. Notably, highly dynamic actions such as *Dance* and *Stumble* benefit from velocity information, improving recognition consistency across frames. Simultaneously, self-contact and multi-person actions remain moderately robust, indicating that root-relative 3D reconstruction preserves essential relational cues even when full world coordinates are unavailable.

When 2D keypoints are predicted, the largest performance drop occurs, reflecting the sensitivity of the pipeline to localization errors. Actions involving subtle interactions or occlusions, including *Hug*, *Fight*, and *Strangle*, experience the most significant decline in accuracy, while common and isolated actions, such as *Walk*, *Run*, and *Stand*, maintain reasonable performance.

These results underscore the importance of accurate 2D detection for robust skeleton-based action recognition in realistic surveillance scenarios.

Overall, the evaluation demonstrates that GTA-RWS provides a comprehensive testbed for rigorous benchmarking of the full perception-to-recognition pipeline. The 2D-HPE subset supports training and evaluation of keypoint detection under crowded and occluded conditions. The 3D-HPE subset enables accurate root-relative depth estimation and multi-person skeleton reconstruction, while the SBAR subset facilitates assessment of individual and pairwise interactions. Extending the evaluation to fully predicted 2D keypoint detections allows for holistic analysis from raw imagery to action recognition, highlighting both strengths and limitations of practical surveillance pipelines.

10.3 Summary

The GTA-RWS dataset provides a comprehensive platform for the study and evaluation of SBAR under realistic, surveillance-style conditions. By offering subsets for 2D keypoint estimation, 3D lifting, and skeleton-based action recognition, it enables systematic assessment of perception-to-recognition pipelines, including the interplay between detection, tracking, 2D keypoint prediction, 3D reconstruction, and action classification.

The end-to-end evaluation conducted using GTA-RWS demonstrates that detecting and notifying salient behaviors in a privacy-preserving manner, through skeleton representations, is a highly challenging task. The stepwise analysis highlights how errors introduced at early stages, such as minor inaccuracies in 2D keypoint localization or partial occlusions, propagate through the system and can substantially affect final action recognition performance. Root-relative 3D reconstructions produced by uplifting 2D keypoints preserve most structural information, enabling robust recognition of individual and multi-person interactions. The inclusion of motion features, such as joint velocities, further supports the identification of complex behaviors, though benefits diminish in the presence of upstream noise.

Human operators and the public hold high expectations for such systems, which must operate transparently and reliably to avoid misuse or perceptions of intrusive surveillance. Achieving trustworthy performance requires careful evaluation of the full pipeline, an inherently complex endeavor given the number of interdependent modules and the diversity of human behaviors encountered in realistic scenarios. GTA-RWS delivers a substantial contribution in this context by providing a unified framework for rigorous, system-level benchmarking of human action recognition.

The dataset integrates naturalistic 3D motions captured from real people, transferred into GTA V, and augmented with external sources such as H4D, Inter-X, and BlindWays. This approach allows the creation of rich, annotated sequences covering both individual and multi-person behaviors, including complex interactions such as fighting or grappling, which are otherwise difficult or ethically challenging to record in real-world surveillance settings. While the visual domain remains synthetic, the dataset's modular subsets support targeted evaluation of each pipeline component—2D keypoint localization, 3D skeleton reconstruction, and skeleton-based action recognition—under realistic occlusions, crowded scenes, and variable viewpoints.

Overall, GTA-RWS establishes a new benchmark for evaluating human action recognition pipelines in privacy-conscious, surveillance-inspired environments. By enabling rigorous testing of 2D-HPE, 3D-HPE, and SBAR modules and capturing the propagation of errors across the full perception-to-recognition pipeline, the dataset provides a foundation for future research aimed at building transparent, reliable, and ethically responsible systems capable of detecting salient behaviors in complex real-world settings.

11 Conclusion and Outlook

After the evaluation of the framework for SBAR proposed in this thesis, the conclusion and outlook chapter summarizes and reflects the main findings and outcomes in Section 11.1. Following this, Section 11.2 provides an outlook on potential future developments.

11.1 Conclusion

In this thesis, a novel deep learning-based framework for SBAR in surveillance-oriented multi-camera networks is proposed. The framework addresses both data scarcity and domain variability, while embedding privacy and auditability as guiding design principles. The modular design of the pipeline supports component-based analysis, logging, and auditing. While this work focuses on three modules of the pipeline, the concept extends naturally to earlier stages as well.

First, the UPAR dataset and its UPAR-Pose extension are introduced, enabling joint studies of attribute recognition and human pose estimation in large-scale, harmonized settings. More than 3.3 million binary attribute annotations and over 71,000 joint pose-and-attribute labels are contributed, establishing a resource for cross-domain generalization experiments in both single-task and multitask learning scenarios. A multitask model using an early fusion strategy demonstrates that attribute and pose modalities are effectively combined, yielding improvements for action-related analysis in surveillance conditions. Attribute recognition profits strongly from shared features, whereas keypoint accuracy suffers, particularly under generalization protocols. Nonetheless,

this approach requires fewer resources and achieves lower latency than deploying separate models for each task.

Second, the LLVIP-P benchmark is contributed as an extension of LLVIP to overcome limitations of visible-spectrum approaches. With more than 26,000 pose-annotated person boxes in paired visible/thermal images, it enables rigorous evaluation under low-light and adverse illumination conditions. Results highlight a clear advantage of thermal over visible models, especially in crowded settings. Building on this, the unified UPPET benchmark harmonizes pose annotations across multiple thermal datasets, resolving discrepancies in keypoint definitions and protocols. With its shared topology and cross-sensor evaluation schemes, it advances thermal pose estimation for surveillance and occupational safety contexts. Data augmentation improves generalizability between sensor modalities, producing gains in mean AP and reduced variance across folds. These findings confirm that strong augmentation acts as an effective regularizer for domain-shift robustness in thermal imagery, often surpassing the benefits of additional model capacity. This also improve generalization protocols in the visible spectrum.

Third, a kinematics-aware 3D-HPE framework is contributed to stabilize temporal sequences and ensure anatomical plausibility. Bone-length constraints, joint-limit regularization, and temporal fusion are integrated into a dual-branch architecture predicting both relative joint coordinates and orientations. Comprehensive ablations and state-of-the-art comparisons highlight favorable accuracy–efficiency trade-offs, particularly on H4D dataset where fights with fast and occlusion-heavy motion is prevalent. The proposed configurations deliver robust kinematics and competitive positional accuracy at low latency, addressing central requirements of surveillance-oriented human analysis. The technique applies flexibly to larger or smaller backbones, consistently producing state-of-the-art results.

Fourth, surveillance-specific data augmentation is introduced to emulate characteristic degradations such as occlusion, truncation, temporal jitter, and fragmented identities. Systematic evaluation confirms that these perturbations enhance robustness in downstream action recognition and offer design guidelines for deployment under adverse surveillance conditions. It is

further shown that 3D joint orientations contribute to improved predictions and interpretability only under ground-truth quality. Current predictions with around 10° MPJAE, although state of the art, remain too noisy to yield reliable benefits in practical pipelines.

Finally, a synthetic surveillance benchmark, GTA-RWS, is contributed, rendering naturalistic human motion in a game engine as a privacy-conscious alternative to sensitive real-world recordings. Modular subsets cover 2D-HPE, 3D-HPE, and SBAR, enabling systematic end-to-end evaluation across detection, tracking, pose estimation, 3D reconstruction, and skeleton based action recognition. Error propagation is quantified across the pipeline, supporting holistic benchmarking in ways that isolated datasets do not permit. To the best of the author's knowledge, this is the first synthetic surveillance dataset that allows systematic study of the complete action recognition pipeline, including the influence of individual components on overall system performance, with realistic movement patterns derived from real human motion. The end-to-end evaluation conducted using GTA-RWS demonstrates that detecting and notifying salient behaviors in a privacy-preserving manner, through skeleton representations, is a highly challenging task. The stepwise analysis illustrates how even minor inaccuracies in early stages—such as small errors in 2D keypoint localization or partial occlusions can propagate through the pipeline, significantly impacting the final performance of SBAR.

Together, these contributions establish a framework that demonstrates competitive accuracy under real-world constraints while aligning with the ethical and legal requirements of surveillance practice. The results provide a foundation for advancing action recognition research in public safety applications, where both technical performance and regulatory compliance remain essential.

11.2 Outlook

While this thesis establishes a framework for SBAR under surveillance constraints, several promising directions remain open for future work. On the

technical side, further exploration of multitask formulations is particularly attractive. The combination of 2D-HPE with attribute recognition, as well as the joint optimization of 3D-HPE and orientation prediction, has demonstrated potential for both efficiency and robustness. Targeted fine-tuning strategies and adaptive balancing mechanisms could refine these interactions and mitigate task interference under domain shift. Moreover, there remains substantial room for improvement in terms of accuracy across all tasks. Future work should aim to develop models that are more generalizable across diverse scenarios, resilient to noise and occlusions, and capable of maintaining strong performance under challenging real-world conditions.

Synthetic data also constitutes a central research direction. The contributed GTA-RWS benchmark represents a first step toward privacy-conscious synthetic surveillance datasets, but it is limited by the context constraints of the game engine, where naturalistic motions may occur without realistic scene interactions. Advances in generative AI, such as diffusion models and neural scene representations, promise to render synthetic data with higher fidelity and richer semantics. Combining such techniques with large-scale motion capture corpora could enable training pipelines that rely primarily on synthetic appearance and real motion, while restricting annotation of real-world footage to evaluation purposes. This shift would substantially reduce the need for sensitive personal data in training, thereby aligning technical development with societal demands for privacy and acceptance.

Beyond technical challenges, responsible deployment requires addressing issues of transparency, governance, and societal trust. The adoption of MLOps principles—encompassing continuous integration, monitoring, reproducibility, and traceability—offers a practical foundation for auditability. Recording how models are trained, what data are processed, and how outputs are derived ensures that systems remain inspectable rather than opaque. At the same time, deployment practices must be aligned with regulatory frameworks such as the GDPR and the forthcoming EU AI Act. Currently, practical guides on how to implement surveillance-oriented AI systems under such regulations remain scarce. Research that translates high-level requirements into human-centered engineering practices is therefore urgently needed.

Equally important is the involvement of stakeholders in the research process itself. Pilot projects where police authorities, data protection officers, and research institutions collaborate—such as those initiated in Mannheim and Hamburg—should be viewed positively, as they allow citizens to have a tangible role in shaping these technologies. This participatory model contrasts with the procurement of narrowly defined commercial products and strengthens the democratic legitimacy of surveillance research.

Finally, credibility and acceptance depend not only on system performance but also on comprehensibility. Transparency and auditability lose their effect if the underlying concepts remain inaccessible to non-experts. In line with science and media communication, the findings, purposes, limitations, and risks of such systems must be explained in language that is understandable to the general public, not only to specialists. Addressing public expectations as well as fears openly helps to maintain and strengthen trust. Credibility is not static; it must be continuously reinforced through reproducibility, open communication, and demonstrable safeguards.

Together, these perspectives indicate that progress in surveillance-oriented action recognition is not solely a technical matter. The future of the field will depend equally on synthetic data innovation, rigorous engineering practice, regulatory alignment, and effective communication with the public. Only by combining these dimensions can systems be developed that are accurate, lawful, and socially acceptable.

Bibliography

[Abd24] Abdelfattah, Mohamed; Hassan, Mariam and Alahi, Alexandre: "MaskCLR: Attention-Guided Contrastive Learning for Robust Action Representation Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 18678–18687 (cit. on p. 29).

[An21] An, Jaeju; Kim, Jeongho; Lee, Hanbeen; Kim, Jinbeom; Kang, Junhyung; Kim, Minha; Shin, Saebyeol; Kim, Minha; Hong, Donghee and Woo, Simon S.: "VFP290K: A Large-Scale Benchmark Dataset for Vision-based Fallen Person Detection". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021. URL: https://openreview.net/forum?id=y2AbfIXgBK3 (cit. on p. 72).

[And14] Andriluka, Mykhaylo; Pishchulin, Leonid; Gehler, Peter and Schiele, Bernt: "2d human pose estimation: New benchmark and state of the art analysis". In: *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*. 2014, pp. 3686–3693 (cit. on pp. 23, 42, 43).

[And18] Andriluka, Mykhaylo; Iqbal, Umar; Insafutdinov, Eldar; Pishchulin, Leonid; Milan, Anton; Gall, Juergen and Schiele, Bernt: "PoseTrack: A Benchmark for Human Pose Estimation and Tracking". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018 (cit. on pp. 23, 37, 42, 43).

[Ang20] Angelini, Federico; Fu, Zeyu; Long, Yang; Shao, Ling and Naqvi, Syed Mohsen: "2D pose-based real-time human action recognition with occlusion-handling". In: *IEEE Transactions on*

Multimedia 22.6 (2020). Publisher: IEEE, pp. 1433–1446 (cit. on pp. 135, 136).

[Bla11] Blanchard, Gilles; Lee, Gyemin and Scott, Clayton: "Generalizing from several related classification tasks to a new unlabeled sample". In: *Advances in neural information processing systems* 24 (2011) (cit. on p. 58).

[Bra14] Brauer, Jürgen: "Human Pose Estimation with Implicit Shape Models". ISBN: 978-3-7315-0184-8 ISSN: 1866-5934 Series: Schriftenreihe Automatische Sichtprüfung und Bildverarbeitung Volume: 6. PhD thesis. KIT Scientific Publishing, 2014. 264 pp. doi: 10.5445/KSP/1000039083 (cit. on p. 24).

[Bra21] Brasó, Guillem; Kister, Nikita and Leal-Taixé, Laura: "The Center of Attention: Center-Keypoint Grouping via Attention for Multi-Person Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 11853–11863 (cit. on p. 21).

[Bun23] Bundesverfassungsgericht: Pressemitteilung Nr. 18/2023: Grundsatzentscheidung zur unzulässigen Videoüberwachung. Publisher: Bundesverfassungsgericht. July 25, 2023. url: https://www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/DE/2023/bvg23-018.html (visited on 09/04/2025) (cit. on p. 4).

[Bun25] Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (BfDI): Videoüberwachung. Publisher: Bundesbeauftragte für den Datenschutz und die Informationsfreiheit. Sept. 4, 2025. url: https://www.bfdi.bund.de/DE/Buerger/Inhalte/Allgemein/Datenschutz/Videoueberwachung.html (visited on 09/04/2025) (cit. on p. 2).

[Bus20] Buslaev, Alexander; Iglovikov, Vladimir I; Khvedchenya, Eugene; Parinov, Alex; Druzhinin, Mikhail and Kalinin, Alexandr A: "Albumentations: fast and flexible image augmentations". In: *Information* 11.2 (2020). Publisher: Multidisciplinary Digital Publishing Institute, p. 125 (cit. on pp. 100, 136).

[Cai19] Cai, Yujun; Ge, Liuhao; Liu, Jun; Cai, Jianfei; Cham, Tat-Jen; Yuan, Junsong and Thalmann, Nadia Magnenat: "Exploiting Spatial-Temporal Relationships for 3D Pose Estimation via Graph Convolutional Networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019 (cit. on pp. 25, 86).

[Cao17] Cao, Zhe; Simon, Tomas; Wei, Shih-En and Sheikh, Yaser: "Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017 (cit. on p. 18).

[Che20] Cheng, Bowen; Xiao, Bin; Wang, Jingdong; Shi, Honghui; Huang, Thomas S. and Zhang, Lei: "HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on p. 19).

[Che21a] Chen, Tianlang; Fang, Chen; Shen, Xiaohui; Zhu, Yiheng; Chen, Zhili and Luo, Jiebo: "Anatomy-aware 3d human pose estimation with bone-based pose decomposition". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.1 (2021). Publisher: IEEE, pp. 198–209 (cit. on pp. 25, 86, 108).

[Che21b] Chen, Yuxin; Zhang, Ziqi; Yuan, Chunfeng; Li, Bing; Deng, Ying and Hu, Weiming: "Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 13359–13368 (cit. on pp. 29, 31, 77, 89, 90, 166).

[Che24] Chen, Sichen; Zhang, Yingyi; Huang, Siming; Yi, Ran; Fan, Ke; Zhang, Ruixin; Chen, Peixian; Wang, Jun; Ding, Shouhong and Ma, Lizhuang: "SDPose: Tokenized Pose Estimation via Circulation-Guide Self-Distillation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 1082–1090 (cit. on p. 20).

[Cho09] Choi, Wongun; Shahid, Khuram and Savarese, Silvio: "What are they doing?: Collective activity classification using spatio-temporal relationship among people". In: *2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops*. IEEE, 2009, pp. 1282–1289 (cit. on p. 30).

[Cho18] Choi, Yukyung; Kim, Namil; Hwang, Soonmin; Park, Kibaek; Yoon, Jae Shin; An, Kyounghwan and Kweon, In So: "KAIST multi-spectral day/night data set for autonomous and assisted driving". In: *IEEE Transactions on Intelligent Transportation Systems* 19.3 (2018). Publisher: IEEE, pp. 934–948 (cit. on pp. 14, 49).

[Cho23] Choudhury, Rohan; Kitani, Kris M. and Jeni, László A.: "TEMPO: Efficient Multi-View Pose Estimation, Tracking, and Forecasting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 14750–14760 (cit. on pp. 22, 25).

[Ci19] Ci, Hai; Wang, Chunyu; Ma, Xiaoxuan and Wang, Yizhou: "Optimizing Network Structure for 3D Human Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019 (cit. on pp. 25, 86).

[Cor21a] Cormier, Mickael: "A Data Annotation Process for Human Activity Recognition in Public Places". In: *Proceedings of the 2020 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Ed.: J. Beyerer; T. Zander*. Vol. 51. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2021, pp. 33–51 (cit. on pp. 43, 260).

[Cor21b] Cormier, Mickael; Röpke, Fabian; Golda, Thomas and Beyerer, Jürgen: "Interactive Labeling for Human Pose Estimation in Surveillance Videos". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2021, pp. 1649–1658 (cit. on pp. 5, 43).

[Cor21c] Cormier, Mickael; Seletkov, Dmitrii and Beyerer, Jürgen: "Towards Lower Precision Quantization for Pedestrian Detection in Crowded Scenario". In: *IEEE EUROCON 2021 - 19th International Conference on Smart Technologies*. 2021, pp. 254–258. doi: 10.1109/EUROCON52738.2021.9535539 (cit. on pp. 36, 38).

[Cor21d] Cormier, Mickael; Wolf, Stefan; Sommer, Lars; Schumann, Arne and Beyerer, Jürgen: "Fast Pedestrian Detection for Real-World Crowded Scenarios on Embedded GPU". In: *IEEE EUROCON 2021 - 19th International Conference on Smart Technologies*. 2021, pp. 40–44. doi: 10.1109/EUROCON52738.2021.9535550 (cit. on pp. 36, 38, 80).

[Cor22a] Cormier, Mickael: "A Simple Pyramid Vision Transformer for Human Pose Estimation in Crowds". In: *Proceedings of the 2021 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory*. Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Vol. 54. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. ISSN: 1863-6489. Karlsruher Institut für Technologie (KIT), 2022, pp. 33–51 (cit. on p. 38).

[Cor22b] Cormier, Mickael; Clepe, Aris; Specker, Andreas and Beyerer, Jürgen: "Where Are We With Human Pose Estimation in Real-World Surveillance?" In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2022, pp. 591–601 (cit. on pp. 2, 38, 42, 132).

[Cor23] Cormier, Mickael; Specker, Andreas; Junior, Julio C. S. Jacques; Florin, Lucas; Metzler, Jürgen; Moeslund, Thomas B.; Nasrollahi, Kamal; Escalera, Sergio and Beyerer, Jürgen: "UPAR Challenge: Pedestrian Attribute Recognition and Attribute-Based Person Retrieval – Dataset, Design, and Results". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of*

Computer Vision (WACV) Workshops. 2023, pp. 166–175 (cit. on pp. 13, 37, 47).

[Cor24a] Cormier, Mickael; Schmid, Yannik and Beyerer, Jürgen: "Enhancing Skeleton-Based Action Recognition in Real-World Scenarios Through Realistic Data Augmentation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2024, pp. 290–299 (cit. on pp. 2, 15, 134).

[Cor24b] Cormier, Mickael; Specker, Andreas; Junior, Julio C. S. Jacques; Moritz, Lennart; Metzler, Jürgen; Moeslund, Thomas B.; Nasrollahi, Kamal; Escalera, Sergio and Beyerer, Jürgen: "UPAR Challenge 2024: Pedestrian Attribute Recognition and Attribute-Based Person Retrieval - Dataset, Design, and Results". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2024, pp. 359–367 (cit. on pp. 13, 38, 47).

[Cor24c] Cormier, Mickael; Yi, Caleb Ng Zhi; Specker, Andreas; Blaß, Benjamin; Heizmann, Michael and Beyerer, Jürgen: "Leveraging Thermal Imaging for Robust Human Pose Estimation in Low-Light Vision". In: *Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops*. 2024, pp. 67–83 (cit. on pp. 14, 38, 50, 52).

[Cor25] Cormier, Mickael; Specker, Andreas and Beyerer, Jürgen: "UPPET: Unified Pedestrian Pose Estimation in Thermal Imaging". In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*. 2025, pp. 4551–4560 (cit. on pp. 38, 51).

[Cor26a] Cormier, Mickael; Specker, Andreas and Beyerer, Jürgen: "UPAR-Pose: Joint pedestrian pose estimation and attribute recognition". In: *Under Review*. 2026 (cit. on pp. 13, 14, 48, 94).

[Cor26b] Cormier, Mickael; Zolk, Jeremy; Laubenheimer, Astrid and Beyerer, Jürgen: "Uplifting 2D to 3D Human Poses with Joint Rotations and Bone Constraints: A Strong Baseline for Sports

and Fitness Applications". In: *2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG)*. 2026 (cit. on pp. 14, 107).

[Das19] Das, Srijan; Dai, Rui; Koperski, Michal; Minciullo, Luca; Garattoni, Lorenzo; Bremond, Francois and Francesca, Gianpiero: "Toyota Smarthome: Real-World Activities of Daily Living". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019 (cit. on pp. 72–74, 131).

[Den14] Deng, Yubin; Luo, Ping; Loy, Chen Change and Tang, Xiaoou: "Pedestrian attribute recognition at far distance". In: *Proceedings of the 22nd ACM international conference on Multimedia*. 2014, pp. 789–792 (cit. on pp. 44–46, 55).

[Do24] Do, Jeonghyeok and Kim, Munchurl: "Skateformer: skeletal-temporal transformer for human action recognition". In: *European Conference on Computer Vision*. Springer, 2024, pp. 401–420 (cit. on p. 29).

[Dör22] Döring, Andreas; Chen, Di; Zhang, Shanshan; Schiele, Bernt and Gall, Jürgen: "PoseTrack21: A Dataset for Person Search, Multi-Object Tracking and Multi-Person Pose Tracking". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 20963–20972 (cit. on pp. 23, 37, 42, 43).

[Dos20] Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain et al.: "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020) (cit. on p. 80).

[Dri66] Drillis, R and Contini, R: Body segment parameters, New York University. NY, Technical Report, 1966 (cit. on p. 269).

[Dua22a] Duan, Haodong; Wang, Jiaqi; Chen, Kai and Lin, Dahua: "Pyskl: Towards good practices for skeleton action recognition". In: *Proceedings of the 30th ACM international conference on multimedia*. 2022, pp. 7351–7354 (cit. on p. 31).

[Dua22b] Duan, Haodong; Zhao, Yue; Chen, Kai; Lin, Dahua and Dai, Bo: "Revisiting Skeleton-Based Action Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 2969–2978. URL: https://github.com/kennymckormick/pyskl (cit. on pp. 2, 28).

[Dwi24] Dwivedi, Sai Kumar; Schmid, Cordelia; Yi, Hongwei; Black, Michael J and Tzionas, Dimitrios: "Poco: 3d pose and shape estimation with confidence". In: *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 85–95 (cit. on p. 24).

[Ein23] Einfalt, Moritz; Ludwig, Katja and Lienhart, Rainer: "Uplift and Upsample: Efficient 3D Human Pose Estimation With Uplifting Transformers". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2023, pp. 2903–2913. URL: https://github.com/goldbricklemon/uplift-upsample-3dhpe (cit. on pp. 86, 87, 161–163).

[Eur25] European Commission: Artificial Intelligence Act: Regulatory Framework. Publisher: European Commission. 2025. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (visited on 09/04/2025) (cit. on p. 2).

[Fab18] Fabbri, Matteo; Lanzi, Fabio; Calderara, Simone; Palazzi, Andrea; Vezzani, Roberto and Cucchiara, Rita: "Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (cit. on pp. 24, 26, 39, 172, 176).

[Fab20] Fabbri, Matteo; Lanzi, Fabio; Calderara, Simone; Alletto, Stefano and Cucchiara, Rita: "Compressed Volumetric Heatmaps for Multi-Person 3D Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on p. 26).

[Fab21] Fabbri, Matteo; Brasó, Guillem; Maugeri, Gianluca; Cetintas, Orcun; Gasparini, Riccardo; Ošep, Aljoša; Calderara, Simone; Leal-Taixé, Laura and Cucchiara, Rita: "MOTSynth: How Can Synthetic Data Help Pedestrian Detection and Tracking?" In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 10849–10859 (cit. on pp. 37, 39, 172).

[Fie21] Fieraru, Mihai; Zanfir, Mihai; Pirlea, Silviu Cristian; Olaru, Vlad and Sminchisescu, Cristian: "AIFit: Automatic 3D Human-Interpretable Feedback Models for Fitness Training". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 9919–9928 (cit. on pp. 61, 62, 65).

[Fis21] Fisch, Martin and Clark, Ronald: "Orientation keypoints for 6D human pose estimation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (2021). Publisher: IEEE, pp. 10145–10158 (cit. on p. 119).

[Geb18] Gebhardt, Evan and Wolf, Marilyn: "Camel dataset for visual and thermal infrared multiple object detection and tracking". In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2018, pp. 1–6 (cit. on pp. 14, 52).

[Gen21] Geng, Zigang; Sun, Ke; Xiao, Bin; Zhang, Zhaoxiang and Wang, Jingdong: "Bottom-Up Human Pose Estimation via Disentangled Keypoint Regression". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 14676–14686 (cit. on pp. 19, 152, 154).

[Gol19] Golda, Thomas; Kalb, Tobias; Schumann, Arne and Beyerer, Jürgen: "Human pose estimation for real-world crowded scenarios". In: *2019 16th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2019, pp. 1–8 (cit. on p. 19).

[Gol22] Golda, Thomas; Thiemich, Johanna; Cormier, Mickael and Beyerer, Jürgen: "For the Sake of Privacy: Skeleton-Based Salient Behavior Recognition". In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 3983–3987. doi: 10.1109/ICIP46576.2022.9897358 (cit. on pp. 2, 72).

[Gol23] Golda, Thomas; Cormier, Mickael and Beyerer, Jürgen: "Intelligente Bild- und Videoauswertung für die Sicherheit". In: *Handbuch Polizeimanagement: Polizeipolitik – Polizeiwissenschaft – Polizeipraxis*. Ed. by Wehe, Dieter and Siller, Helmut. Wiesbaden: Springer Fachmedien Wiesbaden, 2023, pp. 1487–1507. doi: 10.1007/978-3-658-34388-0_87. url: https://doi.org/10.1007/978-3-658-34388-0_87 (cit. on pp. 2, 3).

[Goz25] Gozlan, Yoni; Falisse, Antoine; Uhlrich, Scott; Gatti, Anthony; Black, Michael; Hicks, Jennifer; Delp, Scott and Chaudhari, Akshay: "OpenCapBench: A Benchmark to Bridge Pose Estimation and Biomechanics". In: *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*. Feb. 2025, pp. 4056–4065 (cit. on p. 24).

[Hac23] Hachiuma, Ryo; Sato, Fumiaki and Sekii, Taiki: "Unified Keypoint-Based Action Recognition Framework via Structured Keypoint Pooling". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 22962–22971 (cit. on p. 31).

[Ham25] Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit: Tätigkeitsbericht Datenschutz 2024. Publisher: Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit. Apr. 2025. url: https://datenschutz-hamburg.de/service-information/taetigkeitsberichte/taetigkeitsbericht-datenschutz-2024#IV-9 (visited on 09/04/2025) (cit. on p. 4).

[He16] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing and Sun, Jian: "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on p. 80).

[Hei24a] Heidenheimer Zeitung: Evaluation von Videoüberwachung in Mannheim erst 2027. 2024. URL: https://www.hz.de/suedwest/evaluation-von-videoueberwachung-in-mannheim-erst-2027 (visited on 09/04/2025) (cit. on p. 3).

[Hei24b] Heise Online: Evaluation von intelligenter Videoüberwachung in Mannheim erst 2027. 2024. URL: https://www.heise.de/news/Evaluation-von-intelligenter-Videoueberwachung-in-Mannheim-erst-2027-9580827.html (visited on 09/04/2025) (cit. on p. 3).

[Hen20] Hendrycks, Dan; Mu, Norman; Cubuk, Ekin Dogus; Zoph, Barret; Gilmer, Justin and Lakshminarayanan, Balaji: "Augmix: A simple method to improve robustness and uncertainty under data shift". In: *International conference on learning representations*. Vol. 1. Issue: 2. 2020, p. 5 (cit. on p. 99).

[Her23] Hermann, Prof Dr Dieter: Mannheimer Sicherheitsaudit 2022/23. Institut für Kriminologie, Universität Heidelberg, 2023, p. 75. URL: https://www.mannheim.de/sites/default/files/2023-04/Gutachten-MA-2023_final.pdf (cit. on p. 3).

[Hos18] Hossain, Mir Rayat Imtiaz and Little, James J.: "Exploiting temporal information for 3D human pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (cit. on pp. 25, 86).

[Hsu24] Hsu, Chih-Hsiang and Jang, Jyh-Shing Roger: "Enhancing 3D Human Pose Estimation with Bone Length Adjustment". In: *Proceedings of the Asian Conference on Computer Vision*. 2024, pp. 3723–3738 (cit. on pp. 25, 108).

[Hu21] Hu, Wenbo; Zhang, Changgong; Zhan, Fangneng; Zhang, Lei and Wong, Tien-Tsin: "Conditional directed graph convolution for 3d human pose estimation". In: *Proceedings of the 29th ACM international conference on multimedia*. 2021, pp. 602–611 (cit. on p. 122).

[Hua20] Huang, Junjie; Zhu, Zheng; Guo, Feng and Huang, Guan: "The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on p. 19).

[Hua24] Huang, Qian; Liu, Wenting; Shang, Mingzhou and Wang, Yiming: "Fusing angular features for skeleton-based action recognition using multi-stream graph convolution network". In: *IET Image Processing* 18.7 (2024). Publisher: Wiley Online Library, pp. 1694–1709 (cit. on p. 117).

[Huy09] Huynh, Du Q: "Metrics for 3D rotations: Comparison and analysis". In: *Journal of Mathematical Imaging and Vision* 35.2 (2009). Publisher: Springer, pp. 155–164 (cit. on pp. 118, 119).

[Ibr16] Ibrahim, Mostafa S.; Muralidharan, Srikanth; Deng, Zhiwei; Vahdat, Arash and Mori, Greg: "A Hierarchical Deep Temporal Model for Group Activity Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on p. 30).

[Ing23] Ingwersen, Christian Keilstrup; Mikkelstrup, Christian Møller; Jensen, Janus Nørtoft; Hannemose, Morten Rieger and Dahl, Anders Bjorholm: "SportsPose - A Dynamic 3D Sports Pose Dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2023, pp. 5219–5228 (cit. on pp. 61, 62).

[Ion13] Ionescu, Catalin; Papava, Dragos; Olaru, Vlad and Sminchisescu, Cristian: "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments". In: *IEEE transactions on pattern analysis and machine intelligence* 36.7 (2013). Publisher: IEEE, pp. 1325–1339 (cit. on pp. 61–63, 69, 186).

[Jhu13] Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C. and Black, M. J.: "Towards understanding action recognition". In: *International*

Conf. on Computer Vision (ICCV). Dec. 2013, pp. 3192–3199 (cit. on pp. 72, 73).

[Jia21] Jia, Xinyu; Zhu, Chuang; Li, Minzhen; Tang, Wenqi and Zhou, Wenli: "LLVIP: A visible-infrared paired dataset for low-light vision". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 3496–3504 (cit. on pp. 14, 42, 49).

[Jia24] Jiang, Tianjian; Billingham, Johsan; Müksch, Sebastian; Zarate, Juan; Evans, Nicolas; Oswald, Martin; Pollefeys, Marc; Hilliges, Otmar; Kaufmann, Manuel and Song, Jie: "WorldPose: A World Cup Dataset for Global 3D Human Pose Estimation". In: *eccv* (2024) (cit. on pp. 24, 26).

[Jin23] Jin, Kyung-Min; Lim, Byoung-Sung; Lee, Gun-Hee; Kang, Tae-Kyung and Lee, Seong-Whan: "Kinematic-aware Hierarchical Attention Network for Human Pose Estimation in Videos". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 5725–5734 (cit. on pp. 22, 25, 110, 132).

[Joo15] Joo, Hanbyul; Liu, Hao; Tan, Lei; Gui, Lin; Nabbe, Bart; Matthews, Iain; Kanade, Takeo; Nobuhara, Shohei and Sheikh, Yaser: "Panoptic Studio: A Massively Multiview System for Social Motion Capture". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015 (cit. on pp. 61, 62).

[Kel23] Keller, Marilyn; Werling, Keenon; Shin, Soyong; Delp, Scott; Pujades, Sergi; Liu, C Karen and Black, Michael J: "From skin to skeleton: Towards biomechanically accurate 3d digital humans". In: *ACM Transactions on Graphics (TOG)* 42.6 (2023). Publisher: ACM New York, NY, USA, pp. 1–12 (cit. on pp. 24, 116).

[Kes23] Kessel, Wolfgang: Pilotprojekt Videoüberwachung mit KI in Mannheim: Verlängerung bis 2026. Publisher: SWR Aktuell. Dec. 4, 2023. url: https://www.swr.de/swraktuell/baden-wuerttemberg/mannheim/videoueberwachung-kameras-

videoschutz-polizei-mannheim-innenstadt-sicherheit-strobl-100.html (visited on 09/04/2025) (cit. on p. 3).

[Khi21] Khirodkar, Rawal; Chari, Visesh; Agrawal, Amit and Tyagi, Ambrish: "Multi-Instance Pose Networks: Rethinking Top-Down Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 3122–3131 (cit. on p. 19).

[Khi24a] Khirodkar, Rawal; Bagautdinov, Timur; Martinez, Julieta; Zhaoen, Su; James, Austin; Selednik, Peter; Anderson, Stuart and Saito, Shunsuke: "Sapiens: Foundation for human vision models". In: *European Conference on Computer Vision*. Springer, 2024, pp. 206–228 (cit. on pp. 21, 99).

[Khi24b] Khirodkar, Rawal; Song, Jyun-Ting; Cao, Jinkun; Luo, Zhengyi and Kitani, Kris: "Harmony4D: A Video Dataset for In-The-Wild Close Human Interactions". In: *Advances in Neural Information Processing Systems*. Ed. by Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. and Zhang, C. Vol. 37. Curran Associates, Inc., 2024, pp. 107270–107285. url: https://proceedings.neurips.cc/paper_files/paper/2024/file/c20b843d0c6b1b40a8e6eb9a44e719c9-Paper-Datasets_and_Benchmarks_Track.pdf (cit. on pp. 68, 172, 178).

[Kim24] Kim, Hee Jae; Sengupta, Kathakoli; Kuribayashi, Masaki; Kacorri, Hernisa and Ohn-Bar, Eshed: "Text to Blind Motion". In: *Advances in Neural Information Processing Systems*. Ed. by Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. and Zhang, C. Vol. 37. Curran Associates, Inc., 2024, pp. 16272–16285. url: https://proceedings.neurips.cc/paper_files/paper/2024/file/1d4c4047f5d82699cf01b19c991ec56d-Paper-Datasets_and_Benchmarks_Track.pdf (cit. on pp. 173, 178).

[Koh20] Kohl, Philipp; Specker, Andreas; Schumann, Arne and Beyerer, Jurgen: "The MTA Dataset for Multi-Target Multi-Camera Pedestrian Tracking by Weighted Distance Aggregation". In:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. June 2020 (cit. on p. 39).

[Kre19] Kreiss, Sven; Bertoni, Lorenzo and Alahi, Alexandre: "PifPaf: Composite Fields for Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 18, 22).

[Kre21] Kreiss, Sven; Bertoni, Lorenzo and Alahi, Alexandre: "Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association". In: *IEEE Transactions on Intelligent Transportation Systems* 23.8 (2021). Publisher: IEEE, pp. 13498–13511 (cit. on p. 22).

[Kuz24] Kuzdeuov, Askat; Taratynova, Darya; Tleuliyev, Alim and Varol, Huseyin Atakan: "OpenThermalPose: An Open-Source Annotated Thermal Human Pose Dataset and Initial YOLOv8-Pose Baselines". In: *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*. 2024, pp. 1–8. doi: 10.1109/FG59268.2024.10581992 (cit. on pp. 14, 23, 42, 43, 50, 52, 58).

[Lan21] Landtag Baden-Württemberg: Gefährliche Orte in Baden-Württemberg. 2021. url: https://www.landtag-bw.de/resource/blob/247662/6220f6e6290bef3a48e4e70c61b472cd/16_7437_D.pdf (visited on 09/04/2025) (cit. on p. 4).

[Lan24] Landtag Baden-Württemberg: Innenausschuss befasst sich mit Erkenntnissen aus Projekt „Intelligente Videoüberwachung" in Mannheim. Jan. 18, 2024. url: https://www.landtag-bw.de/de/aktuelles/pressemitteilungen/innenausschuss-befasst-sich-mit-erkenntnissen-aus-projekt-intelligente-videoueberwachung-in-mannheim--419604 (visited on 09/04/2025) (cit. on p. 3).

[Lee23] Lee, Jungho; Lee, Minhyeok; Lee, Dogyoon and Lee, Sangyoun: "Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition". In: *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV). Oct. 2023, pp. 10444–10453 (cit. on pp. 29, 32, 166).

[Lev20] Levinson, Jake; Esteves, Carlos; Chen, Kefan; Snavely, Noah; Kanazawa, Angjoo; Rostamizadeh, Afshin and Makadia, Ameesh: "An Analysis of SVD for Deep Rotation Estimation". In: *Advances in Neural Information Processing Systems*. Ed. by Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F. and Lin, H. Vol. 33. Curran Associates, Inc., 2020, pp. 22554–22565. url: https://proceedings.neurips.cc/paper_files/paper/2020/file/fec3392b0dc073244d38eba1feb8e6b7-Paper.pdf (cit. on p. 118).

[Li15] Li, Dangwei; Chen, Xiaotang and Huang, Kaiqi: "Multi-attribute Learning for Pedestrian Attribute Recognition in Surveillance Scenarios". In: *ACPR*. 2015, pp. 111–115 (cit. on pp. 45, 46, 56).

[Li19a] Li, D.; Zhang, Z.; Chen, X. and Huang, K.: "A Richly Annotated Pedestrian Dataset for Person Retrieval in Real Surveillance Scenarios". In: *IEEE Transactions on Image Processing* 28.4 (2019), pp. 1575–1590 (cit. on pp. 44–46).

[Li19b] Li, Jiefeng; Wang, Can; Zhu, Hao; Mao, Yihuan; Fang, Hao-Shu and Lu, Cewu: "CrowdPose: Efficient Crowded Scenes Pose Estimation and a New Benchmark". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 23, 37, 42, 43).

[Li20] Li, Shichao; Ke, Lei; Pratama, Kevin; Tai, Yu-Wing; Tang, Chi-Keung and Cheng, Kwang-Ting: "Cascaded Deep Monocular 3D Human Pose Estimation With Evolutionary Training Data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on p. 24).

[Li21a] Li, Ruilong; Yang, Shan; Ross, David A. and Kanazawa, Angjoo: "AI Choreographer: Music Conditioned 3D Dance Generation With AIST++". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 13401–13412 (cit. on pp. 61, 62).

[Li21b] Li, Tianjiao; Liu, Jun; Zhang, Wei; Ni, Yun; Wang, Wenqian and Li, Zhiheng: "UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 16266–16275 (cit. on pp. 73, 133).

[Li21c] Li, Yanjie; Zhang, Shoukui; Wang, Zhicheng; Yang, Sen; Yang, Wankou; Xia, Shu-Tao and Zhou, Erjin: "TokenPose: Learning Keypoint Tokens for Human Pose Estimation". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 (cit. on pp. 20, 80).

[Li22a] Li, Wenhao; Liu, Hong; Tang, Hao; Wang, Pichao and Van Gool, Luc: "MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 13147–13156 (cit. on pp. 86, 161, 162).

[Li22b] Li, Yanjie; Yang, Sen; Liu, Peidong; Zhang, Shoukui; Wang, Yunxiao; Wang, Zhicheng; Yang, Wankou and Xia, Shu-Tao: "SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation". In: *Computer Vision – ECCV 2022*. Ed. by Avidan, Shai; Brostow, Gabriel; Cissé, Moustapha; Farinella, Giovanni Maria and Hassner, Tal. Cham: Springer Nature Switzerland, 2022, pp. 89–106 (cit. on pp. 18, 21, 145, 148–152, 154, 156).

[Li23] Li, Wenhao; Liu, Hong; Tang, Hao and Wang, Pichao: "Multi-hypothesis representation learning for transformer-based 3D human pose estimation". In: *Pattern Recognition* 141 (2023). Publisher: Elsevier, p. 109631 (cit. on pp. 25, 86).

[Li24] Li, Wenhao; Liu, Mengyuan; Liu, Hong; Wang, Pichao; Cai, Jialun and Sebe, Nicu: "Hourglass Tokenizer for Efficient Transformer-Based 3D Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition (CVPR). 2024, pp. 604–613. URL: https://github.com/NationalGAILab/HoT (cit. on pp. 25, 86).

[Li25] LI, Zhengcen; CHANG, Xinle; LI, Yueran and SU, Jingyong: "Skeleton-Based Group Activity Recognition via Spatial-Temporal Panoramic Graph". In: *Computer Vision – ECCV 2024*. Ed. by LEONARDIS, Aleš; RICCI, Elisa; ROTH, Stefan; RUSSAKOVSKY, Olga; SATTLER, Torsten and VAROL, Gül. Cham: Springer Nature Switzerland, 2025, pp. 252–269 (cit. on p. 30).

[Lib24] LIBÉRATION: Vidéosurveillance algorithmique dans les gares : la SNCF visée par une plainte devant la CNIL. Publisher: Libération. May 2, 2024. URL: https://www.liberation.fr/societe/police-justice/videosurveillance-algorithmique-dans-les-gares-la-sncf-visee-par-une-plainte-devant-la-cnil-20240502_QLL62YCQMFBBDB7ZLSBOY2MSAY/?redirected=1 (visited on 09/04/2025) (cit. on pp. 4, 5).

[Lib25] LIBÉRATION: Le tribunal administratif de Grenoble ordonne la fin de l'utilisation de BriefCam, un logiciel de vidéosurveillance israélien, à Moirans (Isère). Publisher: Libération. Feb. 1, 2025. URL: https://www.liberation.fr/societe/police-justice/le-tribunal-administratif-de-grenoble-ordonne-la-fin-de-lutilisation-de-briefcam-un-logiciel-de-videosurveillance-israelien-a-moirans-dans-lisere-20250131_TC37ASZAYJA5VCD4X53SGHZZFY/ (visited on 09/04/2025) (cit. on pp. 4, 5).

[Lin14] LIN, Tsung-Yi; MAIRE, Michael; BELONGIE, Serge; HAYS, James; PERONA, Pietro; RAMANAN, Deva; DOLLÁR, Piotr and ZITNICK, C Lawrence: "Microsoft coco: Common objects in context". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755 (cit. on pp. 14, 23, 37, 42, 43, 50, 54, 184).

[Lin19] LIN, Yutian; ZHENG, Liang; ZHENG, Zhedong; WU, Yu; HU, Zhilan; YAN, Chenggang and YANG, Yi: "Improving Person Re-identification by Attribute and Identity Learning". In: *Pattern*

Recognition (2019). DOI: https://doi.org/10.1016/j.patcog.2019.06.006 (cit. on pp. 44, 46).

[Lin23] Lin, Weiyao; Liu, Huabin; Liu, Shizhan; Li, Yuxi; Xiong, Hongkai; Qi, Guojun and Sebe, Nicu: "Hieve: A large-scale benchmark for human-centric video analysis in complex events". In: *International Journal of Computer Vision* 131.11 (2023). Publisher: Springer, pp. 2994–3018 (cit. on pp. 23, 37, 42, 43).

[Liu17] Liu, Xihui; Zhao, Haiyu; Tian, Maoqing; Sheng, Lu; Shao, Jing; Yan, Junjie and Wang, Xiaogang: "HydraPlus-Net: Attentive Deep Features for Pedestrian Analysis". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1–9 (cit. on p. 44).

[Liu18] Liu, W.; W. Luo, D. Lian and Gao, S.: "Future Frame Prediction for Anomaly Detection – A New Baseline". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018 (cit. on p. 72).

[Liu19] Liu, Jun; Shahroudy, Amir; Perez, Mauricio; Wang, Gang; Duan, Ling-Yu and Kot, Alex C: "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding". In: *IEEE transactions on pattern analysis and machine intelligence* 42.10 (2019). Publisher: IEEE, pp. 2684–2701 (cit. on pp. 72, 73, 131).

[Liu20a] Liu, Kenkun; Ding, Rongqi; Zou, Zhiming; Wang, Le and Tang, Wei: "A comprehensive study of weight sharing in graph networks for 3d human pose estimation". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 318–334 (cit. on pp. 25, 86).

[Liu20b] Liu, Ruixu; Shen, Ju; Wang, He; Chen, Chen; Cheung, Sen-ching and Asari, Vijayan: "Attention Mechanism Exploits Temporal Contexts: Real-Time 3D Human Pose Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on pp. 25, 86).

[Liu21] Liu, Zhenguang; Chen, Haoming; Feng, Runyang; Wu, Shuang; Ji, Shouling; Yang, Bailin and Wang, Xun: "Deep Dual Consecutive Network for Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 525–534 (cit. on p. 22).

[Liu22a] Liu, Shuangjun; Huang, Xiaofei; Fu, Nihang; Li, Cheng; Su, Zhongnan and Ostadabbas, Sarah: "Simultaneously-collected multimodal lying pose dataset: Enabling in-bed human pose monitoring". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2022). Publisher: IEEE, pp. 1106–1118 (cit. on p. 50).

[Liu22b] Liu, Zhenguang; Feng, Runyang; Chen, Haoming; Wu, Shuang; Gao, Yixing; Gao, Yunjun and Wang, Xiang: "Temporal Feature Alignment and Mutual Information Maximization for Video-Based Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 11006–11016 (cit. on p. 22).

[Liu23a] Liu, Huan et al.: "Group Pose: A Simple Baseline for End-to-End Multi-Person Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 15029–15038 (cit. on p. 19).

[Liu23b] Liu, Qihao; Kortylewski, Adam and Yuille, Alan L: "PoseExaminer: Automated Testing of Out-of-Distribution Robustness in Human Pose and Shape Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 672–681 (cit. on p. 86).

[Liu24] Liu, Feng et al.: "FarSight: A Physics-Driven Whole-Body Biometric System at Large Distance and Altitude". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2024, pp. 6227–6236 (cit. on p. 4).

[Liu25a] Liu, Feng; Chimitt, Nicholas; Guo, Lanqing; Jain, Jitesh; Kane, Aditya; Kim, Minchul; Robbins, Wes; Su, Yiyang; Ye, Dingqiang; Zhang, Xingguang et al.: "Person Recognition at

Altitude and Range: Fusion of Face, Body Shape and Gait". In: *arXiv preprint arXiv:2505.04616* (2025) (cit. on p. 4).

[Liu25b] Liu, Jiajie; Liu, Mengyuan; Liu, Hong and Li, Wenhao: "TCP-Former: Learning Temporal Correlation with Implicit Pose Proxy for 3D Human Pose Estimation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39.5 (Apr. 2025), pp. 5478–5486. doi: 10.1609/aaai.v39i5.32583. url: https://ojs.aaai.org/index.php/AAAI/article/view/32583 (cit. on pp. 25, 86, 88, 161, 162).

[Lop23] Loper, Matthew; Mahmood, Naureen; Romero, Javier; Pons-Moll, Gerard and Black, Michael J: "SMPL: A skinned multi-person linear model". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023, pp. 851–866 (cit. on pp. 24, 116).

[Los17] Loshchilov, Ilya and Hutter, Frank: "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101* (2017) (cit. on pp. 82, 88).

[Lu24] Lu, Peng; Jiang, Tao; Li, Yining; Li, Xiangtai; Chen, Kai and Yang, Wenming: "RTMO: Towards High-Performance One-Stage Real-Time Multi-Person Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 1491–1500. url: https://github.com/openmmlab/%20mmpose/tree/main/projects/rtmo (cit. on pp. 20, 152, 154).

[Lud25a] Ludwig, Katja; Lorenz, Julian; Kienzle, Daniel; Bui, Tuan and Lienhart, Rainer: "Leveraging Anthropometric Measurements to Improve Human Mesh Estimation and Ensure Consistent Body Shapes". In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*. June 2025, pp. 5872–5881 (cit. on pp. 25, 116).

[Lud25b] Ludwig, Katja; Oksymets, Yuliia; Schön, Robin; Kienzle, Daniel and Lienhart, Rainer: "Efficient 2D to Full 3D Human Pose Uplifting including Joint Rotations". In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*

Workshops. June 2025, pp. 5852–5861 (cit. on pp. 66, 70, 117, 118).

[Lul23] Lulamae, Josephine: In Mannheim, an automated system reports hugs to the police. Publisher: AlgorithmWatch. July 18, 2023. url: https://algorithmwatch.org/en/mannheim-system-reports-hugs-police/ (visited on 09/04/2025) (cit. on p. 4).

[Mah19] Mahmood, Naureen; Ghorbani, Nima; Troje, Nikolaus F.; Pons-Moll, Gerard and Black, Michael J.: "AMASS: Archive of Motion Capture As Surface Shapes". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019 (cit. on pp. 63, 87).

[Maj22] Maji, Debapriya; Nagori, Soyeb; Mathew, Manu and Poddar, Deepak: "YOLO-Pose: Enhancing YOLO for Multi Person Pose Estimation Using Object Keypoint Similarity Loss". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2022, pp. 2637–2646 (cit. on pp. 152, 154).

[Mar17] Martinez, Julieta; Hossain, Rayat; Romero, Javier and Little, James J.: "A Simple yet Effective Baseline for 3D Human Pose Estimation". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017 (cit. on pp. 25, 86).

[McN22] McNally, William; Vats, Kanav; Wong, Alexander and McPhee, John: "Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation". In: *European Conference on Computer Vision*. Springer, 2022, pp. 37–54 (cit. on p. 20).

[Meh17] Mehta, Dushyant; Rhodin, Helge; Casas, Dan; Fua, Pascal; Sotnychenko, Oleksandr; Xu, Weipeng and Theobalt, Christian: "Monocular 3d human pose estimation in the wild using improved cnn supervision". In: *2017 international conference on 3D vision (3DV)*. IEEE, 2017, pp. 506–516 (cit. on pp. 61, 62).

[Meh24] Mehraban, Soroush; Adeli, Vida and Taati, Babak: "MotionAGFormer: Enhancing 3D Human Pose Estimation With a Transformer-GCNFormer Network". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2024, pp. 6920–6930 (cit. on pp. 161, 162).

[MMP20] MMPose-Contributors: OpenMMLab Pose Estimation Toolbox and Benchmark. 2020. URL: https://github.com/open-mmlab/mmpose (cit. on p. 82).

[New17] Newell, Alejandro; Huang, Zhiao and Deng, Jia: "Associative Embedding: End-to-End Learning for Joint Detection and Grouping". In: *Advances in Neural Information Processing Systems*. Ed. by Guyon, I.; Luxburg, U. Von; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S. and Garnett, R. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8edd72158ccd2a879f79cb2538568fdc-Paper.pdf (cit. on p. 18).

[Nib21] Nibali, Aiden; Millward, Joshua; He, Zhen and Morgan, Stuart: "ASPset: An outdoor sports pose video dataset with 3D keypoint annotations". In: *Image and Vision Computing* 111 (2021). Publisher: Elsevier, p. 104196 (cit. on pp. 61, 62).

[Nik21] Nikolov, Ivan Adriyanov; Philipsen, Mark Philip; Liu, Jinsong; Dueholm, Jacob Velling; Johansen, Anders Skaarup; Nasrollahi, Kamal and Moeslund, Thomas B: "Seasons in drift: A long-term thermal imaging dataset for studying concept drift". In: *Thirty-fifth Conference on Neural Information Processing Systems*. Neural Information Processing Systems Foundation, 2021 (cit. on pp. 14, 50).

[Pac25] Pace, Cesare Davide; De Nunzio, Alessandro Marco; De Stefano, Claudio; Fontanella, Francesco and Molinara, Mario: "Poseidon: A ViT-based Architecture for Multi-Frame Pose Estimation with Adaptive Frame Weighting and Multi-Scale Feature Fusion". In: *arXiv preprint arXiv:2501.08446* (2025) (cit. on p. 21).

[Pap18] Papandreou, George; Zhu, Tyler; Chen, Liang-Chieh; Gidaris, Spyros; Tompson, Jonathan and Murphy, Kevin: "PersonLab: Person Pose Estimation and Instance Segmentation with a Bottom-Up, Part-Based, Geometric Embedding Model". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (cit. on p. 18).

[Pas19] Paszke, A: "Pytorch: An imperative style, high-performance deep learning library". In: *arXiv preprint arXiv:1912.01703* (2019) (cit. on pp. 82, 88, 92).

[Pat25] Patel, Priyanka and Black, Michael J: "Camerahmr: Aligning people with perspective". In: *2025 International Conference on 3D Vision (3DV)*. IEEE, 2025, pp. 1562–1571 (cit. on p. 24).

[Pav19a] Pavlakos, Georgios; Choutas, Vasileios; Ghorbani, Nima; Bolkart, Timo; Osman, Ahmed A. A.; Tzionas, Dimitrios and Black, Michael J.: "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image". In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10975–10985 (cit. on p. 116).

[Pav19b] Pavllo, Dario; Feichtenhofer, Christoph; Grangier, David and Auli, Michael: "3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 25, 71, 86).

[Pen24] Peng, Jihua; Zhou, Yanghong and Mok, P. Y.: "KTPFormer: Kinematics and Trajectory Prior Knowledge-Enhanced Transformer for 3D Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 1123–1132 (cit. on pp. 25, 86, 114, 161, 162).

[Pis16] Pishchulin, Leonid; Insafutdinov, Eldar; Tang, Siyu; Andres, Bjoern; Andriluka, Mykhaylo; Gehler, Peter V. and Schiele, Bernt: "DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation". In: *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition (CVPR). June 2016 (cit. on p. 18).

[Pol23a] Polizei Hamburg: IVBeo2 – Intelligente Videobeobachtung in Hamburg. 2023. URL: https://www.polizei.hamburg/ivbeo2-intelligente-videobeobachtung--1066428 (cit. on p. 3).

[Pol23b] Polizei Hamburg: Videoüberwachung Hachmannplatz. 2023. URL: https://www.polizei.hamburg/services/recht/videoueberwachung-hachmannplatz (visited on 09/04/2025) (cit. on p. 4).

[Pol23c] Polizei Hamburg: Videoüberwachung Hansaplatz. 2023. URL: https://www.polizei.hamburg/hansaplatz-buergerinformation-a-787040 (visited on 09/04/2025) (cit. on p. 4).

[Pur25] Purkrabek, Miroslav and Matas, Jiri: "ProbPose: A Probabilistic Approach to 2D Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2025, pp. 27124–27133 (cit. on pp. 19–21).

[Qin22] Qin, Zhenyue; Liu, Yang; Ji, Pan; Kim, Dongwoo; Wang, Lei; McKay, Robert I; Anwar, Saeed and Gedeon, Tom: "Fusing higher-order features in graph neural networks for skeleton-based action recognition". In: *IEEE Transactions on Neural Networks and Learning Systems* 35.4 (2022). Publisher: IEEE, pp. 4783–4797 (cit. on p. 117).

[Raj22] Rajasegaran, Jathushan; Pavlakos, Georgios; Kanazawa, Angjoo and Malik, Jitendra: "Tracking People by Predicting 3D Appearance, Location and Pose". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 2740–2749 (cit. on p. 24).

[Red23] RedaktionsNetzwerk Deutschland: KI-Kameras: Umstrittene Helfer der Polizei. June 15, 2023. URL: https://www.rnd.de/panorama/ki-kameras-umstrittener-freund-und-helfer-der-polizei-Z62ORVRR7VJTZF4MAIQQSXL4RI.html (visited on 09/04/2025) (cit. on p. 3).

[Roc13] Rockstar Games: Grand Theft Auto V. Place: Edinburgh, UK Published: Software. 2013 (cit. on p. 171).

[Rog19] Rogez, Gregory; Weinzaepfel, Philippe and Schmid, Cordelia: "Lcr-net++: Multi-person 2d and 3d pose detection in natural images". In: *IEEE transactions on pattern analysis and machine intelligence* 42.5 (2019). Publisher: IEEE, pp. 1146–1161 (cit. on p. 74).

[Sch23a] Schabacher, Gabriele: "AI and the Work of Patterns: Recognition Technologies, Classification, and Security". In: *Beyond Quantity*. Ed. by Sudmann, Andreas; Echterhölter, Anna; Ramsauer, Markus; Retkowski, Fabian; Schröter, Jens and Waibel, Alexander. Bielefeld: transcript Verlag, 2023, pp. 123–154. doi: 10.14361/9783839467664-008. url: https://doi.org/10.14361/9783839467664-008 (visited on 09/04/2025) (cit. on p. 2).

[Sch23b] Schabacher, Gabriele and Spallinger, Sophie: "Pilotversuche am Bahnhof". In: *ZfM - Zeitschrift für Medienwissenschaft* 15.29 (2023), pp. 35–50. doi: doi:10.14361/zfmw-2023-150205. url: https://doi.org/10.14361/zfmw-2023-150205 (visited on 09/04/2025) (cit. on p. 3).

[Sch24] Schlegel, Kevin; Jiang, Lei and Ni, Hao: "Using joint angles based on the international biomechanical standards for human action recognition and related tasks". In: *arXiv preprint arXiv:2406.17443* (2024) (cit. on p. 117).

[Sch25] Schwarzbeck, Martin: KI-Kameras in Hamburg: „Schaufenster in die Zukunft der Polizeiarbeit". Publisher: netzpolitik.org. Aug. 17, 2025. url: https://netzpolitik.org/2025/ki-kameras-in-hamburg-schaufenster-in-die-zukunft-der-polizeiarbeit/ (visited on 09/04/2025) (cit. on p. 4).

[Sha16] Shahroudy, Amir; Liu, Jun; Ng, Tian-Tsong and Wang, Gang: "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on pp. 72, 73, 131).

[Sha22] Shah, Anshul; Mishra, Shlok; Bansal, Ankan; Chen, Jun-Cheng; Chellappa, Rama and Shrivastava, Abhinav: "Pose and Joint-Aware Action Recognition". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2022, pp. 3850–3860 (cit. on p. 135).

[Shi22] Shi, Dahu; Wei, Xing; Li, Liangqi; Ren, Ye and Tan, Wenming: "End-to-End Multi-Person Pose Estimation With Transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 11069–11078 (cit. on p. 19).

[Sig10] Sigal, Leonid; Balan, Alexandru O and Black, Michael J: "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion". In: *International journal of computer vision* 87.1 (2010). Publisher: Springer, pp. 4–27 (cit. on pp. 61, 62, 69).

[Son21] Song, Yi-Fan; Zhang, Zhang; Shan, Caifeng and Wang, Liang: "Richly Activated Graph Convolutional Network for Robust Skeleton-Based Action Recognition". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.5 (2021), pp. 1915–1925. doi: 10.1109/TCSVT.2020.3015051 (cit. on p. 135).

[Son24] Song, Xingyu; Li, Zhan; Chen, Shi and Demachi, Kazuyuki: "Quater-GCN: enhancing 3D human pose estimation with orientation and semi-supervised training". In: *ECAI 2024*. IOS Press, 2024, pp. 121–128 (cit. on pp. 122, 125, 126).

[Spe22a] Specker, Andreas; Florin, Lucas; Cormier, Mickael and Beyerer, Jürgen: "Improving Multi-Target Multi-Camera Tracking by Track Refinement and Completion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2022, pp. 3199–3209 (cit. on p. 36).

[Spe22b] Specker, Andreas; Moritz, Lennart; Cormier, Mickael and Beyerer, Jürgen: "Fast and Lightweight Online Person Search for Large-Scale Surveillance Systems". In: *Proceedings of the*

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. 2022, pp. 570–580 (cit. on p. 36).

[Spe23] Specker, Andreas; Cormier, Mickael and Beyerer, Jürgen: "UPAR: Unified Pedestrian Attribute Recognition and Person Retrieval". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 981–990 (cit. on pp. 13, 37, 47, 99).

[Spe24] Specker, Andreas Heinrich: "Attribute-Based Person Retrieval in Multi-Camera Networks". PhD thesis. Karlsruher Institut für Technologie (KIT), 2024. 277 pp. doi: 10.5445/IR/1000169837 (cit. on pp. 36, 47, 55).

[Sta17] Statista: FOCUS: Bedeutet für Sie mehr Videoüberwachung eher mehr Sicherheit oder eher einen Eingriff in Ihre persönlichen Freiheitsrechte? 2017. url: https://de.statista.com/statistik/daten/studie/655303/umfrage/sicherheitsgefuehl-durch-videoueberwachung-in-deutschland/ (cit. on p. 3).

[Sta24] Stadler, Daniel Bernhard: "Utilization of Occluded Detections and Target Information in Multi-Person Tracking". PhD thesis. Karlsruher Institut für Technologie (KIT), 2024. 265 pp. doi: 10.5445/IR/1000176028 (cit. on p. 36).

[Sta25] Stadler, Daniel and Specker, Andreas: "A Strong Baseline for Multi-Person Tracking in Thermal Infrared Imagery". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2025, pp. 4549–4559 (cit. on p. 42).

[Str23] StreetPress: À Reims, l'intelligence artificielle filme les gares : la vidéosurveillance policière scrutée. Publisher: StreetPress. July 25, 2023. url: https://www.streetpress.com/sujet/1674664403-reims-intelligence-artificielle-camera-surveillance-police-donnees-thales (visited on 09/04/2025) (cit. on p. 4).

[Str24] StreetPress: Leclerc, Fnac, Biocoop… : des commerces sanctionnés pour vidéosurveillance illégale à l'IA. Publisher: StreetPress. May 2, 2024. URL: https://www.streetpress.com/sujet/1687789862-leclerc-fnac-biocoop-commerces-videosurveillance-intelligence-artificielle-illegal (visited on 09/04/2025) (cit. on p. 4).

[Sul18] Sultani, Waqas; Chen, Chen and Shah, Mubarak: "Real-World Anomaly Detection in Surveillance Videos". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018 (cit. on p. 72).

[Sun19] Sun, Ke; Xiao, Bin; Liu, Dong and Wang, Jingdong: "Deep High-Resolution Representation Learning for Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 19, 80).

[Sun24] Sun, Pengzhan; Gu, Kerui; Wang, Yunsong; Yang, Linlin and Yao, Angela: "Rethinking Visibility in Human Pose Estimation: Occluded Pose Reasoning via Transformers". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2024, pp. 5903–5912 (cit. on p. 19).

[Tan23] Tang, Zitian; Ye, Wenjie; Ma, Wei-Chiu and Zhao, Hang: "What Happened 3 Seconds Ago? Inferring the Past With Thermal Imaging". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 17111–17120 (cit. on p. 43).

[Tan24a] Tan, Dayi; Chen, Hansheng; Tian, Wei and Xiong, Lu: "DiffusionRegPose: Enhancing Multi-Person Pose Estimation using a Diffusion-Based End-to-End Regression Approach". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 2230–2239 (cit. on p. 20).

[Tan24b] Tanaka, Ryota; Suzuki, Tomohiro and Fujii, Keisuke: "3D Pose-Based Temporal Action Segmentation for Figure Skating:

A Fine-Grained and Jump Procedure-Aware Annotation Approach". In: *Proceedings of the 7th ACM International Workshop on Multimedia Content Analysis in Sports*. 2024, pp. 17–26 (cit. on p. 61).

[Tos14] Toshev, Alexander and Szegedy, Christian: "DeepPose: Human Pose Estimation via Deep Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014 (cit. on pp. 18, 145, 148–152, 154, 156).

[Unk25a] Unknown: CodeWalker - GTA V 3D Map, Animations and RPF Explorer. 2025. URL: https://github.com/dexyfex/CodeWalker (cit. on pp. 172, 176).

[Unk25b] Unknown: Script Hook V. 2025. URL: http://www.dev-c.com/gtav/scripthookv/ (cit. on p. 172).

[Von18] Von Marcard, Timo; Henschel, Roberto; Black, Michael J; Rosenhahn, Bodo and Pons-Moll, Gerard: "Recovering accurate 3d human pose in the wild using imus and a moving camera". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 601–617 (cit. on pp. 61, 62, 70).

[Wan14] Wang, Jiang; Nie, Xiaohan; Xia, Yin; Wu, Ying and Zhu, Song-Chun: "Cross-View Action Modeling, Learning, and Recognition". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 2649–2656. DOI: 10.1109/CVPR.2014.339 (cit. on pp. 72, 73, 131).

[Wan21] Wang, Jingdong et al.: "Deep High-Resolution Representation Learning for Visual Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2021), pp. 3349–3364. DOI: 10.1109/TPAMI.2020.2983686 (cit. on pp. 79, 80, 145, 148–152, 154, 156, 268).

[Wan22] Wang, Wenhai; Xie, Enze; Li, Xiang; Fan, Deng-Ping; Song, Kaitao; Liang, Ding; Lu, Tong; Luo, Ping and Shao, Ling: "PVT v2: Improved baselines with pyramid vision transformer". In: *Computational Visual Media* 8.3 (2022), pp. 415–424. DOI: 10.1007/s41095-022-0274-8 (cit. on p. 80).

[Wan24a] Wang, Dongkai and Zhang, Shiliang: "Spatial-Aware Regression for Keypoint Localization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 624–633 (cit. on p. 18).

[Wan24b] Wang, Guoquan; Liu, Mengyuan; Liu, Hong; Guo, Peini; Wang, Ti; Guo, Jingwen and Fan, Ruijia: "Augmented skeleton sequences with hypergraph network for self-supervised group activity recognition". In: *Pattern Recognition* 152 (2024), p. 110478. doi: https://doi.org/10.1016/j.patcog.2024.110478. url: https://www.sciencedirect.com/science/article/pii/S0031320324002292 (cit. on p. 30).

[Win09] Winter, David A: Biomechanics and motor control of human movement. John wiley & sons, 2009 (cit. on p. 269).

[Wu17] Wu, Jiahong; Zheng, He; Zhao, Bo; Li, Yixin; Yan, Baoming; Liang, Rui; Wang, Wenjia; Zhou, Shipei; Lin, Guosen; Fu, Yanwei et al.: "Ai challenger: A large-scale dataset for going deeper in image understanding". In: *arXiv preprint arXiv:1711.06475* (2017) (cit. on pp. 42, 43).

[Xia18] Xiao, Bin; Wu, Haiping and Wei, Yichen: "Simple Baselines for Human Pose Estimation and Tracking". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (cit. on pp. 18, 80).

[Xia25] Xia, Yan; Zhou, Xiaowei; Vouga, Etienne; Huang, Qixing and Pavlakos, Georgios: "Reconstructing Humans with a Biomechanically Accurate Skeleton". In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. June 2025, pp. 5355–5365 (cit. on p. 24).

[Xu22] Xu, Yufei; Zhang, Jing; ZHANG, Qiming and Tao, Dacheng: "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation". In: *Advances in Neural Information Processing Systems*. Ed. by Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K. and Oh, A. Vol. 35. Curran Associates, Inc., 2022, pp. 38571–38584. url: https://proceedings.neurips.cc/paper_

files/paper/2022/file/fbb10d319d44f8c3b4720873e4177c65-Paper-Conference.pdf (cit. on pp. 21, 77, 79, 80, 145, 148–152, 154, 156).

[Xu24] Xu, Liang et al.: "Inter-X: Towards Versatile Human-Human Interaction Analysis". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 22260–22271 (cit. on pp. 172, 178).

[Yan19] Yang, Fan; Wu, Yang; Sakti, Sakriani and Nakamura, Satoshi: "Make skeleton-based action recognition model smaller, faster and better". In: *Proceedings of the 1st ACM International Conference on Multimedia in Asia*. 2019, pp. 1–6 (cit. on p. 28).

[Yan21a] Yang, Di; Wang, Yaohui; Dantcheva, Antitza; Garattoni, Lorenzo; Francesca, Gianpiero and Bremond, Francois: "UNIK: A Unified Framework for Real-world Skeleton-based Action Recognition". In: *BMVC* (2021) (cit. on p. 31).

[Yan21b] Yang, Sen; Quan, Zhibin; Nie, Mu and Yang, Wankou: "TransPose: Keypoint Localization via Transformer". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 (cit. on pp. 20, 80).

[Yan23a] Yan, Hong; Liu, Yang; Wei, Yushen; Li, Zhen; Li, Guanbin and Lin, Liang: "SkeletonMAE: Graph-based Masked Autoencoder for Skeleton Sequence Pre-training". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 5606–5618 (cit. on pp. 31, 32).

[Yan23b] Yang, Jie; Zeng, Ailing; Liu, Shilong; Li, Feng; Zhang, Ruimao and Zhang, Lei: "Explicit Box Detection Unifies End-to-End Multi-Person Pose Estimation". In: *International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=s4WVupnJjmX (cit. on p. 19).

[Yan25] Yang, Jie; Zeng, Ailing; Ren, Tianhe; Liu, Shilong; Li, Feng; Zhang, Ruimao and Zhang, Lei: "ED-Pose++: Enhanced Explicit Box Detection for Conventional and Interactive Multi-Object Keypoint Detection". In: *IEEE Transactions on Pattern*

Analysis and Machine Intelligence 47.7 (2025), pp. 5636–5654. DOI: 10.1109/TPAMI.2025.3555527 (cit. on p. 19).

[Yeu25] YEUNG, Calvin; SUZUKI, Tomohiro; TANAKA, Ryota; YIN, Zhuoer and FUJII, Keisuke: "AthletePose3D: A Benchmark Dataset for 3D Human Pose Estimation and Kinematic Validation in Athletic Movements". In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*. June 2025, pp. 5945–5956 (cit. on pp. 61, 62, 67, 109).

[You13] YOUGOV (VIA IFSEC GLOBAL): 81% of Europeans say CCTV helps reduce crime. 2013. URL: https://www.ifsecglobal.com/infographic-survey-reveals-support-for-surveillance/ (cit. on p. 3).

[YUA21] YUAN, YUHUI; FU, Rao; HUANG, Lang; LIN, Weihong; ZHANG, Chao; CHEN, Xilin and WANG, Jingdong: "HRFormer: High-Resolution Vision Transformer for Dense Predict". In: *Advances in Neural Information Processing Systems*. Ed. by RANZATO, M.; BEYGELZIMER, A.; DAUPHIN, Y.; LIANG, P. S. and VAUGHAN, J. Wortman. Vol. 34. Curran Associates, Inc., 2021, pp. 7281–7293. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/3bbfdde8842a5c44a0323518eec97cbe-Paper.pdf (cit. on pp. 21, 80).

[Zen22a] ZENG, Ailing; JU, Xuan; YANG, Lei; GAO, Ruiyuan; ZHU, Xizhou; DAI, Bo and XU, Qiang: "DeciWatch: A Simple Baseline for 10x Efficient 2D and 3D Pose Estimation". In: *European Conference on Computer Vision*. Springer, 2022 (cit. on pp. 22, 132).

[Zen22b] ZENG, Ailing; YANG, Lei; JU, Xuan; LI, Jiefeng; WANG, Jianyi and XU, Qiang: "SmoothNet: A Plug-and-Play Network for Refining Human Poses in Videos". In: *European Conference on Computer Vision*. Springer, 2022 (cit. on p. 22).

[Zha19a] ZHANG, Song-Hai; LI, Ruilong; DONG, Xin; ROSIN, Paul; CAI, Zixi; HAN, Xi; YANG, Dingcheng; HUANG, Haozhi and HU, Shi-Min: "Pose2Seg: Detection Free Human Instance Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR). June 2019 (cit. on pp. 23, 37, 42, 43).

[Zha19b] Zhao, Long; Peng, Xi; Tian, Yu; Kapadia, Mubbasir and Metaxas, Dimitris N.: "Semantic Graph Convolutional Networks for 3D Human Pose Regression". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 25, 86).

[Zha20] Zhang, Feng; Zhu, Xiatian; Dai, Hanbin; Ye, Mao and Zhu, Ce: "Distribution-Aware Coordinate Representation for Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 (cit. on p. 19).

[Zha22] Zhang, Jinlu; Tu, Zhigang; Yang, Jianyu; Chen, Yujin and Yuan, Junsong: "MixSTE: Seq2seq Mixed Spatio-Temporal Encoder for 3D Human Pose Estimation in Video". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 13232–13242 (cit. on pp. 25, 86, 161, 162).

[Zha23a] Zhao, Qitao; Zheng, Ce; Liu, Mengyuan and Chen, Chen: "A Single 2D Pose with Context is Worth Hundreds for 3D Human Pose Estimation". In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023 (cit. on pp. 25, 86).

[Zha23b] Zhao, Qitao; Zheng, Ce; Liu, Mengyuan; Wang, Pichao and Chen, Chen: "PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 8877–8886 (cit. on pp. 25, 86, 87, 161, 162).

[Zha24] Zhang, Youliang; Liu, Wenxuan; Xu, Danni; Zhou, Zhuo and Wang, Zheng: "Bi-Causal: Group Activity Recognition via Bidirectional Causality". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 1450–1459 (cit. on p. 30).

[Zhe15] Zheng, Liang; Shen, Liyue; Tian, Lu; Wang, Shengjin; Wang, Jingdong and Tian, Qi: "Scalable person re-identification: A benchmark". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015 (cit. on p. 44).

[Zhe21] Zheng, Ce; Zhu, Sijie; Mendieta, Matias; Yang, Taojiannan; Chen, Chen and Ding, Zhengming: "3D Human Pose Estimation With Spatial and Temporal Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 11656–11665 (cit. on pp. 25, 77, 83, 85–88, 161, 162).

[Zho18] Zhou, Xiaowei; Zhu, Menglong; Pavlakos, Georgios; Leonardos, Spyridon; Derpanis, Konstantinos G and Daniilidis, Kostas: "Monocap: Monocular human motion capture using a cnn coupled with a geometric prior". In: *IEEE transactions on pattern analysis and machine intelligence* 41.4 (2018). Publisher: IEEE, pp. 901–914 (cit. on p. 24).

[Zho19] Zhou, Yi; Barnes, Connelly; Lu, Jingwan; Yang, Jimei and Li, Hao: "On the Continuity of Rotation Representations in Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 118, 119).

[Zho22a] Zhou, Honglu; Kadav, Asim; Shamsian, Aviv; Geng, Shijie; Lai, Farley; Zhao, Long; Liu, Ting; Kapadia, Mubbasir and Graf, Hans Peter: "COMPOSER: Compositional Reasoning of Group Activity in Videos with Keypoint-Only Modality". In: *Proceedings of the 17th European Conference on Computer Vision (ECCV 2022)* (2022) (cit. on p. 30).

[Zho22b] Zhou, Yuxuan; Cheng, Zhi-Qi; Li, Chao; Fang, Yanwen; Geng, Yifeng; Xie, Xuansong and Keuper, Margret: "Hypergraph transformer for skeleton-based action recognition". In: *arXiv preprint arXiv:2211.09590* (2022) (cit. on pp. 29, 32, 166).

[Zho25] Zhong, Zeyun; Li, Tianrui; Martin, Manuel; Cormier, Mickael; Wu, Chengzhi; Diederichs, Frederik and Beyerer, Jürgen: "HybridFormer: Bridging Local and Global Spatio-Temporal Dynamics for Efficient Skeleton-Based Action Recognition". In: *European Conference on Computer Vision*. Springer, 2025, pp. 19–35 (cit. on pp. 29, 32, 166).

[Zhu23] Zhu, Wentao; Ma, Xiaoxuan; Liu, Zhaoyang; Liu, Libin; Wu, Wayne and Wang, Yizhou: "MotionBERT: A Unified Perspective on Learning Human Motion Representations". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023 (cit. on pp. 25, 29, 32, 64, 69, 86–89, 161–163, 166).

Own Publications

[1] Cormier, Mickael: "A Data Annotation Process for Human Activity Recognition in Public Places". In: *Proceedings of the 2020 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Ed.: J. Beyerer; T. Zander*. Vol. 51. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. KIT Scientific Publishing, 2021, pp. 33–51.

[2] Cormier, Mickael; Moshkenan, Houraalsadat Mortazavi; Lörch, Franz; Metzler, Jürgen and Beyerer, Jürgen: "Do as we do: Multiple Person Video-To-Video Transfer". In: *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. 2021, pp. 84–90. doi: 10.1109/MIPR51284.2021.00020.

[3] Cormier, Mickael; Röpke, Fabian; Golda, Thomas and Beyerer, Jürgen: "Interactive Labeling for Human Pose Estimation in Surveillance Videos". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2021, pp. 1649–1658.

[4] Cormier, Mickael; Seletkov, Dmitrii and Beyerer, Jürgen: "Towards Lower Precision Quantization for Pedestrian Detection in Crowded Scenario". In: *IEEE EUROCON 2021 - 19th International Conference on Smart Technologies*. 2021, pp. 254–258. doi: 10.1109/EUROCON52738.2021.9535539.

[5] Cormier, Mickael; Wolf, Stefan; Sommer, Lars; Schumann, Arne and Beyerer, Jürgen: "Fast Pedestrian Detection for Real-World Crowded Scenarios on Embedded GPU". In: *IEEE EUROCON 2021 -*

19th International Conference on Smart Technologies. 2021, pp. 40–44. DOI: 10.1109/EUROCON52738.2021.9535550.

[6] Cormier, Mickael: "A Simple Pyramid Vision Transformer for Human Pose Estimation in Crowds". In: *Proceedings of the 2021 Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory*. Joint Workshop of Fraunhofer IOSB and Institute for Anthropomatics, Vision and Fusion Laboratory. Vol. 54. Karlsruher Schriften zur Anthropomatik / Lehrstuhl für Interaktive Echtzeitsysteme, Karlsruher Institut für Technologie ; Fraunhofer-Inst. für Optronik, Systemtechnik und Bildauswertung IOSB Karlsruhe. ISSN: 1863-6489. Karlsruher Institut für Technologie (KIT), 2022, pp. 33–51.

[7] Cormier, Mickael; Clepe, Aris; Specker, Andreas and Beyerer, Jürgen: "Where Are We With Human Pose Estimation in Real-World Surveillance?" In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2022, pp. 591–601.

[8] Golda, Thomas; Thiemich, Johanna; Cormier, Mickael and Beyerer, Jürgen: "For the Sake of Privacy: Skeleton-Based Salient Behavior Recognition". In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 3983–3987. DOI: 10.1109/ICIP46576.2022.9897358.

[9] Specker, Andreas; Florin, Lucas; Cormier, Mickael and Beyerer, Jürgen: "Improving Multi-Target Multi-Camera Tracking by Track Refinement and Completion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2022, pp. 3199–3209.

[10] Specker, Andreas; Moritz, Lennart; Cormier, Mickael and Beyerer, Jürgen: "Fast and Lightweight Online Person Search for Large-Scale Surveillance Systems". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2022, pp. 570–580.

[11] Cormier, Mickael; Specker, Andreas; Junior, Julio C. S. Jacques; Florin, Lucas; Metzler, Jürgen; Moeslund, Thomas B.; Nasrollahi, Kamal; Escalera, Sergio and Beyerer, Jürgen: "UPAR Challenge: Pedestrian Attribute Recognition and Attribute-Based Person Retrieval – Dataset, Design, and Results". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2023, pp. 166–175.

[12] Golda, Thomas; Cormier, Mickael and Beyerer, Jürgen: "Intelligente Bild- und Videoauswertung für die Sicherheit". In: *Handbuch Polizeimanagement: Polizeipolitik – Polizeiwissenschaft – Polizeipraxis*. Ed. by Wehe, Dieter and Siller, Helmut. Wiesbaden: Springer Fachmedien Wiesbaden, 2023, pp. 1487–1507. doi: 10.1007/978-3-658-34388-0_87. url: https://doi.org/10.1007/978-3-658-34388-0_87.

[13] Specker, Andreas; Cormier, Mickael and Beyerer, Jürgen: "UPAR: Unified Pedestrian Attribute Recognition and Person Retrieval". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2023, pp. 981–990.

[14] Cormier, Mickael; Schmid, Yannik and Beyerer, Jürgen: "Enhancing Skeleton-Based Action Recognition in Real-World Scenarios Through Realistic Data Augmentation". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2024, pp. 290–299.

[15] Cormier, Mickael; Specker, Andreas; Junior, Julio C. S. Jacques; Moritz, Lennart; Metzler, Jürgen; Moeslund, Thomas B.; Nasrollahi, Kamal; Escalera, Sergio and Beyerer, Jürgen: "UPAR Challenge 2024: Pedestrian Attribute Recognition and Attribute-Based Person Retrieval - Dataset, Design, and Results". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. 2024, pp. 359–367.

[16] Cormier, Mickael; Yi, Caleb Ng Zhi; Specker, Andreas; Blaß, Benjamin; Heizmann, Michael and Beyerer, Jürgen: "Leveraging Thermal Imaging for Robust Human Pose Estimation in Low-Light Vision". In: *Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops*. 2024, pp. 67–83.

[17] Cormier, Mickael; Specker, Andreas and Beyerer, Jürgen: "UPPET: Unified Pedestrian Pose Estimation in Thermal Imaging". In: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*. 2025, pp. 4551–4560.

[18] Zhong, Zeyun; Li, Tianrui; Martin, Manuel; Cormier, Mickael; Wu, Chengzhi; Diederichs, Frederik and Beyerer, Jürgen: "Hybrid-Former: Bridging Local and Global Spatio-Temporal Dynamics for Efficient Skeleton-Based Action Recognition". In: *European Conference on Computer Vision*. Springer, 2025, pp. 19–35.

[19] Cormier, Mickael; Specker, Andreas and Beyerer, Jürgen: "UPAR-Pose: Joint pedestrian pose estimation and attribute recognition". In: *Under Review*. 2026.

[20] Cormier, Mickael; Zolk, Jeremy; Laubenheimer, Astrid and Beyerer, Jürgen: "Uplifting 2D to 3D Human Poses with Joint Rotations and Bone Constraints: A Strong Baseline for Sports and Fitness Applications". In: *2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG)*. 2026.

List of Figures

List of Tables

Acronyms

2D-HPE	2D Human Pose Estimation
3D-HPE	3D Human Pose Estimation
3FCV	3-Fold Cross-Validation
4FCV	4-Fold Cross-Validation
AAD	Average Angular Distance
ACC	Top-1 Accuracy
Acc@1	Top-1 Accuracy
AMASS	Archive of Motion Capture as Surface Shapes
Antonn	Annotations Tool over Neural Network
AP	Average Precision
AP3D	AthletePose3D
AP_{50}	Average Precision @ 0.50
CAMEL-P	CAMEL - Pose
CNN	Convolutional Neural Network

DA	Data Augmentation
DoF	Degree of Freedom
DSTformer	Dual-stream Spatio-Temporal Transformer
EMA	Exponential Moving Average
F1	F1
FC	Fully-Connected
GCN	Graph Convolutional Neural Network
GT	Ground Truth
GTA-RWS	GTA Real World Surveillance
H36M	Human3.6M
H4D	Harmony4D
HPE	Human Pose Estimation
IK	Inverse Kinematics
IoU	Intersection Over Union
ISM	Implicit Shape Model
LLVIP	Low-Light Visible-Infrared Paired
LLVIP-P	Low-Light Visible-Infrared Paired - Pose
LOOCV	Leave-One-Out Cross-Validation

mA Mean Accuracy

MAPCA Mean Average Per-Class Accuracy

Market-1501 Market-1501

MoCap Motion Capture

MotionBERT Motion Bidirectional Encoder Representations from Transformers

MPBLE Mean Per Bone Length Error

MPJAE Mean Per Joint Angle Error

MPJPE Mean Per Joint Position Error

MPJVE Mean Per Joint Velocity Error

MSE Mean Squared Error

N-MPJPE Normalized MPJPE

OTP OpenThermalPose

PA-100K Pedestrian Attribute 100K

PA-24K Pedestrian Attribute and Pose Estimation 24K

PAR Pedestrian Attribute Recognition

PED Pedestrian

PETA Pedestrian Attribute

P-MPJPE Procrustes-aligned Mean Per Joint Position Error

RAPv1 Richly Annotated Pedestrian v1

RAPv2 Richly Annotated Pedestrian v2

SBAR Skeleton Based Action Recognition

Smarthome Toyota Smarthome

SMPL Skinned Multi-Person Linear

TPE Thermisch-PE

TSA Task-specific Adapter

UPAR Unified Pedestrian Attribute Recognition

UPAR-Pose Unified Pedestrian Attribute Recognition and Pose Estimation

UPPET Unified Pedestrian Pose Estimation in Thermal Imaging

Symbols

ACC	Top-1 Accuracy, the proportion of correctly classified samples based on the top predicted class
$\mathbf{R}'$	Alignment rotation matrix between predicted and ground truth rotations for a joint
AP	Average Precision, the area under the precision-recall curve for joint detection across OKS thresholds
$\mathbf{AP}_L$	Average Precision for large-scale persons (bounding box area greater than 96^2)
$\mathbf{AP}_M$	Average Precision for medium-scale persons (bounding box area between 32^2 and 96^2)
$\mathbf{AP}_S$	Average Precision for small-scale persons (bounding box area less than 32^2)
AR	Average Recall, the average fraction of correctly localized joints across all recall levels
$\mathbf{B}$	Ground truth bone vector in 3D space
$\hat{\mathbf{B}}$	Predicted bone vector in 3D space

$\mathcal{D}_{2D}$	Dataset of 2D cropped images and corresponding 2D keypoint annotations
$\mathcal{D}_{3D}$	Dataset of 2D keypoints and corresponding 3D joint annotations for temporal sequences
$d_{L2}(\hat{\mathbf{k}}_s^r, \mathbf{k}_s^r)$	Euclidean distance between predicted and ground truth coordinates for keypoint s in sample r
$\mathbf{F1_{PAR}}$	Instance-based F1 score for pedestrian attribute recognition
FN	False negatives for a given threshold
$\mathbf{FN}_c$	False negatives for class c in skeleton-based action recognition
FP	False positives for a given threshold
f	Function that returns the recognized person attributes for an image
J	Ground truth joint coordinates in 3D space
$\mathbf{K}^r$	Ground truth set of 2D keypoint coordinates for image r, i.e., $\mathbf{K}^r = \{\mathbf{k}_s^r\}_{s=1}^{\mathrm{N_{keypoints}}}$
R	Ground truth rotation matrices for all joints
I	Image
$\mathbf{IoU_{set}}$	Set of evaluated IoU thresholds (e.g., [0.50, 0.95])
$\hat{\mathbf{J}}^{\mathbf{aligned}}$	Procrustes-aligned predicted joint position in 3D space
L	Number of semantic attributes

$\mathcal{L}_{3d}$	Overall loss function for training
$\mathcal{L}_{\textbf{angular}}$	Average Angular Distance (AAD) loss function for quaternions
$\mathcal{L}_{\textbf{geodesic}}$	Geodesic loss function for rotation matrices
$\mathcal{L}_{K_{v2}}$	Kinematics loss with flow, velocity and acceleration
$\mathcal{L}_{\textbf{MSE}}$	MSE loss function
$\mathcal{L}_{\textbf{MPJPE}}$	MPJPE loss
$\mathcal{L}_{\textbf{scale}}$	Scale loss (normalized MPJPE)
mA	Label-based mean accuracy
MAPCA	Mean Average Per-Class Accuracy, measuring the average classification performance per class, addressing class imbalance
f	Mapping function from cropped image to 2D keypoints
MPBLE	Mean Per Bone Length Error, measuring the average discrepancy in predicted and ground truth bone lengths
MPJAE	Mean Per Joint Angle Error
MPJPE	Mean Per Joint Position Error
MPJVE	Mean Per Joint Velocity Error
$\mathbf{N}_{\textbf{bones}}$	Total number of bones in the skeleton
N	Number of negative samples

$\mathbf{N_{joints}}$	Number of joints per 3D pose
$\mathbf{N_{keypoints}}$	Number of keypoints per 2D pose
$\mathbf{N^{2D}_{samples}}$	Number of samples in the 2D dataset
$\mathbf{N^{3D}_{samples}}$	Number of samples in the 3D dataset
f	Number of frames in the input sequence
OKS	Object Keypoint Similarity, a normalized measure of similarity between predicted and ground truth joints
ϕ_{2D}	Learnable parameters of the 2D Human Pose Estimation (2D-HPE) model
ϕ_{3D}	Learnable parameters of the 3D Human Pose Estimation (3D-HPE) model
P-MPJPE	Procrustes-aligned Mean Per Joint Position Error
P	Number of positive samples
$\mathbf{Prec_{PAR}}$	Instance-based precision score for pedestrian attribute recognition
$\hat{\mathbf{J}}$	Predicted joint coordinates in 3D space
$\hat{\mathbf{K}}^r$	Predicted set of 2D keypoint coordinates for image r, i.e., $\hat{\mathbf{K}}^r = \{\hat{\mathbf{k}}^r_s\}_{s=1}^{\mathrm{N_{keypoints}}}$
$\tilde{\mathbf{R}}$	Predicted rotation matrices for all joints
$\bar{q}_{t,j}$	Ground truth quaternion for joint j at frame t

$q_{t,j}$	Predicted quaternion for joint j at frame t
$\mathbf{R_{set}}$	Set of evaluated recall thresholds
$\mathbf{Rec_{PAR}}$	Instance-based recall score for pedestrian attribute recognition
$\mathbf{J_{root}}$	3D position of the root joint (e.g., pelvis)
$\tilde{\mathbf{J}}$	Sequence of 3D joint positions relative to the root joint
s	Scale of the person, typically defined as a function of the bounding box area (used in OKS computation)
σ_s	Per-keypoint constant σ_s accounting for localization uncertainty, as defined in COCO evaluation protocol
SO(3)	The rotation group of 3D rotations
TN	Number of true negatives
C_{SBAR}	Total number of action classes in the SBAR dataset
$\mathbf{N_{samples}}$	Total number of samples in the evaluation set for skeleton-based action recognition tasks
TP	Number of true positives
$\mathbf{TP}_c$	True positives for class c in skeleton-based action recognition
vis	Visibility of a joint, where 0: not labeled, 1: labeled but not visible, 2: labeled and visible
Y	Set of ground truth positive labels

A Antonn

Data annotation is a complex and costly process, particularly for large-scale, multi-view datasets. Designing dedicated tools to enable partial or full automation is therefore a critical step toward efficient data production. Nevertheless, ensuring the quality of the annotations requires continuous control and validation. Consequently, a transparent and reliable annotation process is essential.

In the context of human activity recognition in public spaces, this work formalizes five levels of automation for multi-view data annotation:

- **Level 0: No Automation**
  All annotations are performed manually. Annotation tools provide only basic functions for creating annotations.
- **Level 1: Tool Assistance**
  The annotator is assisted by tools that reduce manual effort. Given multiple views of the same scene, the annotator may switch between views while annotating, and annotations are propagated across views. Additionally, partially annotated sequences can be interpolated automatically into subsequent frames.
- **Level 2: Partly Automated Annotation**
  The annotator defines a region of interest using points, scribbles, or polygons. The tool automatically generates proposed annotations, which are manually corrected if needed.

- **Level 3: Highly Automated Annotation**
  Entire multi-view sequences are automatically annotated for the selected label types. Human annotators primarily conduct quality control and corrections.
- **Level 4: Highly Automated Annotation with Pre-Checking**
  Complete sequences are automatically annotated, and quality checks are generated for human validators to review and extend.

In this chapter, the software Antonn is introduced as a pivotal contribution that enabled the annotation of multiple large-scale datasets. The formalized annotation processes were initially proposed in [Cor21a], with substantial contributions from student assistants.

The primary goal of Antonn is to provide a web-based annotation platform with advanced control over image-based datasets. The system architecture comprises a backend, an API, and two frontends: a static Django frontend for data and annotator management, and a dynamic frontend forked from CVAT[1] for interactive annotation. The platform is designed to support all levels of automation described above, with current functionality between Level 2 and Level 3.

Key contributions of the platform include:

- Semi-automatic generation of annotation jobs,
- Functionality for creating dataset datasheets,
- Secure access control via Role-Based Access Control with Attributes (RBAC-A) with an open-source implementation[2],
- A secure and extensible API,
- An SDK for integration into external workflows,
- Distributed processing tools for large-scale annotation pipelines.

[1] https://github.com/cvat-ai/cvat
[2] https://github.com/MickaelCormier/django-rbaca

Overall, Antonn provides a scalable, flexible, and secure framework for multi-view human activity annotation, significantly reducing manual effort while ensuring high-quality, consistent labels across large datasets. Annotation workspaces for 2D-HPE are illustrated in Figures A.1 to A.3.

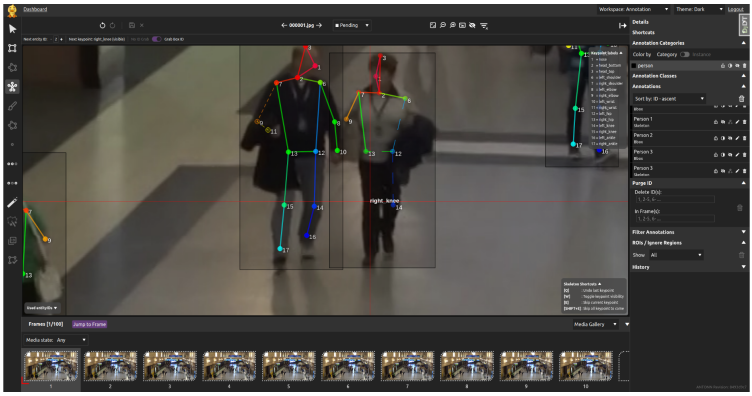


Figure A.1: Annotation workspace in Antonn – Screenshot of the standard annotation workspace used for detection, tracking, and keypoint annotations.

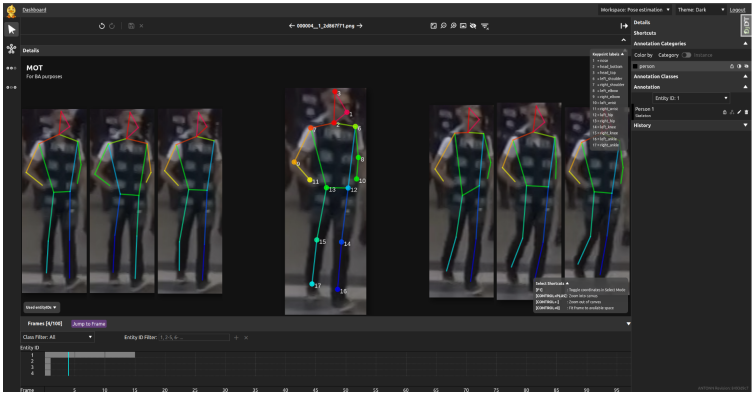


Figure A.2: Pose estimation workspace in Antonn – Screenshot of the advanced workspace dedicated to 2D-HPE.

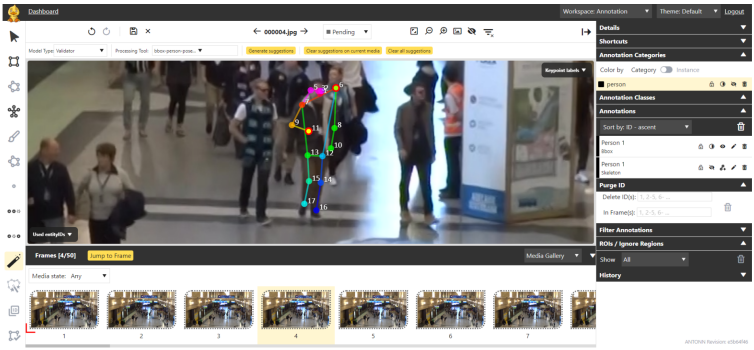


Figure A.3: Pose estimation workspace validation in Antonn – Screenshot of the interactive validation interface for 2D-HPE. Errors are circled in red.

B Annotation Process

A crucial element of the contributions presented in this thesis lies in the creation of novel datasets. To ensure their scientific value and reproducibility, a transparent documentation of the underlying annotation workflows is provided. This appendix therefore outlines the procedures, tools, and quality control steps that were employed in the generation of the annotated data.

Two datasets are discussed in detail: the UPAR dataset, which harmonizes semantic attributes across multiple sources, and the LLVIP-P-Pose benchmark, which extends visible–thermal pairs with fine-grained pose information. The annotation processes are described in Appendix B.1 and Appendix B.2, respectively, highlighting both methodological choices and practical considerations.

B.1 UPAR Annotation Process

This section details the harmonization strategy, annotation interfaces, and validation procedures that were used to construct the UPAR dataset.

To deliver UPAR, a large amount of supplementary annotations were contributed. The annotation process is further detailed and challenges are illustrated in Figure B.1. From the start annotators are informed that the datasets comprise camera shots of people under varying viewpoints, exposures, and crowd densities; the same individual may appear across frames. Attributes are labeled per frame, strictly based on visible evidence in that frame; thus, the same person can receive different labels across frames due to lighting, shadows, or partial visibility. Only perceived visual appearance is considered.

Figure B.1: UPAR Annotation challenges – For the first three images, the real color of the clothing differs from the color seen. In the fourth image, the upper body is missing. Sometimes the primary person is difficult to tell, as can be seen in the first two images of the second row. Do leggings make the lower-body clothing long? (yes) Do long boots make the lower body clothing long? (no).

B.1.1 UpperBodyClothingColor and LowerBodyClothingColor

For each frame, a primary person is selected and the body is conceptually split at the hips into upper-body and lower-body regions. Each region is assigned a color class for `UpperBodyClothingColor` and `LowerBodyClothingColor`, respectively. Eleven unique color classes are defined, supplemented by the labels *other* for unassignable colors and *unknown* for missing evidence.

- **Selection of the primary person:**
  - If a single person appears, this is the primary person.
  - If multiple people are present, the person most centrally framed and least occluded is selected.

    - If no unambiguous primary person can be identified, both regions are labeled *unknown*.

- **Selection of the primary color:**
    - The primary color is the class covering at least 50% of the visible clothing pixels in the respective region.
    - Different shades of the same color count toward the same class.
    - Colors that do not fit a unique class (*e.g.*, metallic tones, beige) are labeled *other*.
    - If the region is not visible or rendered in false colors (*e.g.*, grayscale), the label is *unknown*.
- **Additional color evidence:** If another color covers at least 10% of the region, the *mixture* flag is set.
- **Special cases (color):**
    - If no unique primary color exists (*e.g.*, uniformly striped or densely dotted clothing), the label is *other* and the *mixture* flag is set.
    - Infants are not selected as primary persons; instead, the accompanying adult is annotated, even if partially occluded.

B.1.2 LowerBodyClothingLength

Lower-body clothing length is annotated as *long*, *short*, or *unknown*.

- **Selection of the length class:**
    - Assign *long* if thighs and knees are clothed.
    - Assign *short* if parts of the thigh or knee are uncovered.
    - If the lower body is not fully visible and a clear decision is impossible, assign *unknown*.

- **Special cases (length):** If only one leg is fully visible, it is assumed to be representative of the entire lower-body clothing.

B.1.3 Age

A primary person is selected as above. If the person is clearly underage (child/teenager) or clearly adult/elderly, the corresponding label is assigned; otherwise, unknown is assigned.

B.1.4 Hair length

A primary person is selected as above. Annotators answer: (1) short: is the hair shorter than shoulder length? (2) long: is the hair at least shoulder length? (3) bald: is the person (partly) bald? (4) unknown: not clear (*e.g.*, hat, low resolution)? If multiple apply (*e.g.*, bald patch and short/long), all applicable cues are recorded per the scheme.

B.1.5 Glasses

A primary person is selected as above. Annotators answer: (1) glasses: is the person wearing glasses? (2) sunglasses: if yes, are these sunglasses? (3) unknown: difficult to tell (*e.g.*, head turned away).

B.1.6 Shortcuts for fast annotation

Annotations are created using Appendix A with task-specific shortcuts to improve speed without increasing errors. First, the color label distributions from Market-1501 Attribute, PETA, and RAPv2 are analyzed to identify the most frequent colors. Second, in dialogue with annotators, an ergonomic, keyboard-centric workflow is designed that avoids mandatory mouse use. The left hand switches images (a–d) and sets status (r: review, f: finished). The right hand, positioned on the numpad (thumb free; 4, 8, 6, Enter), handles

classification. As shown in Figure B.2, color codes are mapped to UpperBodyClothingColor and LowerBodyClothingColor: the first code sets the upper-body color, the second the lower-body color (*e.g.*, 5 then 88 yields upper-body white, lower-body black). Annotators study the printed color matrix and typically reach 12–20 frames/minute for first-pass color annotation and 6–12 frames/minute for validation.

Figure B.2: UPAR Annotation Shortcuts – Shortcuts keys designed to improve annotation speed. Given a numpad and the use of five fingers, we associate the most often used keys with the most common colors. Standart position for the finger are for the right hand: *thumb free, four, eight, six, enter.* The upperBodyColor is annotated with the first code, the lowerBodyColor is annotated with the second code. For instance, the sequence *5* and *88* result in annotating upperBodyColor-white and lowerBodyColor-black.

B.1.7 Validation process

Validation is performed by a separate team member (four-eyes principle) after initial labeling. Two error types are addressed: (1) operator errors—software misoperations (*e.g.*, duplicated primary color)—are detected and corrected directly; (2) perceptual inconsistencies—color perception can vary due to external factors (room lighting, monitor calibration) and internal factors (contrast sensitivity, color vision). Randomized frame ordering can thus yield inconsistent labels for borderline colors. A shared database of inconsistent frames is maintained. Through group discussion, each ambiguous case is resolved to a unique, consistent label; regular updates ensure consistent future decisions across annotators and sequences. Annotators are responsible for both labeling assigned sequences and validating peers' work while coordinating to maintain consistency.

B.2 LLVIP-Pose

This section describes the annotation of visible–thermal image pairs with person bounding boxes and skeletal keypoints, including guidelines for annotators and consistency checks across modalities.

B.2.1 Annotation Process

A semi-automated pipeline was used to reduce manual effort. Initial pose predictions were generated by a top-down HPE model (HRNet-W48-UDP [Wan21]) pre-trained on COCO. Bounding boxes were padded by 10% in width and height prior to inference to better enclose body parts. Predictions were visualized on thermal images and corrected for 2D keypoint locations and visibility. Since RGB-based estimators perform suboptimally in low light, an initial set of poses was annotated manually to bootstrap training; a thermal-based pose estimator was then trained and used to pre-annotate the remaining, more challenging sequences for correction.

B.2.2 Validation Process

Manual validation is time-consuming; an outlier detector was therefore designed to prioritize review and ensure consistency.

B.2.2.1 Anthropometric Detection of Outliers

Anthropometric regularities were used to evaluate pose plausibility via proportions of adjacent body segments. Following Drillis and Contini [Dri66], segment lengths relative to body height H were reparameterized as ratios that do not require H, accommodating truncation or occlusion (see Figure B.3). Ratios involving the hip are used due to anatomical coupling: hip–shoulder, thigh–torso, and torso–leg. Arm-related ratios were omitted due to high 3D variability not captured in 2D. Additionally, the hip–shoulder angle was included; in LLVIP, it is near zero for standing, walking, and riding.

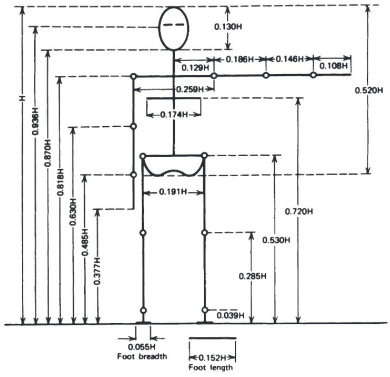


Figure B.3: Comparison of Body segments – Body segments with respect to height H (adapted from [Dri66, Win09])

Ground-truth values are listed in Tab. B.1. Individual poses are scored by averaging per-value distances between pose-derived values x_i and ground-truth y_i:

$$\text{Metric} = \frac{\sum_{i=1}^{n_{\text{values}}} d_i(x_i, y_i)}{n_{\text{values}}} \tag{B.1}$$

$$d_i(x_i, y_i) = \begin{cases} 1.0 - f(|x_i - y_i| \mid 0, \sigma_{|x_i - y_i|}) \cdot \sigma_{|x_i - y_i|}\sqrt{2\pi}, & \text{if } x_i \text{ present} \\ 1.0 - f(|x_{\text{approx},i} - y_i| \mid 0, \sigma_{|x_i - y_i|}) \cdot \sigma_{|x_i - y_i|}\sqrt{2\pi}, & \text{if } x_i \text{ missing} \end{cases} \tag{B.2}$$

with x_i: calculated values; y_i: ground-truth values; $x_{\text{approx},i}$: approximations for missing values; $\mu_{|x_i - y_i|}, \sigma_{|x_i - y_i|}$: mean and standard deviation of errors; μ_{x_i}, σ_{x_i}: mean and standard deviation of calculated values; n_{values}: number of values; n_x: number of available values. The Gaussian is

$$f(z \mid \mu_{|x_i - y_i|}, \sigma_{|x_i - y_i|}) = \frac{1}{\sigma_{|x_i - y_i|}\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{z - \mu_{|x_i - y_i|}}{\sigma_{|x_i - y_i|}}\right)^2} \tag{B.3}$$

Outliers are flagged when the pose score exceeds a data-driven threshold reflecting variability:

$$\text{Threshold} = \frac{\sum_{i=1}^{n_{\text{ratios}}} \left(1 - \frac{\sigma_i}{\mu_i}\right)}{n_{\text{ratios}}} \tag{B.4}$$

For the hip–shoulder angle, the inverse relation is used due to $\mu_{x_i} \gg \sigma_{x_i}$.

Missing values are approximated to avoid pose-dependent thresholds. Approximation uses the distribution of available values:

$$\frac{|x_{\text{approx},i} - y_i|}{\sigma_{|x_i - y_i|}} = \frac{\sum_{n=1}^{n_x} \frac{|x_i - y_i|}{\sigma_{|x_i - y_i|}}}{n_x} \tag{B.5}$$

Table B.1: Ratios used in the pose validator – Ground-truth values used in the pose validator

Ratio	Value
Hip–Shoulder	0.737
Hip–Shoulder Angle	0.0
Thigh–Torso	0.851
Torso–Leg	0.587

B.2.2.2 Validation Workflow

After scoring, poses are ranked to prioritize human validation. Validators correct the highest-scoring outliers per sequence. Mean and standard deviation of error terms before and after validation (Fig. B.4) consistently decrease, indicating improved annotation quality. Correcting roughly one quarter to one half of poses typically improves the overall metric by 0.135 points on average, with gains up to 0.277 in difficult cases, thereby reducing validation cost while preserving quality.

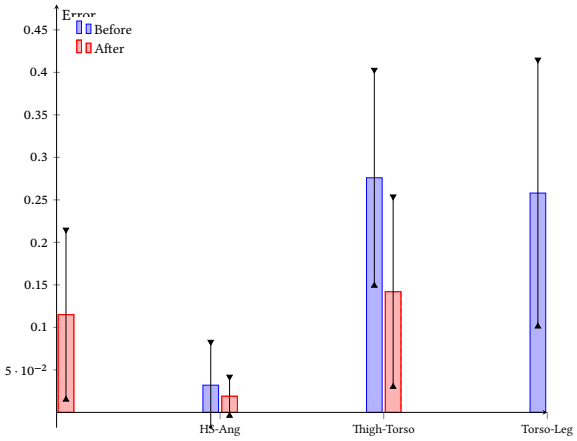


Figure B.4: Showcase of the Pose Validator' efficiency – Mean and standard deviation of error terms before and after validation, evidencing the effectiveness of the pose validator.)