

LEVERAGING SYNTHETIC DATA FOR VIDEO-BASED HUMAN ACTIVITY RECOGNITION

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)

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

Dissertation von
DAVID SCHNEIDER

geboren in Wolfach



Tag der mündlichen Prüfung: 15. Januar 2026

Hauptreferent: Prof. Dr.-Ing. Rainer Stiefelhagen
Korreferent: Prof. Dr. rer. nat. Abhinav Valada

ABSTRACT

Deep learning has advanced human activity recognition, enabling applications from healthcare to assistive robotics. Yet, the success of systems hinges on large-scale recordings of people in real environments, raising privacy concerns and failing to deliver diversity needed for robust deployment. When examining datasets and training pipelines, four limitations become apparent: (1) staged recordings are expensive, restricting scale and motivating a shift to web-sourced video; (2) social media data is strongly biased toward shareable content, omitting sensitive, mundane, or unflattering activities that are relevant in practice; (3) individuals appearing in such recordings rarely provide informed consent for machine learning use; and (4) data collection cannot keep pace with advances in compute, pushing the field toward a data-scarce regime where architectures cannot be exploited. This raises the question: can synthetic data serve not merely as a surrogate but as a principled tool to overcome limitations of data collected from the real world?

This thesis investigates synthetic data as a foundation for activity recognition along three directions. First, to study the synthetic-to-real domain gap, we introduce a benchmark pairing synthetic data from a life simulation game with real-world evaluation. Our analysis reveals performance degradation when models trained on synthetic videos are confronted with deployment conditions. To address this, we explore strategies: domain adaptation methods that calibrate model predictions to target distributions and multi-modal domain generation synthesizing diverse new modalities without direct access to the target domain. Together, these methods improve cross-domain robustness.

Second, we leverage parametric human body models for controllable large-scale rendering of activities. These models enable manipulation of pose, shape, viewpoint, and environment, yielding synthetic datasets that transfer to real benchmarks. By coupling parametric avatars with biomechanical simulation, we generate physiological quantities such as joint torques and muscle activations that are invisible to cameras. This opens up new research directions for estimating physical effort, injury risk, or rehabilitation-relevant quantities from vision, where large-scale measurement is infeasible.

Third, we target deployment scenarios where real data collection faces ethical and practical barriers, such as fall detection for elderly individuals. We show that synthetic fall videos with variation in demographics, body types, and environments can outperform real staged data when evaluated on accident datasets, challenging the assumption that synthetic data is inferior to real recordings. To enable privacy-preserving deployment, we develop anonymization methods that protect individuals and living spaces by combining avatar-based replacement of subjects with privacy mechanisms for background regions.

Overall, this thesis demonstrates that synthetic data can induce severe domain gaps and limit performance, but considering more advanced data generation methods it can also rival and even exceed real recordings. By providing controllability, demographic diversity, and privacy, synthetic data emerges as a central tool for building robust, ethically acceptable, and practically deployable human activity recognition systems.

ZUSAMMENFASSUNG

Deep Learning hat die Erkennung menschlicher Aktivitäten erheblich vorangebracht und Anwendungen von der Gesundheitsüberwachung bis zur assistiven Robotik ermöglicht. Der Erfolg heutiger Systeme beruht jedoch auf umfangreichen Aufnahmen realer Personen in realen Umgebungen, die gravierende Datenschutzbedenken aufwerfen und dennoch nicht die Vielfalt bieten, die für einen robusten Einsatz erforderlich wäre. Bei der Analyse bestehender Datensätze und Trainingspipelines treten vier zentrale Einschränkungen zutage: (1) Gestellte Aufnahmen sind kostspielig und logistisch aufwendig, was den Umfang begrenzt und den Einsatz webbasierten Videomaterials motiviert; (2) Daten aus sozialen Medien sind stark auf teilbare Inhalte verzerrt und blenden sensible, alltägliche oder unvorteilhafte Aktivitäten aus, die in der Praxis häufig besonders relevant sind; (3) die in solchen Aufnahmen gezeigten Personen erteilen nur selten eine informierte Einwilligung zur nachgelagerten Nutzung in Machine-Learning-Systemen; und (4) die Datenerhebung hält nicht mit den Fortschritten in Rechenleistung und Modellkapazität Schritt, sodass das Feld in ein absehbares Regime der Datenknappheit gedrängt wird, in dem verbesserte Architekturen nicht voll ausgeschöpft werden können. Daraus ergibt sich die grundlegende Frage, ob synthetische Datengenerierung nicht nur als Ersatz für reale Aufnahmen dienen kann, sondern als prinzipielles Werkzeug, um Einschränkungen zu überwinden, die mit realer Datenerhebung nicht lösbar sind.

Diese Dissertation untersucht synthetische Daten als systematische Grundlage für die menschliche Aktivitätserkennung entlang dreier Forschungsrichtungen. Erstens führen wir zur rigorosen Untersuchung der Synthetic-to-Real-Domänenlücke einen Benchmark ein, der synthetische Daten aus einem Lebenssimulationsspiel mit realweltlicher Evaluierung verknüpft. Unsere Analyse zeigt erhebliche Leistungseinbußen, wenn Modelle, die ausschließlich auf synthetischen Videos trainiert wurden, mit realen Einsatzbedingungen konfrontiert werden. Um dem zu begegnen, untersuchen wir komplementäre Strategien: Domänenanpassungsmethoden, die Modellvorhersagen an Zielverteilungen kalibrieren, sowie multimodale Domäengenerierung, die vielfältige neue Modalitäten synthetisiert, ohne direkten Zugriff auf die Ziel-Domäne zu benötigen. Zusammengekommen steigern diese Ansätze die Robustheit gegenüber Domänenverschiebungen deutlich.

Zweitens nutzen wir parametrische menschliche Körpermodelle für kontrollierbares, umfangreiches Rendering menschlicher Aktivitäten. Diese Modelle ermöglichen eine präzise Steuerung von Pose, Körperform, Kameraperspektive und Umgebung und erzeugen synthetische Datensätze, die sich effektiv auf reale Benchmarks übertragen lassen. Durch die Kopplung parametrischer Avatare mit biomechanischer Simulation generieren wir zudem latente physiologische Größen wie Muskelaktivierungen, die für Kameras unsichtbar sind. Dies eröffnet neue Forschungsrichtungen zur Schätzung körperlicher Belastung, Verletzungsrisiko oder rehabilitationsrelevanter Größen aus visuellen Daten.

Drittens adressieren wir Einsatzszenarien, in denen reale Datenerhebung auf grundlegende ethische und praktische Hürden stößt, etwa bei der Sturzerkennung für ältere Menschen. Wir zeigen, dass synthetische Sturzvideos mit systematischer Variation in Demografie, Körpertypen und Umgebungen reale gestellte Daten übertreffen können, wenn sie auf Datensätzen mit echten Unfällen evaluiert werden, und damit die verbre-

itete Annahme in Frage stellen, synthetische Daten seien realen Aufnahmen grundsätzlich unterlegen. Um einen datenschutzwahrenden Einsatz in sensiblen Kontexten zu ermöglichen, entwickeln wir Anonymisierungsmethoden, die sowohl Personen als auch ihre Wohnumgebungen schützen, indem sie avatarbasierte Ersetzungen der Probanden mit selektiven Privacy-Mechanismen für Hintergrundbereiche kombinieren.

Insgesamt zeigt diese Dissertation, dass sorgfältig entworfene synthetische Daten realen Aufnahmen ebenbürtig sein und sie in entscheidenden Dimensionen sogar übertreffen können. Durch Kontrollierbarkeit, demografische Vielfalt und eingebaute Datenschutzgarantien etablieren sich synthetische Daten als zentrales Werkzeug für den Aufbau von Systemen zur Aktivitätserkennung, die robust, ethisch vertretbar und praktisch in realen Umgebungen einsetzbar sind.

ACKNOWLEDGEMENTS

I feel very fortunate to have carried out my PhD in computer vision at a time when the entire field is evolving at a breathtaking pace. At the same time, it is easy to see how such rapid progress can be perceived as a burden, with new developments appearing almost continuously. All the more, I am grateful for the people who created an environment in which this fast pace felt like an opportunity rather than a source of strain!

First and foremost, I would like to express my gratitude to my supervisor, Prof. Rainer Stiefelhagen. Whenever advice was needed, Rainer was available to share his perspective, yet he always gave me the freedom to explore my own research. During the entire time as a PhD student I felt that I could come to him with any idea or issue and he would support me and provide options on how to pursue or solve it. We have a great environment and team in our lab credit to Rainer and I believe that cannot be taken for granted.

I am sincerely thankful to my doctoral committee, to whom I will have the opportunity to present and discuss my work. In particular, I am honored by and grateful to Prof. Abhinav Valada for offering his time and effort to be my second advisor. I am very much looking forward to his and the committee's feedback. I am also thankful to Prof. Bruno and Prof. Bellosa for inviting me to the Professorengespräch and for the interesting discussions there. Furthermore, I would like to thank Ms. Ribet and Ms. Bohlinger from the Dekanat for helping me in the dissertation process and providing great support for all questions I had. My time at the CV:HCI lab has been shaped by many colleagues and friends. I remember the evenings we spent building PhD hats for those who defended before us, which somehow always turned into much more elaborate projects than planned. I will not forget our Bad Herrenalb retreats with games, and most importantly, table soccer, and the bouldering sessions with Alex and Constantin (and once upon a time we even convinced Simon to join) as well as countless shared breaks and discussions. I am also grateful to Kunyu, with whom I shared an office for a long time, and now to Di Wen, who has to tolerate my growing collection of caffeine beverage containers. I would like to thank Alina, who particularly in the early phase of my PhD was a supervisory figure in addition to Rainer and an important source of advice and perspective. I am grateful to all my co-authors and collaborators, many of whom are also my lab colleagues. In particular, I would like to thank Kunyu, Alexander, Zdravko, and Simon for our joint work and many discussions, as well as Zeyun for our joint collaborations in JuBot and on papers. I also would like to thank Vivek for joint collaborations and for giving me the opportunity to explore computer vision research as a PhD intern at Sony in Zürich. Beyond these names, I am thankful to all past and present colleagues and friends I met, Yufan, Ruiping, Omar, Saquib, Jiale, Di, Junwei, and Angela, and of course Corinna who holds the lab together.

I am also deeply thankful to Saskia and to Benjamin and Lucie for their support during this time. Finally, I want to thank my parents, Inge and Helmut, who have always supported me unconditionally and never gave me a reason to doubt that they would stand behind me, no matter which path I chose.

CONTENTS

I	BACKGROUND	1
1	INTRODUCTION	3
1.1	Challenges in Human Activity Recognition Datasets	3
1.1.1	Data Scarcity and Scale	3
1.1.2	Selection Biases and Dataset Limitations	5
1.1.3	Ethical and Privacy Constraints	5
1.2	Synthetic Data for Human Activity Recognition	6
1.2.1	Why Synthetic Data?	6
1.3	Thesis Outline and Contributions	9
2	RELATED WORK	13
2.1	Deep Learning for Human Activity Recognition	13
2.1.1	Vision-based Action Recognition Architectures	13
2.1.2	Domain Adaptation and Transfer Learning	14
2.2	Human Activity Recognition Datasets	15
2.2.1	General HAR Datasets	15
2.2.2	HAR Datasets for Activities of Daily Living	15
2.2.3	Fall Detection Datasets	15
2.3	Simulation-Based Data Synthesis for HAR	16
2.3.1	General Rendering-Based Solutions	16
2.3.2	SMPL-Based Data Synthesis	18
2.3.3	Generative Models for Data Synthesis	19
2.4	Synthetic-to-Real Transfer Challenges	20
2.5	Synthetic Data for Biological Motion Analysis	21
2.5.1	Computational Biomechanics and OpenSim Integration	21
2.5.2	EMG and Muscle Activation Datasets	22
2.6	Privacy-Preserving Methods	23
II	LEARNING FROM SYNTHETIC DATA	25
3	THE SYNTHETIC-TO-REAL DOMAIN GAP	27
3.1	Learning from Life Simulation Video Games	28
3.1.1	Sims4Action	29
3.1.2	Neural Architectures	33
3.1.3	Experiments	34
3.2	Domain Adaptive Knowledge Distillation	36
3.2.1	Problem Formulation	37
3.2.2	Optimal Transport for Domain Adaptation	38
3.2.3	Experiments	39
3.3	Multimodal Domain Generation for Domain-Transfer	42
3.3.1	Motivation	43
3.3.2	Multimodal Framework for Activity Domain Generation	46

3.3.3	Experimental Results	51
3.4	Chapter Conclusion	56
3.4.1	Contributions to Cross-Domain Synthetic Data Transfer	56
4	SYNTHETIC DATA FROM 3D HUMAN BODY MODELS	57
4.1	Motion Capture-Based Synthetic Data Generation	58
4.1.1	Implementation	59
4.1.2	Experiments	62
4.1.3	Synthetic data in Downstream Datasets	65
4.2	Biomechanical Synthetic Data Generation	67
4.2.1	The Muscles In Time dataset	68
4.2.2	Motion to Muscle Activation Estimation Benchmark	74
4.2.3	Experiments	74
4.3	Chapter Conclusion	76
4.3.1	Contributions to SMPL-based Synthetic Data Generation	77
5	SYNTHETIC DATA FOR REAL-WORLD PROBLEMS	79
5.1	OmniFall	80
5.1.1	Dataset	82
5.1.2	Experiments	89
5.2	Avatar-Based Anonymization with Privacy Guarantees	94
5.2.1	Preliminaries	96
5.2.2	Masked Differentially Private Training	98
5.2.3	Anonymization	100
5.2.4	Experiments	102
5.2.5	Results	103
5.2.6	Limitations	106
5.3	Chapter Conclusion	107
5.3.1	Contributions to Real World Problems	108
III	INSIGHTS	109
6	IMPACT ON THE FIELD	111
6.1	Research Directions	111
6.2	New Datasets	112
6.3	Methods, Frameworks, and Evaluations	113
7	OPEN RESEARCH QUESTIONS AND FUTURE WORK	115
7.1	Extending Core Contributions	115
7.1.1	Multi-Modal Learning from Shared Motion Representations	115
7.2	Towards General Synthetic Data Superiority	116
7.2.1	Generalization Beyond Specialized Tasks	116
7.2.2	Long-Term Coherent Video Generation	116
7.3	Understanding and Validating Synthetic Data	116
7.3.1	Crossing the Synthetic-to-Real Domain Gap for Biomechanical Signals	116
7.3.2	Generative Model Biases: Opportunity and Challenge	117
7.3.3	Quality Assessment and Failure Modes for Generative Models	117
7.4	Differentially Private Anonymization	117
7.5	Active and Adaptive Synthetic Data Generation	118
7.6	Concluding Synthesis	118

IV	APPENDIX	119
A	APPENDIX	121
A.1	Deferred Theoretical Proofs	121
A.1.1	Proof of Theorem 5.4	121
A.2	OmniFall Synthetic Data Generation Methodology	122
A.2.1	Demographic Attribute Design	123
A.2.2	Environmental and Contextual Diversity	123
A.2.3	Action-Specific Attributes	124
A.2.4	Camera Specification	124
A.2.5	Prompt Construction and Generation	124
A.2.6	Safety Filtering and Iterative Refinement	125
A.3	OmniFall Dense Annotation Process	125
A.3.1	Annotation Software Selection and Customization	125
A.3.2	Annotation Team and Training	126
A.3.3	Multi-View Annotation Strategy	126
A.3.4	Quality Assurance Protocol	126
A.4	MinT Dataset Simulation Quality Analysis	126
A.4.1	Quality Metrics and Thresholds	126
A.4.2	Tracking Error Results	127
A.4.3	Coordinate-Specific Errors	127
A.4.4	Knee Loading Metrics	128
A.4.5	Summary Dashboard	129
A.4.6	Summary Statistics	129
	Curriculum Vitae	131
B	AUTHORED PUBLICATIONS	133
C	BIBLIOGRAPHY	137

LIST OF FIGURES

Figure 1	Evolution of video action recognition architectures and datasets from 2011 to 2025. CNN-based methods are marked in blue while attention based methods are marked in green. Datasets for Activities of Daily Living are marked in orange, general datasets are unmarked.	14
Figure 2	Timeline of synthetic datasets for human action recognition (2016-2025). Approaches are categorized as game-based (red), simulator-based (blue), and generative (green) based on their underlying generation methodology.	17
Figure 3	This work leverages the life simulation video game <i>THE SIMS 4</i> for inexpensive ADL training data which is then utilized in the <i>GAMING</i> → <i>REAL</i> setting to suit real-world applications, such as smart homes and assistive robots.	28
Figure 4	Snapshots of the virtual apartments we designed to simulate household situations. The two homes cover a dining room, a living room and a kitchen. <i>C1</i> - <i>C24</i> marks the 24 camera placements used in the static camera mode recordings.	29
Figure 5	Eight subjects (four elderly people, two adults and two young adults) are covered in the <i>SIMS4ACTION</i> dataset.	30
Figure 6	Examples of subjects being engaged into different domestic activities in the <i>SIMS4ACTION</i> dataset and the real-life counterparts used for the <i>GAMING</i> → <i>REAL</i> Benchmark, which are derived from Toyota Smarthome [52].	31
Figure 7	Sample frequency of the activities in <i>train</i> , <i>test</i> and <i>val</i> . In contrast to real-life data collection [52], life simulation video games give us control of the data distribution resulting in a well-balanced dataset. (A sample corresponds to a 3s snippet with the assigned label.)	32
Figure 8	<i>Where did the activities take place?</i> The cord diagram visualizes the correspondences of the captured behaviours and the rooms it took place in. Best viewed in color.	32
Figure 9	Overview of our domain adaptation method. We evaluate the batch distribution of label predictions and use the Sinkhorn-Knopp algorithm to calculate target predictions which conform to a desired distribution.	37
Figure 10	Comparison of direct transfer (top) with unsupervised distribution training assuming the source label distribution (uniform), or using the target cross-subject train label distribution.	40

Figure 11	An overview of the proposed SYNTHETIC→REAL activity recognition pipeline improved through neural network-based generation of novel activity domains. In addition to synthetic RGB videos, the training data is enriched with modalities explicitly extracted from videos (<i>source</i> modalities) and action representations from the <i>novel</i> modalities generated with our model. Such multimodal diversification of the training samples significantly mitigates adverse effects of the SYNTHETIC→REAL domain shift.	43
Figure 12	Architecture of our multimodal framework for generation of novel ADL modalities. First, multiple modalities related to the body pose and movement are extracted and used alongside the original RGB videos as our source modalities X_k . Activity examples in these source modalities are then fed into the <i>Domain Generator (DG)</i> as single images in order to generate novel modalities. The distribution divergence between each source-novel domain pair is estimated via the <i>Novelty Loss</i> , whereas the <i>Diversity Loss</i> computes the divergence for each novel-novel domain pair. The source domains and the generated novel domains are both used as inputs to the <i>Action Classifier</i> , which makes separate predictions for both. The same generator is used to reconstruct the original input and the <i>Reconstruction Loss</i> is computed. The error signals for training our model via backpropagation are depicted as red arrows.	45
Figure 13	Examples of the <i>source</i> X_k (top row) and <i>novel</i> modalities $\tilde{X}_k$ (bottom row). H: Heatmaps, L: Limbs, OF: Optical Flow. We use all 8 modalities $X_k \cup \tilde{X}_k$ for training and evaluate our models on REAL data.	46
Figure 14	Domain Classifier (DC): Overview of inference and training. The DC model is pre-trained end-to-end on Sims4Action.	47
Figure 15	Domain Classifier: Computation of the <i>Novelty Loss</i> for all source-novel modality pairs $(X_k, \tilde{X}_k)$. Note that the <i>Diversity Loss</i> is computed the same way with all novel-novel pairs $(\tilde{X}_k, \tilde{X}_l)_{k \neq l}$	48
Figure 16	Example of AC training with three modalities ($k = 2$). The input sequences are concatenated along the channel dimension C . The action classifier AC is trained on both the source X_k and novel $\tilde{X}_k$ modalities for action recognition and makes separate predictions for both. Note that the input to the AC model is sequence-based ($S = 16$) as opposed to the image-based DC and DG ($S = 1$). Its frozen copy AC_f is pre-trained on the source modalities from Sims4Action and is only used for the computation of $\mathcal{L}_{Class}$ in Equation 4. The error gradient from AC_f flows back to the novel modalities as its prediction is used to train DG with $\mathcal{L}_{Class}$ (Eq. 4).	49
Figure 17	t-SNE visualization of the domain generator’s bottleneck layer embeddings. Each modality (both source and novel) forms distinct clusters, confirming that the generator has learned to create diverse, non-overlapping modality distributions.	51

Figure 18	The four camera views of SynthAct (top) and corresponding depth maps (below).	59
Figure 19	Random camera positions.	59
Figure 20	Overview of our data generation pipeline using multiple publicly available data sources.	60
Figure 21	Gain of pre-training on AMARV over OMNIVORE baseline on NTU RGB+D [228] using depth only. We plot the gain in per-class Top-1 accuracy for the top ten and bottom ten classes. Pre-training on AMARV improves the Top-1 accuracy on 52 out of the 60 total classes.	63
Figure 22	Comparison between synthetically supplemented training with real-world-only experiments considering varying shares of real world data on Toyota Smarthome.	65
Figure 23	Simulation pipeline of the Muscles in Time dataset. The SMPL representation is extracted from videos, then, the SMPL represented motions are mapped to bio-mechanically validated human body models to simulate fine-grained muscle activation, connecting computer vision with biomechanical research. Bottom right: two activation sequences for exemplary muscles. Images from [58, 158]	67
Figure 24	The AMASS body model with specific indices mapped onto the OpenSim lower body model by Lai <i>et al.</i> [135] (middle) and model of the thoracolumbar region by Bruno <i>et al.</i> [23] (right). Best viewed by zooming in.	69
Figure 25	Prevalence of different motions in the MinT dataset.	70
Figure 26	Approximated weight and height distribution of the analysed subjects in the MinT dataset.	71
Figure 27	Left: Clustering of multiple activities within the BMLmovi dataset by muscle activation features. Right: Column-wise color coded histograms of areas under muscle activation curves for 402 muscle strands, sorted by histogram medians. Log-normalized color map.	72
Figure 28	Example lower body muscle activations (split in left and right muscle strands) for the actions <i>kick</i> and <i>jumping jacks</i> . It is clearly visible that the kick is performed with the left leg. During <i>jumping jacks</i> , <i>gluteus medius</i> and <i>rectus femoris</i> are activated alternately for both legs.	75
Figure 29	OmniFall comprises OmniFall-Staged, OF-Synthetic, and OF-In-the-Wild with cross-dataset compatible labels.	81
Figure 30	Our annotations on an example video of CMDFall [256].	82
Figure 31	Label distribution and single-view duration per dataset: eight OF-Sta constituents, OF-ItW, and OF-Syn (33h total).	82
Figure 32	Pairwise Fréchet Video Distances calculated on features from I3D, VideoMAE-K400 and SigLip2, visualised as heat-maps. Darker colors indicate higher similarity (smaller distances) between datasets.	84
Figure 33	2D projections, colour-coded by dataset and activity. Rows correspond to feature type, columns show T-SNE and H-NNE clustering	85

Figure 34	Class-wise <i>sensitivity</i> , <i>specificity</i> and <i>F1-score</i> on OF-In-the-Wild for the eight most frequent classes, cross-domain generalization. . . .	86
Figure 35	OF-Syn demographic splits; <i>sensitivity</i> , <i>specificity</i> , <i>F1</i>	91
Figure 36	In our experiments we apply MaskDP on anonymized datasets containing non-sensitive synthetic humans. Only real-world data is considered private and needs protection.	94
Figure 37	Overview of our video anonymization framework. The process includes: ① 3D pose estimation, ② Synthetic human rendering using SMPL-H, ③ Background inpainting to remove original persons, ④ Merging of synthetic avatar and inpainted background, and ⑤ Diffusion-based refinement. This pipeline effectively anonymizes individuals while preserving video quality and context.	95
Figure 38	Anonymization examples from <i>NTU RGB+D 60</i> (a), <i>Toyota Smarthome</i> (b) and <i>UCF-101</i> (c). Our anonymization method can be applied to difficult poses (a) and maintains important context information such as food on a table or a kitchen environment (b).	100
Figure 39	Examples for the data used in each row of Table 24.	103
Figure 40	Class-wise accuracy loss on Toyota Smarthome compared to real world non-private training when applying DP (upper) and regain of accuracy upon applying MaskDP (lower) on anonymized data for $\epsilon = 5$. Class distribution marked in dark grey bars within. . . .	106
Figure 41	Distribution of maximum tracking errors across 3,947 simulated sequences. (A) Rotational tracking errors show median 0.40° with 95th percentile at 0.59° . All sequences remain below the 5° threshold. (B) Translational tracking errors show median 1.20 cm with 95th percentile 4.87 cm. 37.4% of sequences (1,476 sequences) exceed the 2 cm threshold.	127
Figure 42	The 20 rotational coordinates (joint angles) with highest median tracking error. The metatarsophalangeal (MTP) joint angles exhibit the highest errors. All values remain well below the 5° threshold (dashed line).	128
Figure 43	Knee loading metric distributions across all sequences. (A) Peak knee adduction moment (KAM) distribution shows median 0.023 BW·ht with 95th percentile 0.043 BW·ht. The shaded region (0.02–0.04 BW·ht) indicates the range reported for normal walking [133]. 0.2% of sequences exceed the 0.08 BW·ht threshold. (B) Peak medial contact force (MCF) distribution shows median 2.69 BW with 95th percentile 6.22 BW. 20.6% of sequences exceed the 4.0 BW threshold.	129
Figure 44	Summary of key quality metrics across 3,947 sequences. (A) Maximum rotational tracking error distribution (median 0.40° , threshold 5° shown as dashed red line). (B) Peak knee adduction moment distribution (median 0.023 BW·ht, threshold 0.08 BW·ht). (C) Peak medial contact force distribution (median 2.69 BW, threshold 4.0 BW). Black dashed lines indicate medians.	130

LIST OF TABLES

Table 1	Overview of synthetic action recognition datasets, ordered chronologically.	17
Table 2	Comparison between recent muscle activation datasets and <i>Muscles in Time</i>	22
Table 3	Statistics of the Sims4Action dataset. Samples represent 3-second video segments with activity labels.	33
Table 4	Per-class performance of I3D on the Sims4Action dataset. Precision, recall, per activity class.	33
Table 5	Evaluation of the ADL recognition in the GAMING→GAMING setting of our benchmark. The performance metrics are standard- and balanced accuracies.	34
Table 6	Recognition results in the GAMING→REAL setting. Such transfer is a hard task for current recognition models, although all methods surpass the random baseline.	34
Table 7	Domain generalization baseline results. Pre-training with pose-based representations improves transfer from synthetic to real domains.	41
Table 8	Domain adaptation results from Sims4Action to Toyota Smarthome.	42
Table 9	Comparison of our model to state-of-the-art results on the SYNTHETIC→REAL benchmark [207] and to three additional domain generalization methods - TA ³ N [35], APN [293], and VideoDG [293].	52
Table 10	Evaluation results for all 15 modality combinations on the SYNTHETIC→REAL benchmark for the real Toyota Smarthome [52] and ETRI [112] datasets and on the SYNTHETIC→SYNTHETIC benchmark for Sims4Action [207]. Our method of generating novel modalities $\tilde{X}_k$ for the training data leads to a significant improvement of the vast majority of the models on the SYNTHETIC→REAL benchmark. The performance on Sims4Action is listed to illustrate the domain gap. The performance boost by adding novel modalities $\tilde{X}_k$ to the training data is indicated in brackets for each modality combination.	55
Table 11	Supervised training on AMARV.	61
Table 12	Synthetic data approaches fine-tuned on NTU RGB+D.	61
Table 13	Cross-view-subject evaluation on NTU RGB+D, comparison with SURREACT.	62
Table 14	NTU RGB+D [228] and AMARV using depth only. Results demonstrate that the models pre-trained on AMARV are more generalizable and show superior fine-tuning performance on NTU.	64

Table 15	Toyota Smarthome comparison on cross-subject and cross-view 1 splits.	66
Table 16	Similarity of synthetic datasets (including our proposed AMARV dataset) and three real datasets with OMNIVORE features.	66
Table 17	Human motion-to-muscle activation prediction results for the lower- and upper body model.	73
Table 18	Datasets in the OmniFall benchmark. The first eight rows constitute OmniFall-Staged (OF-Sta), followed by OmniFall-In-the-Wild (OF-ItW) and OmniFall-Synthetic (OF-Syn).	83
Table 19	VMAE-K400 [253] trained on CMDFall CS and CV.	87
Table 20	VMAE-K400 [253] trained on OF-Staged (merge of all staged datasets) CS and CV and tested on those splits respectively.	88
Table 21	VMAE-K400 [253] trained on OF-Syn. Bal. Accuracy, Sensitivity, Specificity.	89
Table 22	VMAE-K400 [253] trained on OF-Sta(CS)+OF-Syn(Ra). Bal. Accuracy, Sensitivity, Specificity.	90
Table 23	ASQuery [82] timeline segmentation results.	92
Table 24	Performances on NTU RGB+D, Toyota Smarthome (TS) and UCF-101 without differentially private training.	104
Table 25	Accuracy for NTU RGB+D and mean Per Class Accuracy for Toyota Smarthome contrasting the model’s performance with synthetic and anonymized data under Differential Privacy (DP) and Masked Differential Privacy (MaskDP) across various epsilon values.	105
Table 26	Accuracy on UCF-101 contrasting the model’s performance with synthetic and anonymized data under Differential Privacy (DP) and Masked Differential Privacy (MaskDP) across various epsilon values.	106
Table 27	Summary statistics for MinT dataset quality metrics across 3,947 sequences.	129

ACRONYMS

OT-DAKD	Optimal Transport-based Domain Adaptive Knowledge Distillation
MMGen	Multimodal Domain Generation
AMARV	Archive Of Motion capture As Rendered Videos
MinT	Muscles in Time
OF-ItW	OmniFall-In-the-Wild
OF-Sta	OmniFall-Staged
OF-Syn	OmniFall-Synthetic
MaskDP	Masked Differential Privacy
ADL	Activities of Daily Living
CNN	Convolutional Neural Network
DANN	Domain-Adversarial Neural Network
DP	Differential Privacy
DP-SGD	Differentially Private Stochastic Gradient Descent
EMG	Electromyography
sEMG	Surface Electromyography
GAN	Generative Adversarial Network
HAR	Human Activity Recognition
LSTM	Long Short-Term Memory
mPCA	Mean Per-Class Accuracy
OOD	Out-of-Distribution
RNN	Recurrent Neural Network
SMPL	Skinned Multi-Person Linear model
SMPL-X	SMPL eXpressive
VAE	Variational Autoencoder
ViT	Vision Transformer

Part I

BACKGROUND

INTRODUCTION

This thesis investigates whether synthetic data can match or exceed real data for human activity recognition. For this, we take on challenges in cross-domain transfer as well as in controllable data generation. We pursue three main directions: First, we present benchmarks which quantify the synthetic-to-real domain gap and develop domain adaptation and generation methods which improve cross-domain robustness. Second, we use parametric body models to enable controllable large-scale data generation through rendering-based approaches and couple them with biomechanical simulation to access physiological signals cameras cannot capture. This opens research directions in effort estimation as well as rehabilitation monitoring. Third, we demonstrate that synthetic data with demographic diversity outperforms real data for human fall detection on genuine accidents. This shows that synthetic data can exceed real data for specialized recognition tasks.

In Plato's *allegory of the cave*, prisoners mistake shadows on a wall for reality itself. Modern machine learning systems face a similar challenge when learning human behavior. They observe the world through training data such as video recordings of humans performing activities and construct their entire understanding from these limited observations. Each video is a projection: a single viewpoint of a complex three-dimensional world, captured by a sensor at one moment in time.

Consider what these recordings capture. A camera in a laboratory films volunteers walking, sitting, cooking, or exercising under controlled conditions. Security cameras in public spaces record daily commutes and shopping trips. Smartphones capture social interactions and recreational activities. These recordings comprise the experiences from which deep learning systems learn human motion and behavior. Yet like the shadows in Plato's cave, they represent only a fraction of human experience, filtered through what is publicly observable and ethically permissible to collect. The following sections examine how these shadows, the data available for learning human activity recognition, are constrained by scarcity, systematic biases, and ethical boundaries.

1.1 CHALLENGES IN HUMAN ACTIVITY RECOGNITION DATASETS

1.1.1 Data Scarcity and Scale

Deep learning progress was driven forward by large and high-quality datasets. Apart from being caused by algorithmic improvements, deep learning breakthroughs such as *AlexNet* [130] were just as much enabled by datasets such as *ImageNet* [59]. Later computer vision datasets such as *COCO* [146] for object detection and *Kinetics-400* [123] for activity

recognition have likewise transformed research trajectories. They implement evaluation standards and incite new methodological directions. Thereby, these contributions show how datasets drive progress in artificial intelligence development. Beyond serving as benchmarks, such datasets also influence research agendas: Since they define which problems are considered tractable, publicly available datasets induce model development and architectural innovations. Recent discussions around data-centric AI explicitly recognize that the scale and quality of datasets are just as important as model design [165, 185].

Dataset acquisition for human activity recognition is especially challenging. Short internet clips frequently suffice for large-scale representation learning. At the same time, applications in health monitoring, elderly care, or robotic assistance require recordings of activities of daily living in authentic contexts. Such data is difficult to acquire: collection in private homes or care facilities is subject to privacy regulations; recruiting participants from diverse age groups and cultural backgrounds is logistically demanding.

Fine-grained annotation requirements further complicate the situation: Large video benchmarks rely on clip-level labels, whereas activity datasets for assistance or clinical use require temporally precise annotation and additional detail about the scene or interaction partners. More detailed in-depth activity analysis, for example, studies that include inertial sensors or EMG muscle activity [257], requires specialized equipment and expert labeling. Such datasets therefore stay small and are costly to scale. In fact, large diverse, and richly annotated datasets are rare over broad RGB-only datasets [200, 234].

Large-scale activity datasets such as *Kinetics-400/AVA* [93, 123] or *Moments in Time* [171] have become standard benchmarks for training and pretraining deep models in computer vision. Built primarily from internet video, they contain hundreds of thousands of clips across hundreds of activity categories. Despite their impact, these datasets do not address major challenges of human activity recognition in *sensitive domains*: categories are skewed toward visually distinctive and entertainment-oriented activities, whereas *activities of daily living (ADLs)* or rare yet sensitive events such as human falls are underrepresented. They also rarely contain modalities beyond RGB and lack the annotation detail needed for healthcare and assistive applications. In other words, although important for benchmarking general-purpose architectures, they remain insufficient for capturing the diversity and detail required in *human activity analysis for everyday assistance* situations.

The field is facing systematic data scarcity. While image recognition data has expanded through web scraping, we find that everyday activities in homes as well as activities in hospitals, and care facilities remain scarce, with privacy restricting their use. Dataset refinement approaches can alleviate quality issues [69, 272], but such methods can not compensate the lack of recordings from sensitive or infrequent settings. In similar manner to natural language processing, where models approach the limits of available high-quality training data [269], activity recognition is confronted with finite motion datasets limited by ethical and legal concerns as well as distributional boundaries [185]. As a result, the activities most needed for high-impact applications, such as falls, mobility impairments, or rehabilitation exercises, are precisely those least available. Consequently, this scarcity motivates exploring synthetic data to expand available training distributions [117, 178].

1.1.2 *Selection Biases and Dataset Limitations*

Besides data scarcity, dataset limitations extend to systematic biases that shape the set of capturable activities and their representation. Many datasets overrepresent young, healthy adults in public or semi-public environments. Meanwhile, vulnerable populations such as older adults or people with disabilities are less frequently included [3, 69, 119, 168]. One example for this is NTU RGB+D [228], a widely used multimodal human activity recognition (HAR) dataset. It covers 60 actions but consists almost entirely of young volunteers recorded in a controlled lab. Likewise, the Oops! dataset [68] focuses on accidental actions and yet it lacks vulnerable populations as well as settings in clinical environments. Domain-specific datasets for daily living activities, such as Toyota Smarthome [52], remain between one to two orders of magnitude smaller than web-scale datasets.

Large-scale web-collected datasets contain similar biases: visually distinctive activities shared online are overrepresented, whereas routine ADLs as well as caregiving tasks and interactions in domestic settings are relatively rare [123, 171]. The private area of human life remains mostly hidden, e.g. how elderly individuals navigate their homes and adapt movements to impairment, or likewise how families interact in intimate settings. These limitations reflect structural factors including ethical and privacy regulations as well as human dignity considerations that restrict what can be captured at scale [170].

Even where recordings exist, they often capture appearance, only. A video may show someone standing up from a chair, but not whether they require support, e.g. since they feel dizzy or adapt their movement to manage pain. Large-scale benchmarks focus on typical and successful task executions, while edge cases including near-failures or impaired performances remain underrepresented. Such cases are exactly where recognition systems could provide benefit: by identifying the need for assistance e.g. due to unsafe conditions or early mobility decline. Synthetic data can generate such underrepresented scenarios in a controlled, reviewable manner.

1.1.3 *Ethical and Privacy Constraints*

Privacy and ethical considerations additionally limit large-scale data collection of sensitive activities. Datasets such as EPIC-Kitchens [48] and Charades [236] cover household environments, and yet these remain small in size and are not representative of vulnerable populations such as the elderly or individuals with health conditions. In sensitive settings such as hospitals and care facilities, data collection is challenging: privacy regulations and informed consent requirements restrict what may be captured. Such constraints apply in order to avoid conflicting with ethical norms designed to protect human dignity [170].

Note, that many activities relevant to human activity recognition cannot be recorded systematically: falls in frail populations, medical emergencies, intimate or stigmatized behaviors, and scenes involving minors. Their collection is ethically constrained, requiring consent- and dignity-related safeguards, raising liability concerns, and is in addition limited by the rarity of such events. By modeling such scenarios computationally rather than filming real individuals, synthetic data enables research in domains where conventional collection is ethically or legally infeasible while respecting the ethical and legal requirements that rule out direct observation.

1.2 SYNTHETIC DATA FOR HUMAN ACTIVITY RECOGNITION

Recently, synthetic data became an increasingly important concept in computer vision and human activity recognition [117, 176, 178]. In this context, it refers to data generated by computational processes instead of being captured by physical sensors. We differentiate two major approaches: simulation-based pipelines and generative model-based methods.

Simulation-based pipelines make use of physics-based models and rendering engines to create human motion with fitting sensor outputs such as RGB, depth, or pose data [267]. Datasets such as SURREAL [267] and ElderSim [109] simulate activities in virtual environments, thereby providing control over scene setup and annotations. At the same time, generative model-based methods build on improvements of GANs, diffusion models, and neural rendering. This allows to generate motions from learned distributions or textual descriptions [106, 107, 249]. Although both approaches increase options for dataset design, they differ in what they offer: simulation allows for explicit control, while generative models provide larger flexibility regarding realism and diversity.

Augmentation techniques alter existing samples through geometric or photometric transformations. Still, such methods cannot introduce new behaviors or contexts. Synthetic data approaches, by contrast, allow the generation of new subjects and activities within new environments. This includes setups which are rare or ethically infeasible to record. In this sense, synthetic data complements real-world datasets: it extends the experimental setup and opens domains which remain out of reach for conventional collection.

1.2.1 *Why Synthetic Data?*

In the following, we discuss several potential advantages of synthetic data for HAR, covering both simulation-based and generative-model-based approaches.

COST EFFECTIVENESS AND SCALABILITY With simulation pipelines, large datasets can be generated at negligible marginal cost. Labels such as skeletons, action categories, or segmentation masks are generated automatically. Human annotation, frequently the dominant expense in large-scale video datasets [93], is therefore not needed. This cost saving matters most for temporal data, as frame-accurate labeling of action boundaries is slow and error-prone. Still, each target domain requires high entry investments: realistic 3D assets, detailed environments, and physics-based animation controllers are costly to design. Besides, unrealistic simulations affect performance once deployed in the real world.

Generative models change the cost structure of synthetic data generation. Training large-scale GANs or diffusion models requires both, large compute budgets and diverse training datasets. Still, once trained, these models can generate very large amounts of new data at low marginal cost: Text-to-motion models can generate thousands of variants of a rehabilitation exercise simply by varying textual prompts or latent seeds. This ability to scale makes them well suited to continual data generation, although dependence on real training datasets means that they do not remove data collection costs entirely.

CONTROLLING FACTORS OF VARIATION In simulation environments, the factors of variation are explicit and scriptable: researchers can systematically vary avatar body morphology and environment setups. Additional controllable factors include lighting as well as

camera geometry and object presence. Compositional datasets become possible this way: rare yet important scenarios can be generated easily, e. g. a fall while holding an object or sit-to-stand transitions with impaired balance. Such controlled sampling is not possible in real-world data collection, in which event frequencies follow Zipf-like distributions.

In contrast, the control offered by generative models is not deterministic since data diversity is introduced through conditioning mechanisms: Text-to-motion models e. g. MotionCLIP [249], allow us to specify an action such as “fall while carrying groceries” and to sample diverse stylistic variants. In order to alter generated sequences, diffusion pipelines can make use of pose constraints or other guidance signals. Note, that this is subject to model biases: prompts to create elderly avatars may default to stereotypes or fail, if such populations are underrepresented in the training data. Hence, both approaches support variation: simulation allows for explicit control of generation parameters, meanwhile generative models support larger stylistic and semantic diversity.

RELIABILITY AND ROBUSTNESS Robustness in HAR requires the appearance of disturbing factors in training. Examples for this are occlusion, clothing diversity, unusual camera angles or environmental noise. At the same time, simulation pipelines can incorporate such factors and create edge-case training data. For example, fall events can be displayed with stark occlusion or under uncommon sensor placements, e.g., top-down ceiling cameras in care facilities. Studies in robotics and autonomous driving show that such “domain randomization” methods boost generalization to unseen real-world settings [252].

Generative models add additional robustness by enlarging the range of realistic appearances and contexts. Diffusion-based video generation can alternate environmental conditions, e. g. falls in different scenes or weather conditions. This diversity helps models to become invariant to disturbing factors. Still, this applies only if synthetic realism is sufficient to overlap with the real distribution. Otherwise, artifacts may harm training rather than improve it. For this reason, robustness gains from generative models depend on ongoing progress in high quality human video synthesis.

LABEL QUALITY A clear advantage of simulation is its ability to provide temporally consistent, noise-free labels across multiple modalities. From a single simulation run, researchers can extract RGB, depth, and skeletal keypoints, but also latent biomechanical quantities, e. g. ground-reaction forces or joint torques, via physics-based models like OpenSim [58]. Note, that such labels are synchronized across frames and free from human annotator error, a feat not achievable through manual annotation alone. They are therefore attractive for tasks such as fine-grained rehabilitation monitoring or gait analysis.

Generative models do not inherently provide labels. Existing approaches impose labels using conditional inputs, e. g. “sitting down” as a text prompt or conditioning on a target pose. Still no guarantees on the semantics of the generated samples are given. Some approaches combine generative models with pretrained classifiers in order to validate or relabel generated outputs. However, this additional step often hampers label quality, leading to error-prone annotation pipelines. In other words, generative models offer flexible data, while simulation remains the gold standard, when exact labels are important.

PAIRED MULTIMODALITY Multimodal data is valuable for HAR, since RGB video alone can fail to capture internal states, e.g. to detect subtle motoric deficits. At the same time,

simulation environments can generate aligned paired modalities: RGB, depth, surface normals, optical flow, and IMU tracks can all be extracted simultaneously. Meanwhile, biomechanical simulators produce ground-truth muscle activations allowing for the calculation of forces and torques. Datasets such as AMASS [155] demonstrate the value of such multimodal biomechanics.

Generative models currently do not cover this area: Existing pipelines focus on visual modalities, despite works which increasingly explore audio-visual co-generation or other modalities. True multimodal diffusion pipelines for HAR, such as the generation of synchronized video containing inertial traces or EMG remain an open challenge. At present, co-aligned multi-modal data generation is mostly done by data simulation methods.

PRIVACY PRESERVATION AND ETHICAL DATA COLLECTION Privacy and ethics are major challenges in human activity recognition. In this field, data recordings may reveal identifiable appearance and behavioral patterns, and can even enable re-identification via gait. Synthetic data reduces these risks but does not eliminate them completely: Simulation pipelines reduce identifiability by replacing real subjects with avatars and synthetic environments. Nevertheless, if avatar motion is derived from motion capture of real individuals, gait or movement patterns may still enable re-identification. Generative models, in turn, produce novel samples rather than direct recordings. Still, these are subject to memorization and unintended reproduction of training data samples [29]. Mitigation strategies exist: hybrid pipelines can re-render real motion into synthetic avatars. This allows for obscuring appearance and context, while differential privacy can reduce memorization risks in generative models [1], albeit often at the expense of fidelity.

Beyond privacy, synthetic data enables research in domains in which conventional data collection is ethically or legally infeasible. Many intimate activities relevant to HAR as well as medical emergencies or scenes involving minors, cannot be recorded due to consent and dignity constraints or rarity. Hence, when modelling such scenarios without filming real individuals, synthetic pipelines enable the design of balanced and reviewable datasets, e.g. for safety-critical applications. At the same time, ethical and legal requirements that complicate real-world recordings remain intact.

ACTIVE DATA GENERATION AND KNOWLEDGE DISTILLATION A major promise of synthetic data lies in its potential for adaptive dataset design. Simulation environments allow iterative refinement: when models fail on a particular context (e.g., occluded sit-to-stand), new synthetic samples can be scripted to specifically target that weakness. This aligns with active learning, where data acquisition policies adapt to model uncertainty.

More recent generative models further facilitate the data generation process, allowing for synthesis of data samples based on textual representations or other constraints.

Techniques such as membership query synthesis and data-free knowledge distillation enable models themselves to request synthetic samples that reduce their uncertainty or mimic missing source data [10, 226]. Such techniques are established in the field of Natural Language Processing (NLP) and image classification; in HAR, however, their application remains limited, as actions are spatio-temporally complex and labels in long video sequences are difficult to validate. Still, they point toward HAR datasets which are dynamic and co-evolve with the models trained on them.

1.3 THESIS OUTLINE AND CONTRIBUTIONS

This dissertation aims to develop methods for generating and using synthetic data to advance video-based human activity recognition while addressing challenges in cross-domain transfer, controllable data generation, and privacy-preserving deployment. Specifically, we identify three research directions crucial for establishing synthetic data as a viable and, in certain contexts, superior alternative to real-world data collection. The main part of this thesis contains three chapters dedicated to these directions: We establish benchmarks and methods for synthetic-to-real transfer (Chapter 3), demonstrate controllable generation through parametric body models (Chapter 4), and solve real-world deployment problems where real data collection faces inherent limitations (Chapter 5).

The Synthetic-to-Real Domain Gap

How can we evaluate and bridge the domain gap between synthetic training data and real-world deployment? Synthetic data is rarely generated for synthetic-world deployment but rather to train models that perform in real-world settings. This inherent cross-domain nature requires both systematic benchmarks to quantify domain gaps and methods to bridge them, which Chapter 3 addresses through three complementary contributions. We begin by establishing Sims4Action (Section 3.1), the first cross-domain benchmark specifically designed for synthetic-to-real transfer in activities of daily living. While prior synthetic datasets existed, they were not developed with cross-domain evaluation as a primary design criterion. Sims4Action pairs gaming-based synthetic data from The Sims 4 with real-world evaluation on Toyota Smarthome, providing standardized protocols for quantifying domain gaps and evaluating cross-domain methods (published in *IROS 2021* [207]).

We continue by addressing the identified domain gaps through two complementary approaches. Section 3.2 presents Optimal Transport-based Domain Adaptive Knowledge Distillation (OT-DAKD), which improves results on the Sims4Action benchmark through pose-based feature alignment and knowledge distillation (published in *CVPRW 2022* [223]). Section 3.3 introduces MMGen, a framework that generates synthetic training domains to improve cross-domain robustness. Rather than adapting to specific real-world distributions, MMGen creates novel training modalities through generative adversarial networks that maximize distribution divergence while preserving semantic content. Evaluated on the Synthetic→Real ADL recognition benchmark, MMGen demonstrates that learned synthetic domain augmentation reduces cross-domain performance gaps. This positions synthetic domain generation as an alternative solution to cross-domain robustness (published in *IROS 2022* [162]).

Synthetic Data from Human Pose and Shape Models

How can parametric body models enable large-scale synthetic data generation, and does training on such data improve real-world performance? Can parametric body models combined with biomechanical simulation enable muscle activation datasets at scales sufficient for training deep learning models? SMPL parametric human body models enable controllable, scalable synthetic data generation through bidirectional mapping: neural networks estimate SMPL parameters

from video, while rendering engines reconstruct photorealistic synthetic humans from these parameters. Chapter 4 presents two applications of this approach.

Section 4.1 introduces SynthAct, a Unity-based rendering pipeline which generates large-scale multi-view synthetic data from motion capture sequences. By rendering SMPL bodies from AMASS motion sequences across four synchronized camera perspectives, the framework produces AMARV (Archive Of Motion capture As Rendered Videos). Experiments demonstrate that models pre-trained on AMARV transfer effectively to real-world action recognition benchmarks. This makes parametric body model rendering a practical approach for large-scale synthetic data generation. At the same time it addresses the privacy concerns of web-scraped video datasets (published in *ICRA 2024* [219]).

Section 4.2 extends SMPL-based synthesis to biomechanical simulation and introduces Muscles in Time (MinT). SMPL’s consistent surface correspondence enables mapping 3D points across frames. This allows for the transfer of motion data to OpenSim musculoskeletal models for biomechanical simulation. MinT generates 10 hours of 402-muscle activation data across 227 subjects, a scale unreachable with real-world EMG measurement. Experiments on pose-to-muscle activation prediction demonstrate that deep learning models can effectively learn these mappings from synthetic biomechanical data, which shows that parametric models combined with biomechanical simulation yield training data at the scale needed to train deep models from scratch on physiological tasks (published in *NeurIPS 2024* [221]). This biomechanical enrichment opens research directions in effort estimation and injury risk assessment as well as rehabilitation monitoring from vision, which were so far constrained by the scarcity of physiological measurement data.

Synthetic Data for Real-World Problems

Can synthetic data with systematic variation outperform real data for out-of-distribution generalization, and what does this reveal about the maturity of modern video generation foundation models? When does synthetic data enable real-world deployment capabilities that real data collection cannot provide? Chapter 5 explores two scenarios where real data collection faces fundamental barriers: performance limitations of staged laboratory data and privacy concerns in sensitive monitoring applications.

Section 5.1 addresses fall detection for elderly care, in which real staged datasets feature healthy actors in controlled environments with safety equipment. This creates unknown generalization gaps to genuine accidents. We present OmniFall, a unified fall detection benchmark which contains OF-Staged (eight public datasets with standardized annotations), OF-In-the-Wild (genuine accident videos for out-of-distribution evaluation), and OF-Synthetic (12,000 videos generated using Wan 2.2 with systematic variation). Our main finding: synthetic video with systematic variation in age, ethnicity, body type, and environment surpasses real staged datasets in out-of-distribution generalization. This demonstrates that generatively produced synthetic video can be preferable to real-world data and signals the maturity of modern video diffusion models. It points to a major shift in which synthetic video generation becomes a primary data collection method, chosen for performance advantages rather than only for cost savings. This shows, that synthetic data generation can surpass real data for training when real data exhibits limited variation in its distribution (published in *ECCV 2026* [220]).

Can we combine synthetic avatar anonymization with selective differential privacy to provide privacy for both the appearance of individuals and their environments and living spaces? Section 5.2 addresses privacy-preserving activity recognition in home monitoring, where real video data contains identifiable information about individuals and their living environments. We introduce MaskDP (Masked Differential Privacy), a hybrid privacy framework decomposing videos into synthetic avatar regions (providing visual anonymization through parametric body model rendering) and sensitive background regions (receiving formal differential privacy protection during training). By protecting only explicitly defined sensitive spatio-temporal areas rather than entire videos, MaskDP achieves (ϵ, δ) -differential privacy guarantees for background contexts while significantly improving recognition accuracy compared to standard differential privacy. While not providing the same strong guarantees as full differential privacy, this selective protection improves privacy significantly compared to pure anonymization while being less sensitive to the strong utility penalties incurred by full DP training, making hybrid privacy approaches practical for video-based activity recognition, particularly in strongly private low- ϵ regimes (presented at the *2nd International Workshop on Privacy-Preserving Computer Vision* [222]).

A NOTE ON IMPLEMENTATION Section 3.3, Section 4.1, and Section 4.2 include components resulting from close collaboration between David Schneider and the Master’s students Zdravko Marinov, Marco Keller, and Marco Kugler under his supervision. While precise boundaries are difficult to draw, implementation work for Section 3.3 was rather carried out by the student based on ideas and guidance from David Schneider. For Section 4.1 and Section 4.2, David Schneider contributed both the core conceptual development and major parts of the implementation, with the students substantially contributing to the implementation of the data generation setup. Almost all experiments in Section 4.1 and Section 4.2 were implemented, conducted, and evaluated by David Schneider. David Schneider is the sole main responsible of all remaining parts of this thesis.

A NOTE ON TABLE MARKINGS Where results warrant comparison, we mark best values in tables in bold and, optionally, second best values underlined throughout this thesis.

RELATED WORK

This thesis builds upon prior work in five areas: deep learning architectures, human activity datasets, synthetic data generation, biomechanical simulation, and privacy preservation. Foundational architectures and domain adaptation methods (Section 2.1) are the basis for the models we apply in this work. Human activity recognition datasets (Section 2.2) span general benchmarks, activities of daily living, and fall detection. They represent the evaluation settings which our contributions address. We distinguish between two paradigms of synthetic data generation: simulation-based methods, which use game engines and SMPL parametric models (Section 2.3), and generative model-based techniques, which build on GANs and diffusion models (Section 2.3.3). Biomechanical simulation (Section 2.5) bridges motion with physiological signals. Muscle activation analysis thus becomes possible at scales which experimental measurement cannot reach. Privacy-preserving techniques (Section 2.6), from visual anonymization to differential privacy, address the requirements of deployment in sensitive environments.

2.1 DEEP LEARNING FOR HUMAN ACTIVITY RECOGNITION

Vision-based deep learning has progressed from hand-crafted feature pipelines to end-to-end trainable models which jointly learn spatial and temporal representations from video. In the following, we study the evolution of primary architectural paradigms, from early convolutional networks to modern transformers and self-supervised approaches. We highlight models which have become standard backbones for action recognition and downstream tasks in this thesis (Fig. 1).

2.1.1 *Vision-based Action Recognition Architectures*

Two-Stream Convolutional Neural Networks (CNNs) [237] introduced the separation of appearance and motion in modern deep learning for action recognition, as they process spatial and temporal information through two separate streams, RGB and optical flow.

C3D [255] built on these early convolutional approaches and learned spatiotemporal features directly from video through 3D convolutions. Complexity and limited temporal receptive fields, however, constrained its application to shorter sequences. I3D [30] inflated ImageNet-pretrained 2D filters to 3D and made use of large-scale pretraining. This showed how important pretraining is for spatiotemporal understanding.

To further improve efficiency and temporal modeling, SlowFast [75] processed video through dual pathways at different temporal resolutions, which enabled multi-scale mod-

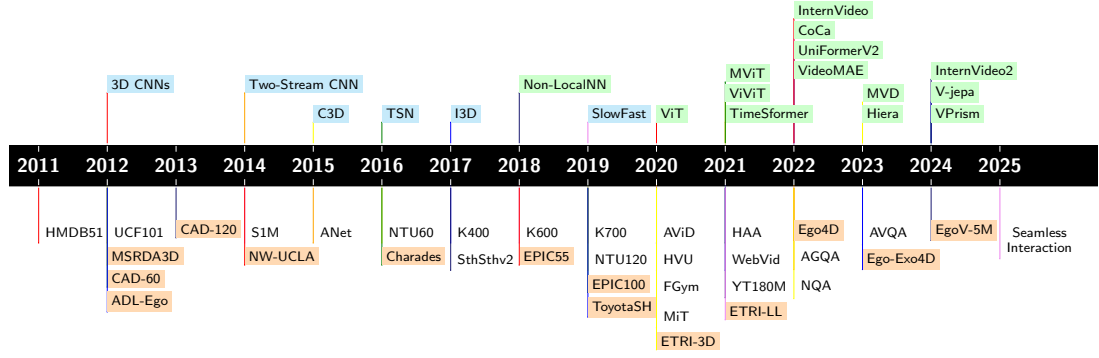


Figure 1: Evolution of video action recognition architectures and datasets from 2011 to 2025. CNN-based methods are marked in blue while attention based methods are marked in green. Datasets for Activities of Daily Living are marked in orange, general datasets are unmarked.

eling. TSM [145] enabled 2D CNNs to model temporal relationships at zero cost by shifting channels across time and achieved real-time mobile performance.

Vision Transformers moved beyond convolutional backbones and adapted attention mechanisms to video understanding. ViT [65] introduced self-attention for spatial modeling, which video architectures extended through various temporal factorization strategies: TimeSformer [15] separated temporal and spatial attention for longer sequences, ViViT [9] demonstrated multiple factorization approaches and the importance of image pretraining, while MViT [72] achieved improved efficiency through hierarchical feature pyramids.

In parallel, self-supervised pretraining has become central for scaling video models. VideoMAE [253] applied masked autoencoding for self-supervised video understanding and addressed data scarcity through reconstruction of masked spatiotemporal patches. InternVideo [278] unified multiple tasks through massive-scale pretraining with strong transfer capabilities, though deployment remains computationally challenging.

2.1.2 Domain Adaptation and Transfer Learning

Domain adaptation addresses distribution shift through techniques that learn domain-invariant representations or explicitly align source and target distributions. Classical approaches include DANN’s gradient reversal for domain-invariant features [83] and Deep CORAL’s second-order statistic alignment [245]. Domain adaptation approaches are explored for synthetic-to-real transfer in Sims4Action (Section 3.1), optimal transport knowledge distillation (Section 3.2), and MMGen (Section 3.3).

Knowledge distillation transfers representations across domains and modalities through soft output matching [104], hint-based intermediate features [210], and spatial attention transfer [296]. Together with optimal transport, which provides principled distribution alignment via the efficient entropic regularization of the Sinkhorn-Knopp algorithm [46], these techniques form the basis for the multimodal video and pose based cross-domain adaptation in Section 3.2. Damodaran *et al.* [49] explored similar ideas for domain adaptation in the image domain. This shows the broader applicability of optimal transport for distribution alignment.

2.2 HUMAN ACTIVITY RECOGNITION DATASETS

Human activity recognition datasets enable training and evaluation of deep learning models across different scales and domains.

2.2.1 General HAR Datasets

For brevity, we review general human activity recognition datasets which have shaped the field. Figure 1 displays additional important contributions in this area, yet without being exhaustive. Early controlled datasets such as KTH (600 videos, 6 actions) [225] and Weizmann (90 videos, 10 actions) [91] enabled initial development, yet lacked real-world diversity. UCF101 [242] (13,320 videos, 101 categories) and HMDB51 [131] (6,849 clips, 51 actions) moved towards YouTube and movie sources, which are more realistic, yet not representative of daily life. Kinetics [123] advanced the field with 300,000 videos across 400 classes (extended to 600 and 700 classes in later versions). This enabled large-scale pretraining despite web-sourced biases. Something-Something [92] (220,847 videos) emphasized fine-grained temporal reasoning with everyday objects, while AVA [93] (430 video clips, 1.6M action labels at 1 Hz) addressed spatiotemporal localization and multiple simultaneous actions.

2.2.2 HAR Datasets for Activities of Daily Living

Activities of Daily Living (ADL) datasets address assistive technologies and smart home applications with longer, complex activities in domestic environments.

Charades [236] contains 9,848 videos of 157 daily activities recorded in homes by crowd workers following scripts. The dataset features temporal annotations for multiple concurrent activities, which reflect the complexity of real household behaviors.

Taking a multi-modal approach, NTU RGB+D [228] provides 56,880 sequences of 60 action classes performed by 40 subjects, captured with Kinect sensors providing RGB, depth, and skeleton data. The dataset offers both cross-subject and cross-view evaluation.

Focusing more directly on assistive monitoring, Toyota Smarthome [52] provides 16,115 videos of 31 activities from 18 elderly subjects with multiple camera views which emphasize realistic behaviors, under cross-subject and cross-view protocols. ETRI-Activity3D [112] captures 112,620 RGB-D samples from 100 subjects across 55 activities from multiple viewpoints for human-robot interaction.

2.2.3 Fall Detection Datasets

Fall detection datasets [2, 11, 33, 94, 134, 163, 256, 302] represent a specialized subset of HAR datasets which consists of specialized activities with high relevance but low occurrence frequency. Fall events are difficult to capture in real-world settings due to ethical and safety as well as privacy concerns.

In Section 5.1 we introduce OmniFall as a new fall detection dataset with synthetically generated data. Since this work systematically collects and unifies existing fall detection datasets, the respective related work is detailed and deferred to Section 5.1.1.

Our Contribution: To address the generalization gap between real staged fall datasets (featuring healthy actors with safety equipment) and genuine accidents, we establish OmniFall, a unified benchmark comprising OF-Staged (eight public datasets with standardized annotations), OF-In-the-Wild (genuine accident evaluation), and OF-Synthetic (controlled demographic diversity). The critical finding: synthetic data with systematic diversity dominates real staged data for out-of-distribution generalization on genuine accidents. Thereby, synthetic video generation becomes a primary data collection method for specialized recognition tasks.

2.3 SIMULATION-BASED DATA SYNTHESIS FOR HAR

Scope: This section focuses specifically on synthetic data generation methods aimed at providing training data for human activity recognition. While substantial work exists on SMPL-based synthetic data generation and synthetic video generation for other purposes, such as training human pose and shape estimation models or advancing generation techniques themselves, we review only approaches that explicitly target human activity recognition as their primary application. The exception is video diffusion models (Section 2.3.3), which we include because OmniFall makes use of recent video diffusion models for activity data generation, directly relating these foundation models to activity recognition within this thesis.

Simulation-based synthetic data generation makes use of 3D rendering engines and physics-based modeling to create controlled training data for human action recognition. This paradigm addresses the dual challenges of data scarcity and privacy concerns and at the same time provides explicit parameter control over appearance, motion, viewpoint, and environmental factors. The approach has proven effective across computer vision domains such as object recognition [31, 188, 246], semantic segmentation [205, 206], and body pose estimation [85, 155, 183, 189, 195, 244, 267], with growing applications in human activity recognition [56, 198]. In Figure 2, we summarize simulation-based datasets and approaches from 2016-2025, organized by generation methodology.

Simulation-based approaches span two primary technical paradigms: general rendering-based solutions, which make use of game engines and specialized platforms [7, 109, 193], and SMPL-based parametric human modeling, which enables systematic control over body shape, pose, and appearance [219, 267]. The trade-offs are distinct: gaming and specialized platforms provide visual diversity and environment complexity, whereas SMPL-based methods offer precise control over human body parameters and multi-modal outputs (depth, segmentation, pose). How effective synthetic-to-real transfer is depends on the balance of motion authenticity, visual realism, and data diversity.

2.3.1 General Rendering-Based Solutions

In Table 1, we provide an overview of the publicly available datasets for synthetic-to-real activity recognition. These datasets effectively address their intended applications; systematic cross-domain evaluation, however, was not a central design objective. ElderSim [109],

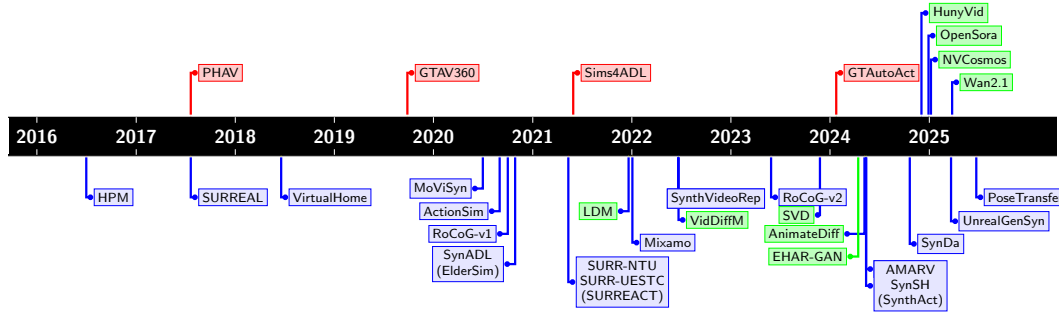


Figure 2: Timeline of synthetic datasets for human action recognition (2016-2025). Approaches are categorized as game-based (red), simulator-based (blue), and generative (green) based on their underlying generation methodology.

Table 1: Overview of synthetic action recognition datasets, ordered chronologically.

Dataset	Year	Acts.	Acq.	Source	Size	Mod.
PHAV [56]	2017	35	MC	CMU Mocap	55h	RGB
ActionSim [151]	2021	5	MC	Mocap	0.1k / 1h	RGB, 2DP
SynADL [109] ¹	2021	55	PE	EtriAct.3D	462k	RGB, 3DP
Sims4ADL [207]	2021	10	-	Ingame	1k / 10h	RGB
SURR. NTU [266]	2021	60	PE	NTU RGB+D	105k	RGB, OF, 2DP
SURR. UESTC [266]	2021	40	PE	UESTC	3k	RGB, OF, 2DP
Mixamo [44]	2022	14	PE	Kinetics	25k / 70h	RGB, 2DP
BEDLAM [18]	2023	-	MC	AMASS	10k	RGB, IS, D, 3DP
AMARV	2023	260	MC	AMASS	800k	RGB, IS, SN, D, 3DP
Synthetic SH	2023	31	PE	Toyota ADL	45k	RGB, IS, SN, D, 3DP

IS Instance segmentation OF Optical Flow D Depth Maps 2DP 2D-Poses 3DP 3D-Poses

SN Surface Normals

MC Motion Capture PE Pose Estimation

¹ While Eldersim is the dataset generator, SynADL is the dataset name.

for instance, covers elderly care scenarios with 55 activities. PHAV [56] and Mixamo [44] offer 35 and 14 action categories, respectively. Standardized protocols which quantify and compare synthetic-to-real transfer across these datasets remained absent. This motivated dedicated benchmarks with explicit cross-domain evaluation frameworks.

De Souza et al. [56] first addressed synthetic-to-real activity recognition with PHAV (Procedural Human Action Videos), which uses motion capture data and computer graphics techniques to generate single-view videos. The dataset comprises 39,000 videos across 35 actions. Viewpoint, lighting, and appearance are systematically controlled, whereas motion authenticity is maintained. Subsequent work extended synthetic generation to specialized application domains and technical paradigms.

Domain-specific platforms address targeted use cases. Puig et al. [193] introduced VirtualHome, which simulates household activities through symbolic program execution. Hwang et al. [109] developed ElderSim for eldercare with 55 daily activities and age-appropriate motion patterns as well as multimodal outputs (RGB, depth, skeletal data).

De Melo et al. [167] developed RoCoG for human-robot gesture recognition; their ablation studies reveal, that character diversity and motion variation impact transfer effectiveness more than environmental factors. Reddy et al. [201] extended this direction through RoCoG-v2 as well as enhanced synthetic-to-real evaluation protocols.

Accessible alternatives came from gaming environments, i. e. from engine extraction or modification. Ardianto and Hang [7] extracted omnidirectional videos from Grand Theft Auto V for 360-degree viewpoint coverage, whereas Song et al. [240] developed GTAu-toAct through systematic game engine redevelopment, which enables explicit control over character behavior, environment, and camera parameters.

Parallel efforts explored augmentation strategies and representation learning. Ludl et al. [151] demonstrated through ActionSim that simulation-based augmentation especially benefits difficult corner-case poses and motions, whereas Guo et al. [97] developed SynthVideoRep for representation learning pretraining. This shows that generalizable representations enable effective real-world transfer. Recent work has focused on end-to-end optimization: Rajendran et al. [199] integrated motion synthesis, rendering, and training in SynDa’s unified pipeline; Mulero-Pérez et al. [175] developed UnrealGenSyn for rare event generation; and Knapp and Bohacek [126] introduced PoseTransfer for motion pattern transfer across visual contexts.

Our Contribution: To evaluate synthetic-to-real transfer in activities of daily living, we present Sims4Action (*IROS 2021*), which pairs life simulation game-based synthetic data from *The Sims 4* with real-world evaluation on Toyota Smarthome. Sims4Action quantifies the domain gap and provides cross-domain evaluation protocols. Thus, synthetic→real transfer capabilities become measurable.

2.3.2 SMPL-Based Data Synthesis

Parametric human body models, especially SMPL (Skinned Multi-Person Linear model), enable systematic control over body shape, pose, and appearance. These models provide explicit parameter control and multi-modal outputs (RGB, depth, segmentation, pose) and at the same time maintain realistic human proportions and motion plausibility.

Rahmani and Mian [197] established synthetic data for viewpoint-invariant recognition by systematically rendering 3D human models across multiple viewpoints. Their demonstration that synthetic viewpoint variation could address real-world viewpoint challenges influenced subsequent parametric approaches, particularly SURREAL [267].

Varol et al. [267] developed SURREAL with SMPL-based parametric synthesis and generated 6.5 million frames with pose, depth, and segmentation annotations. The systematic variation of body shape, texture, and environment demonstrated parametric scalability and laid the basis for subsequent action recognition work.

Building on SURREAL [267], Varol et al. [266] developed SURREACT with SURR-NTU and SURR-UESTC datasets specifically for action recognition. The work demonstrated that parametric body models generate effective training data for real-world benchmarks when combined with domain adaptation techniques. Turrisi da Costa et al. [44] extended this direction through Mixamo-Kinetics, applying dual-head contrastive domain adaptation to bridge the gap between Mixamo animations and real video data.

Our *SynthAct* [219] integrated AMASS motion capture with scalable rendering and generated 800,000+ multi-view clips which preserve motion authenticity. This yielded cross-domain improvements on NTU RGB+D and Toyota Smarthome benchmarks.

In Section 4.1 one of our contributions is the AMARV synthetic video dataset which builds onto the AMASS motion capture dataset. The concurrently developed BEDLAM [18] bears similarities to AMARV, as both make use of the AMASS dataset. While our data generator *SynthAct* builds on Unity’s perception library [263], BEDLAM makes use of the Unreal Engine. Despite a number of technical and content generation related differences, the datasets serve different purposes and scale: BEDLAM targets 3D human pose and shape estimation and samples 2,311 motions, which result in 10,000 motion clips optimized for body model fitting tasks. AMARV, in contrast, aims to provide comprehensive training data for generalizable multi-view activity recognition. It covers more than 10,000 motion sequences labeled by BABEL [195] and adds additional modalities for action understanding tasks. Since AMARV also provides synchronized simultaneous recordings from four perspectives, this results in a clearly larger dataset of 800,000 motion clips.

Our Contribution: To enable large-scale multi-view synthetic data generation without the privacy concerns of web-scraped datasets, we introduce *SynthAct* (ICRA 2024), which renders AMASS motion sequences through parametric body models across synchronized camera perspectives. The resulting AMARV dataset demonstrates that parametric rendering transfers effectively to real-world benchmarks. It offers a privacy-preserving alternative to web-sourced video and at the same time enables multi-view learning without expensive multi-camera infrastructure.

2.3.3 Generative Models for Data Synthesis

Generative models have evolved from discriminative approaches to powerful synthesis methods capable of creating realistic training data at scale. Generative Adversarial Networks (GANs) [90] introduced the basic adversarial training approach, in which a generator and discriminator engage in a min-max game to produce samples indistinguishable from real data. Subsequent variants such as DCGAN [196], Wasserstein GANs [8, 95], BigGAN [21] and StyleGAN [122] progressively improved training stability, controllability and image fidelity. This framework spawned numerous architectural innovations over the following years. It extended from image generation to video synthesis through temporal modeling in early video GANs [213, 270].

Video GANs addressed the challenge of temporal coherence and modeled both, appearance and dynamics. Early approaches such as VGAN [270] and TGAN [213] modeled spatio-temporal structure with 3D convolutions or separate temporal generators. MoCoGAN [259] introduced explicit motion-content decomposition, using separate latent codes to control temporal dynamics and visual appearance, establishing a foundational architecture for controllable video synthesis, while MoCoGAN-HD [250] leveraged a strong pretrained image generator to scale GAN-based video synthesis to higher resolutions. These GAN-based approaches demonstrated the feasibility of synthetic video generation, yet faced challenges in temporal consistency, mode collapse, and scaling to complex scenes. For action recognition specifically, Pulakurthi et al. [194] introduced EHAR-GAN

for dataset augmentation. It preserves action semantics and introduces variations as supplementation rather than replacement.

Diffusion models emerged as a more stable alternative and first succeeded in image generation with denoising diffusion probabilistic models and their improved variants [105, 177], later with latent diffusion for efficient high-resolution, text-conditioned synthesis [209] before extending to video domains. Recent text-to-video diffusion models such as Imagen Video, VideoLDM, and Make-A-Video [19, 107, 238] have enabled large-scale synthetic data generation for action recognition. CogVideoX [292] makes use of a 3D causal variational autoencoder with expert transformers to generate high-quality videos from text prompts. AnimateDiff [98] extends text-to-image diffusion models with motion modules, which enables animation of characters, a capability we made use of in privacy-preserving approaches like MaskDP [222] for avatar refinement via video-to-video style transfer. Lumiere [14] introduced a space-time U-Net architecture that generates videos at full frame rates in a single pass. This advances beyond the cascaded generation approaches of earlier diffusion models. Wan [271] introduced a foundation model for video generation which provides high-quality videos covering diverse scenes and actions, a capability we used for the generation of OmniFall-Synthetic in Section 5.1. The quality and realism of generated videos remain challenging to evaluate. Fréchet Video Distance (FVD) [264] has become a standard metric for assessing generative video models. FVD extends the Inception-based FID metric to video by computing distances between feature distributions extracted from I3D video understanding models. This provides better alignment with human judgment of video quality than frame-based metrics. In OmniFall, FVD quantifies the domain gap between synthetic and real fall detection datasets.

While these generative approaches enable large-scale synthetic data creation, their practical utility depends on effective transfer to real-world tasks, further motivating research into synthetic-to-real domain adaptation and generalization.

2.4 SYNTHETIC-TO-REAL TRANSFER CHALLENGES

Synthetic-to-real transfer challenges manifest as appearance variations, motion timing differences, and environmental discrepancies [44, 167, 207]. From the methodological perspective, prior work addresses these challenges through two primary paradigms: (1) complementing real training data by mixing it with generated data, often posed as a domain adaptation problem [56, 109, 151, 223, 266], or (2) learning action categories on synthetic data only [161, 162, 207], termed synthetic-to-real domain generalization.

Mixing strategies combine synthetic and real data: De Souza et al. [56] proposed mixing within minibatches, whereas Varol et al. [266] and Liu et al. [147] augmented real data with synthetic variants from new viewpoints. Adaptation methods address distributional shift: Turrissi da Costa et al. [44] improved transfer through dual-head contrastive adaptation with pose-based feature alignment, and Varol et al. [266] combined parametric models with domain adaptation for effective transfer. Ablation studies by De Melo et al. [167] revealed that character diversity and motion variation impact synthetic effectiveness more than environmental factors. This provides design insights for subsequent work.

Transfer effectiveness varies across generation paradigms: gaming-based approaches face substantial challenges [207], motion capture methods show better generalization characteristics [219, 267], while domain-specific platforms achieve targeted improvements

for their intended applications [109, 167]. Puig et al. [193] addressed learning compositions of actions within virtual domains only, demonstrating that synthetic data can support learning in purely simulated environments without real-world transfer.

Our Contribution: To improve cross-domain robustness through synthetic data, we introduce MMGen (IROS 2022), a framework generating novel training modalities from existing ones (RGB, flow, pose) through adversarial optimization. By creating diverse synthetic training domains, MMGen demonstrates that learned domain-level augmentation reduces cross-domain performance gaps.

A fundamental assumption underlying this body of work is that synthetic data is inherently inferior to real data, with research focused on closing the performance gap through adaptation and domain generation techniques. However, this assumption remains largely untested: prior work has not systematically evaluated whether synthetic data, when generated with controlled demographic diversity, could match or exceed real staged datasets for out-of-distribution generalization. Section 5.1 challenges this assumption by demonstrating that synthetic video can outperform real staged data for specialized recognition tasks, suggesting that controlled synthetic diversity can exceed real staged uniformity when the real data itself exhibits limited variation.

Beyond visual appearance transfer, synthetic data generation can address modalities that are difficult or impossible to capture at scale in real settings, particularly physiological signals during human motion.

2.5 SYNTHETIC DATA FOR BIOLOGICAL MOTION ANALYSIS

Deep learning for biomechanical analysis requires large training sets, yet experimental EMG collection remains constrained to dozens of subjects with limited motions [76, 103].

2.5.1 Computational Biomechanics and OpenSim Integration

OpenSim [58, 227] enables biomechanical simulation at scales impossible to achieve experimentally, transforming kinematic data through inverse kinematics and dynamics into muscle activations via physiologically-constrained optimization. Recent advances [27, 77, 101, 115, 309] demonstrate growing interest in musculoskeletal control. In Section 4.2, we combine OpenSim with AMASS [155], which provides large amounts of motion capture data [86, 159]. Vertex-to-marker mappings [17] allow us to bridge representations, with trajectory optimization [260] estimating missing ground reaction forces.

Validation of OpenSim [58, 227] simulations against experimental EMG shows good timing agreement but variable amplitude across muscles [4, 88]. EMG-informed models [24] and optimization frameworks [13] improve alignment by incorporating recorded EMG. This provides a basis for simulations as pseudo-ground truth in data-scarce regimes.

Video-to-biomechanics pipelines estimate kinematics and ground reaction forces from vision. OpenCap [260] derives joint moments and muscle activations from multi-view video with accuracy approaching clinical-grade measurements. Physics-informed net-

Table 2: Comparison between recent muscle activation datasets and *Muscles in Time*.

	Year	Subjects	Act. Vals.	Duration	Actions	GRF	RGB	Depth	Activation	Skeleton	Description
Camargo <i>et al.</i> [28]	2021	22	11	10 min	4	×	×	×	✓	×	×
Feldotto <i>et al.</i> [76]	2022	5	7	10 min	4	×	×	×	✓	×	×
KIMHu [103]	2023	20	4	10 h	3	✓	✓	✓	✓	✓	×
MuscleMap [187]	2023	N/A	20 ^a	~25 h ^b	135	✓	✓	✓	×	✓	×
MiA [38]	2023	10	8	12.5 h	15	✓	✓	×	×	✓	×
MinT (ours)	2024	227	402 ^{c,d}	10 h	187	✓ ^d	✓ ^{d,e}	✓ ^d	✓ ^d	✓	✓

^a Clip-wise binary labels.^b Coarse estimation based on 15,004 clips of 3-9s.^c Muscle strands, some muscles represented by multiple strands.^d Simulated data.^e From [219]

works embed musculoskeletal constraints [299], whereas neural enhancement refines OpenSim outputs to match EMG [248].

2.5.2 EMG and Muscle Activation Datasets

Human Motion Analysis and Datasets Publicly available sEMG databases remain severely limited in scale and muscle coverage [76, 81, 103, 113, 136, 157, 164, 173, 208, 224, 300]: Zhang *et al.* [300] (5 subjects, 8 sensors), KIMHu [103] (20 subjects, 4 upper limb muscles), MiA [38] (10 subjects, 8 muscles across 15 exercises). MuscleMap [187] assigns binary muscle activation labels to 135 actions across 20 muscle groups. Table 2 compares recent EMG datasets to MinT, which features a clearly larger scale in subjects, activation measurements, and motion diversity.

Our Contribution: To enable large-scale training data for muscle activation prediction from vision, we introduce Muscles in Time (MinT) (*NeurIPS 2024*), which couples SMPL motions with physiological signals through biomechanical simulation. By transferring AMASS motion sequences to OpenSim models, MinT provides 10 hours of 402-muscle activation data across 227 subjects, unprecedented for real EMG datasets. This vision-physiology coupling opens research directions in effort estimation, injury risk assessment, and rehabilitation monitoring from vision that were previously constrained by the scarcity of physiological measurement data.

2.6 PRIVACY-PRESERVING METHODS

Deployment of activity recognition systems, whether trained on real or synthetic data, must address privacy concerns that arise when monitoring human behavior in sensitive environments. Privacy preservation techniques in this field include data transformation, obfuscation, adversarial training, and differential privacy methods.

DATA TRANSFORMATION, ANONYMIZATION AND OBFUSCATION-BASED METHODS Downsampling-based methods anonymize data via low-resolution inputs, as demonstrated by Chou et al. [39], Srivastava et al. [243], and Butler et al. [25], who show that action recognition remains feasible even when fine-grained identity cues are removed by aggressive spatial downsampling. Obfuscation-based approaches use off-the-shelf object detectors or segmentation networks to identify privacy-sensitive regions and modify or remove them. Ren et al. [203] synthesize fake faces in place of detected real faces for action detection, while Zhang et al. [301] employ semantic segmentation followed by blurring to hide people or specific body parts for video privacy preservation.

Complementary techniques remove or obfuscate people directly in the pixel space. Image and video inpainting models such as LaMa [247] and ProPainter [306] can erase humans or faces and hallucinate plausible backgrounds from context. Combined with pose-based overlays or 3D avatars, these methods yield visually convincing anonymized videos that preserve scene semantics and motion patterns. However, they do not provide formal privacy guarantees: small residual artifacts or background patterns across scenes can still be exploited for re-identification. These limitations motivate hybrid approaches combining visual anonymization with full-video privacy mechanisms (Section 5.2).

ADVERSARIAL TRAINING Adversarial methods trade action utility against privacy leakage through minimax objectives. Image-based approaches [190, 287] optimize task classifiers and at the same time discourage adversarial privacy classifiers from inferring sensitive attributes. Video extensions [283, 284] minimize action loss and maximize privacy loss on shared representations. Variants selectively preserve attributes [34], mask video tubelets [140], or retain motion while they obfuscate spatial semantics [36, 55, 78].

DIFFERENTIAL PRIVACY FOR VIDEO ANALYSIS Differential Privacy (DP) [67] is widely regarded as a gold standard [57] for formalizing privacy guarantees of algorithms operating on sensitive datasets. DP has been integrated into a variety of machine learning tasks, such as image classification [1, 182, 229, 254, 308], activity recognition [152], language modeling [63, 233, 295], graph learning [214–216], and speech recognition [230]. The most popular method for training deep neural networks with DP is Differentially Private Stochastic Gradient Descent (DP-SGD) [1], which masks the contribution of any single data point to the model updates by clipping per-sample gradients and then adding Gaussian noise to the sum of the clipped gradients. However, this added noise frequently hampers optimization, which leads to performance degradation over non-private training [64, 132]. To address this, various methods have been proposed to improve the utility of DP-SGD, such as adaptive clipping [6], adaptive noise scaling [80], and gradient compression [273]. Alternatively, some works have proposed relaxed notions of DP to protect only a data subset or to provide different privacy levels per data attribute [116, 129, 233].

Empirical work studies how these mechanisms behave in practice. Jayaraman and Evans systematically evaluate differentially private machine learning under membership-inference attacks and show that commonly used large privacy budgets can provide limited effective protection, while tighter budgets often entail substantial accuracy loss [114]. De et al. [57] demonstrate that scaling model and data size can clearly narrow this utility gap for image classification, and Li et al. [141] show that large pretrained language models can be fine-tuned with DP-SGD to reach strong non-private baselines. Both results underscore the importance of scale and pretraining. Practical toolkits such as Opacus [294] and design guides like “How to DP-fy ML” [192] distill implementation best practices, from accounting and clipping strategies to optimizer choices. This makes DP-SGD more accessible to practitioners. In addition, Morsbach et al. [174] conduct a replication study on DP-SGD hyperparameters and find that the interaction between clipping threshold and learning rate dominates other choices. This provides concrete guidance for practitioners.

Luo et al. [152], most relevant to our work, demonstrated the application of DP-SGD to visual HAR. The anonymization and obfuscation methods discussed further above protect individual appearance; they do not provide privacy guarantees for backgrounds and living environments, which reveal sensitive contextual information. Section 5.2 explores hybrid approaches combining avatar-based anonymization with selective differential privacy for background protection. This improves privacy over pure anonymization and is less sensitive to the utility penalties incurred by full DP training.

Our Contribution: To provide privacy protection for both individuals and their environments in video-based activity recognition, we introduce Masked Differential Privacy (MaskDP) (presented at the 2nd International Workshop on Privacy-Preserving Computer Vision @ ECCV 2024), a hybrid framework decomposing videos into non-sensitive anonymized avatar regions and sensitive subject-related background regions, subsequently protected by Differential Privacy. This selective approach provides improved privacy over subject-only anonymization methods, whereas it clearly improves recognition accuracy over standard differential privacy. Thereby, full-video privacy becomes practical for video analysis.

Part II

LEARNING FROM SYNTHETIC DATA

THE SYNTHETIC-TO-REAL DOMAIN GAP

The central challenge in synthetic data generation is the synthetic-to-real domain gap, i. e. models trained on synthetic data must transfer to real-world deployment. To address this gap, we need systematic benchmarks which quantify transfer deterioration, as well as methods which bridge the identified gaps. In this chapter, we introduce infrastructure for cross-domain evaluation and demonstrate two different mitigating approaches. We first establish Sims4Action (Section 3.1), a benchmark which pairs life-simulation-game-based synthetic data with real-world evaluation. This benchmark reveals significant appearance and motion gaps [207]. Having quantified these gaps, we explore two alleviating strategies. One is adaptation through knowledge distillation (OT-DAKD, Section 3.2) [223]. The other, multimodal domain augmentation, builds on findings from Section 3.2. Here we evaluate 15 modality combinations, and our results show, that synthetic multimodal diversity drives cross-domain robustness (MMGen, Section 3.3) [162].

At its root, this cross-domain challenge is created by asymmetry: Training occurs in synthetic domains, whereas deployment occurs in complex, unconstrained real-world environments. The domain gap is present across multiple dimensions, e. g. appearance distributions, motion characteristics, environmental complexity, and sensor noise patterns. Its quantification and mitigation require systematic approaches.

Addressing this synthetic-to-real gap needs two complementary components. Benchmarks must provide controlled environments which quantify domain gaps and evaluate cross-domain transfer methods under standardized protocols. Methods, in turn, must bridge the identified gaps, either through adaptation techniques which align synthetic and real distributions, or through generalization approaches which create robust synthetic training data. Prior to this work, synthetic datasets for human activity recognition existed, yet cross-domain evaluation for activities of daily living was not a primary design criterion. This limited the systematic assessment of synthetic→real transfer.

We take on these challenges through three complementary contributions. To enable systematic evaluation, we establish cross-domain benchmark infrastructure which pairs gaming-based synthetic data with real-world activities of daily living and quantifies the magnitude of the appearance and motion gaps. To mitigate the identified gaps, we explore two complementary approaches. The first is adaptation through pose-based prediction calibration using optimal transport; it requires distributional knowledge of the target domain. The second is generalization through systematic multimodal domain augmentation, which improves robustness without target-domain access, as it generates diverse synthetic training variations.

3.1 LEARNING FROM LIFE SIMULATION VIDEO GAMES

This section is based on our publication at IROS 2021 [207]. © 2021 IEEE.

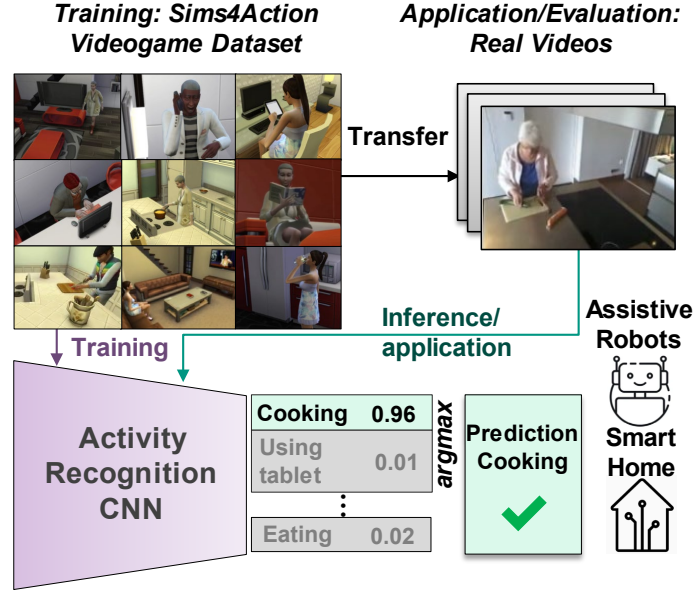


Figure 3: This work leverages the life simulation video game THE SIMS 4 for inexpensive ADL training data which is then utilized in the GAMING→REAL setting to suit real-world applications, such as smart homes and assistive robots.

Can synthetic training data from life simulation video games transfer to real-world activity recognition systems? Activity recognition for assistive household robots relies on deep neural networks trained on carefully curated and manually labelled datasets [52, 54, 112, 228, 236]. While synthetic data offers potential advantages—controllable parameters, unlimited scale, freedom from privacy constraints—the transfer from synthetic training to real-world deployment remains poorly quantified. Evaluating synthetic→real transfer requires systematic benchmarks with controlled evaluation protocols and paired synthetic-real data, which have been largely absent in ADL recognition research.

Life simulation video games provide an opportunity to address this benchmarking challenge. Unlike purpose-built synthetic data generation, which lacks corresponding real-world datasets, commercial life simulation games can be configured to match existing real ADL datasets. THE SIMS 4¹ is specifically designed to give players full control of everyday situations, enabling flexible design of environments, camera angles, and human appearances. This controllability allows creating synthetic data that mirrors the activities, environments, and camera configurations of existing real ADL benchmarks.

We introduce Sims4ACTION, a 10-hour synthetic dataset designed to pair gaming-based synthetic data with real-world evaluation from Toyota Smarthome [52] (Figure 3). We create six virtual environments, capture data from four fixed cameras per location plus a moving camera mode, and design eight subjects with diverse appearances (two young

¹ The Sims 4 is a commercial game by Electronic Arts (EA):

<https://web.archive.org/web/20260117190507/https://www.ea.com/games/the-sims/the-sims-4>

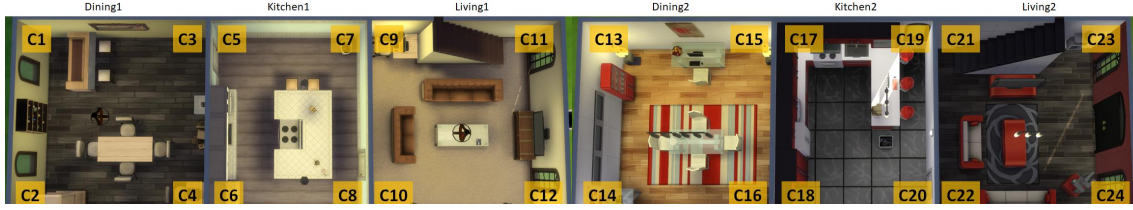


Figure 4: Snapshots of the virtual apartments we designed to simulate household situations. The two homes cover a dining room, a living room and a kitchen. C1 - C24 marks the 24 camera placements used in the static camera mode recordings.

adults, two adults, four elderly people, equal gender distribution). The dataset features ten household activities matching Toyota Smarthome’s activity taxonomy, which we trigger by “playing” the desired behaviours in the game. This paired dataset enables systematic Gaming→Real evaluation, where models trained on synthetic data are evaluated on real videos from Toyota Smarthome. As our visual model, we adopt two off-the-shelf video classification architectures, which we train with the *Sims4Action* data only. Results provide systematic quantification of the synthetic→real domain gap with two key findings: (1) models trained on gaming-based synthetic data achieve above-random performance on real-world evaluation, demonstrating that cross-domain transfer is possible, yet (2) substantial performance degradation reveals that modern recognition algorithms are highly susceptible to domain shifts. This systematic benchmark enables controlled evaluation of domain adaptation methods and motivates the pose-based representations and optimal transport techniques introduced in Section 3.2. By establishing paired synthetic-real evaluation protocols, *Sims4Action* provides infrastructure for research on cross-domain transfer that was previously difficult to assess systematically.

3.1.1 *Sims4Action*

This section details the dataset collection process and characteristics. The main dataset statistics are provided in Table 3.

The *Sims4Action* dataset is collected using *THE SIMS 4* life-simulation video game, where players guide virtual humans through everyday activities. The simulation captures complex character traits, varying emotional states, diverse subject appearances, and different age groups. We leverage this environment to induce the situations we want to recognize and use the built-in video capture tool for the recordings. The tool allows us to select the time frame, capture size, and image resolution. We configure recordings at 30 FPS with 640×368 pixel resolution. The main stages of our data collection comprise three steps: 1) designing virtual subjects, 2) designing home environments, and 3) actively playing behaviour categories we want to recognize while recording the gameplay.

VIRTUAL SUBJECTS For systematic cross-domain evaluation, subject diversity is essential. *THE SIMS 4* enables detailed subject personalization, capturing characteristics such as age, gender, body type, clothing, facial features, and even walking style. We design eight virtual subjects diversifying appearance, gender (four male and four female), and



Figure 5: Eight subjects (four elderly people, two adults and two young adults) are covered in the SIMS4ACTION dataset.

age groups (two young adults, two adults, four elderly people) to enable subject-wise cross-validation splits. An overview of the created subjects is provided in Figure 5.

INDUCING VIRTUAL BEHAVIOURS In THE SIMS 4, the virtual subject (Sim) follows given instructions. We explicitly trigger desired behaviours by guiding the Sim to engage in specific activities, which range from cooking or taking pills to rare events such as having a wedding party or morning sickness during pregnancy. In this work, we focus on ten everyday household activities popular in previous works on recognition in real smart homes. Examples for each of the classes are provided in Figure 6. While balanced datasets are rare in real-life [52], synthetic data collection enables equal distribution of classes with approximately one hour of footage per category. The number of videos per activity varies with activity duration: *drink* clips are typically 15-30 seconds, while *cook* and *eat* range from 30 seconds to one minute. Some activities can be controlled by the player (*get up/sit down*, *use computer*, *use tablet*, *walk*, *watch TV*), whereas others complete automatically after a certain duration (*cook*, *drink*, *eat*, *read book*, *use phone*). Figure 7 illustrates sample statistics for the individual categories.

VIRTUAL HOMES As users can create customized interiors, we can generate a large variety of environments for data collection. We design virtual environments to match the Toyota Smarthome [47, 52] framework, recreating digital versions of their households with the game’s build mode. Our dataset covers two simulated apartments with three rooms: a kitchen, a dining room, and a living room. While some activities (such as *talking on the phone*) happen at various locations, other events, such as *cooking* have natural location biases (the location distribution of the captured activities is provided in Figure 8).

The simulation environment allows us to freely choose the camera angle. We record the videos with two different camera modes: (1) the *fixed camera* mode, where the camera stays in one determined position for the entire video, and (2) the *moving camera* mode, where the camera is set at various random angles and moves every few seconds during the video. For the *fixed camera* mode, we define four different camera angles in each room, resulting in a total of 24 distinct camera views. An overview of the modeled homes and the camera views is shown in Figure 4.

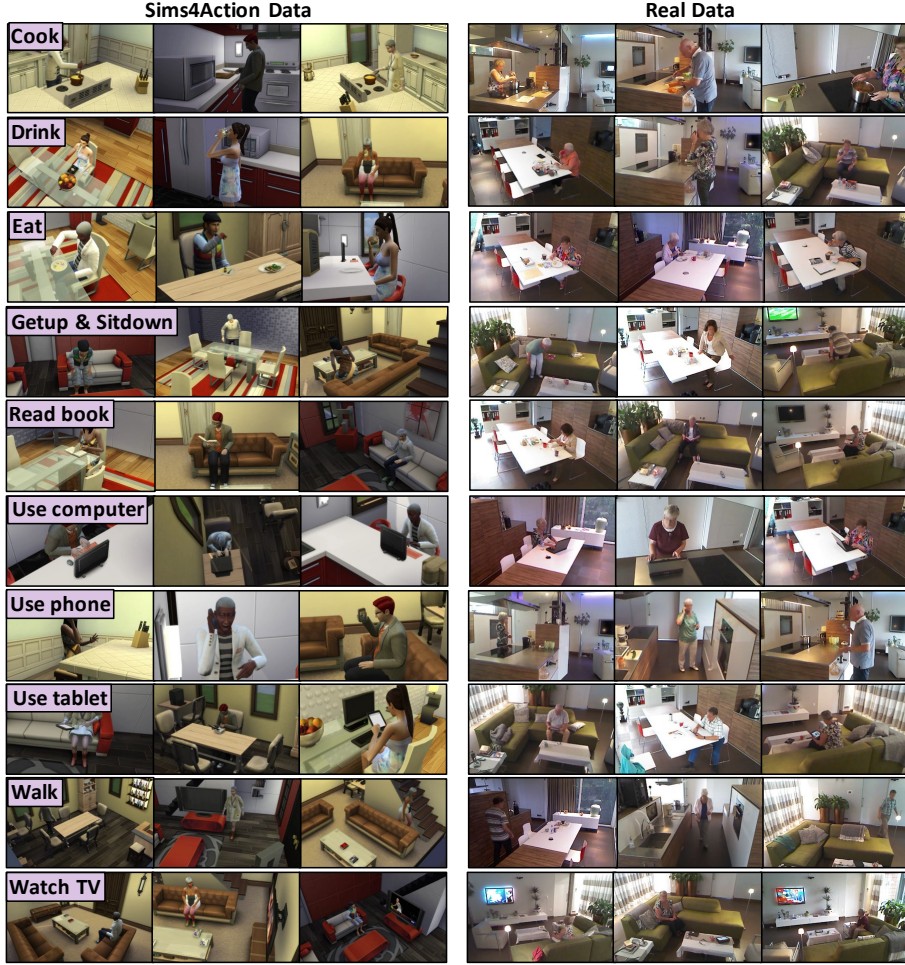


Figure 6: Examples of subjects being engaged into different domestic activities in the Sims4Action dataset and the real-life counterparts used for the GAMING→REAL Benchmark, which are derived from Toyota Smarthome [52].

3.1.1.1 Dataset and Splits Statistics

The dataset comprises 942 videos with a total length of 10 hours, 25 minutes, and 32 seconds and a roughly equal representation of activity classes (ca. 1 hour each) and subjects (ca. 1 hour, 17 minutes each).

SIMS4ACTION SPLITS Generalization to unseen subjects is essential in ADL recognition. We therefore divide our dataset *subject-wise* into *train* (5 subjects), *val* (1 subject), and *test* (2 subjects). We learn model parameters on *train*, select checkpoints and hyperparameters on *val*, and present final evaluation on *test* (overview of the split statistics provided in Figure 7). Since behaviour complexity and duration vary strongly, we divide the videos into chunks of three seconds and use them as samples in our benchmark. Hence, the classification task on our dataset amounts to assigning the correct activity label to a three second video.

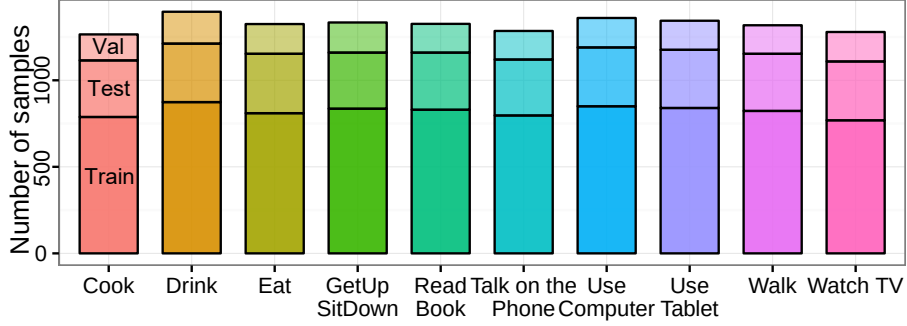


Figure 7: Sample frequency of the activities in *train*, *test* and *val*. In contrast to real-life data collection [52], life simulation video games give us control of the data distribution resulting in a well-balanced dataset. (A sample corresponds to a 3s snippet with the assigned label.)

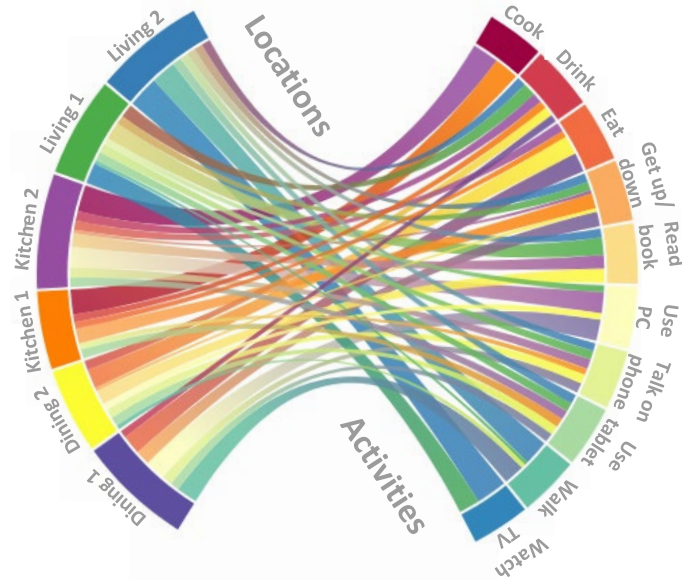


Figure 8: *Where did the activities take place?* The cord diagram visualizes the correspondences of the captured behaviours and the rooms it took place in. Best viewed in color.

GAMING→REALBENCHMARK Since neural networks are sensitive to distribution shift, transferring recognition capabilities from the GAMING→REAL domain is challenging. To enable sustained transfer from video games to actual robotics or smart home applications, SIMS4ACTION is accompanied by a real data benchmark. To achieve this, we build correspondences between our categories and activities captured in the real Toyota Smarthome dataset [52] and use the Toyota Smarthome cross-subject test set with the corresponding classes for evaluation. Note that we do not directly take Toyota Smarthome classes as-is, and certain modifications and consolidations are required to match the SIMS4ACTION classes. For example, we consolidate different variants of cooking activities into a single cooking class. Figure 6 compares the videogame examples and their real-life counterparts for every category. Overall, the GAMING→REAL test set comprises 4412 videos with the corresponding ten activities inside a real smart home, where each video is used as a test sample, as in [52]. This real-life test data is not evenly distributed, with the number of samples ranging between 15 for *use tablet* and 1225 for the *walk* category.

Table 3: **Statistics of the Sims4Action dataset.** Samples represent 3-second video segments with activity labels.

Property	Train	Val	Test	Total
Duration (min)	388.5	79.6	157.5	625.6
Samples*	8,217	1,684	3,331	13,232
Subjects	5	1	2	8
Male/Female	2/3	1/0	1/1	4/4
Static cam.	✓	✓	✓	✓
Moving cam.	✓	✓	✓	✓
Static views	24	24	24	24
Activity classes	10	10	10	10
Min. vids/class	769	150	324	1,265
Avg. vids/class	822	168	333	1,323

Table 4: **Per-class performance of I3D** on the Sims4Action dataset. Precision, recall, per activity class.

True Activity	Prec. [%]	Rec. [%]	F1 [%]
Cook	98.99	93.61	96.22
Drink	84.74	94.44	89.33
Eat	93.90	95.65	94.77
Getup / Sitdown	78.44	85.37	81.76
Read book	97.15	88.35	92.54
Use computer	99.34	94.06	96.63
Talk on phone	78.17	96.67	86.44
Use tablet	96.43	85.71	90.76
Walk	100.00	57.37	72.91
Watch TV	76.74	100.00	86.84

3.1.2 Neural Architectures

To evaluate our idea of ADL recognition by learning from life simulation video games and provide a competitive benchmark based on the Sims4Action dataset, we adapt two off-the-shelf CNNs for general activity recognition.

INFLATED 3D CNN The Inflated 3D ConvNet (I3D) by Carreira and Zisserman [30] is a widely adopted video classification architecture utilized for both general- and ADL recognition [30, 52]. The model benefits from the object recognition architecture Inception-v1 [110] and modifies it to deal with the temporal domain via expansion of 2D- to 3D kernels and ablations on how to integrate pooling operations.

Inherited from the ancestor network, I3D is 27 layers deep and stacks three convolution layers at the beginning, nine inception modules which themselves are two layers deep, one fully-connected layer at the end as well as five max- and average-pooling layers.

SEPARABLE 3D CNN While I3D offers excellent classification accuracy, it is a heavy architecture with > 12 million parameters and might be impractical in robotic applications. With this in mind, Xie et al. [288] redesigned the I3D architecture by considering (1) switching to a top-heavy model, (2) temporally separable convolutions, and (3) spatio-temporal feature gating. These alterations led to a major reduction in floating point operations (FLOPs) and culminated in the new Separable 3D architecture (S3D). The so-called top-heavy model property reduces FLOPs by switching early 3D convolutional blocks in the I3D architecture with standard 2D convolutions and only in later layers involves temporally separable 3D convolutions.

IMPLEMENTATION AND TRAINING DETAILS In all experiments except for the baselines, we train our networks end-to-end on Sims4Action for 100 epochs. We use the I3D architec-

Table 5: Evaluation of the ADL recognition in the GAMING→GAMING setting of our benchmark. The performance metrics are standard- and balanced accuracies.

Model	Weight init.	Normal Acc [%]			Balanced Acc [%]		
		Top 1	Top 3	Top 5	Top 1	Top 3	Top 5
Random	None	10.0	33.3	50.0	10.0	33.3	50.0
Sep. 3D Net	Random	56.7	79.4	91.2	56.5	79.3	91.3
Sep. 3D Net	K400	84.7	96.7	98.7	84.6	96.7	98.7
I3D Net	Random	67.0	87.8	94.3	66.9	87.8	94.3
I3D Net	K400	89.1	98.4	99.6	89.1	98.3	99.5

Table 6: Recognition results in the GAMING→REAL setting. Such transfer is a hard task for current recognition models, although all methods surpass the random baseline.

Model	Weight init.	Normal Acc [%]			Balanced Acc [%]		
		Top 1	Top 3	Top 5	Top 1	Top 3	Top 5
Random	None	10.0	33.3	50.0	10.0	33.3	50.0
Sep. 3D Net	Random	20.0	45.9	64.6	12.4	34.4	58.7
Sep. 3D Net	K400	34.1	58.6	75.9	23.2	43.9	62.3
I3D Net	Random	18.0	44.4	66.1	10.9	31.5	53.6
I3D Net	K400	22.8	47.8	65.0	23.3	47.0	64.4

ture with an input frame size of 224×224 or the S3D architecture with an input frame size of 128×128 . The network architecture-related parameters (e.g., dropout, input size) were set according to the original architectures [30, 288]. We apply multiple augmentations during training including random rotations ($\pm 20^\circ$), hue changes ($\pm 180^\circ$), saturation changes ($\pm 100\%$), brightness changes ($\pm 80\%$), and random cropping. For testing, we only scale the input clips to the desired input size and take a center crop. All videos are divided into 90 frame chunks. During training, we randomly select single 30 frame clips from each chunk and classify them separately. For the final classification on videos, we only classify the center chunk of each video. We use the Adam optimizer with a learning rate of $1e-4$ and weight decay $1e-4$. For training on the Sims dataset, we used a mini-batch size of 48 for S3D and 20 for I3D. For our transfer learning experiments on ADL, we exchanged the last layer to match the number of 31 classes and then trained on that layer for one epoch in order to fine-tune the parameters. We used a batch size of 64 chunks for S3D and a batch size of 32 chunks for I3D, all other settings are equal to the Sims training.

3.1.3 Experiments

This section investigates how well off-the-shelf activity recognition models described in Section 3.1.2 can leverage our synthetic Sims4ACTION data to recognize activities of daily living. To gather this insight, we present results for two scenarios: (1) recognition of ADL in the video game domain and (2) recognition transfer from video game-trained models

to real-life ADL classification. We use the standard accuracy and the balanced multi-class accuracy (average of the individual class recalls) as our main evaluation metrics.

3.1.3.1 ADL Recognition on Video Game Data

First, we explore how well the described CNN architectures handle the video game data in the GAMING→GAMING setting, where we train and evaluate in the synthetic domain. In Table 5 we present the results of the S3D and I3D networks both trained on the SIMS4ACTION dataset. Similarly to Toyota Smarthome [52], we initialize the networks using the Kinetics-400 classification models [123], but also report the results when training from scratch. As expected, the accuracy of the pretrained models is considerably higher, since 3D CNNs are known to strongly benefit from pretraining. I3D initialized with Kinetics-400 achieves 89.1% Top-1 accuracy, representing strong performance for 10-class ADL recognition. The benefit of pretraining is substantial: I3D improves from 67.0% when trained from scratch to 89.1% with Kinetics initialization, a 22.1 percentage point gain that underscores the value of large-scale video pretraining even for synthetic gaming data.

I3D achieves 98.4% accuracy in the Top-3 evaluation scenario, approaching perfect recognition. This indicates that when the model errs, the correct class typically appears among the top predictions, suggesting confusion between semantically similar activities rather than fundamental misclassification. In addition to the overall accuracy, we examine model performance for the individual classes, with the exact precision, recall and F1-score (harmonic mean of the precision and recall) provided in Table 4.

Interestingly, the best recognition in terms of F1-score is achieved for the activities *cook* and *use computer*. This is likely due to the distinctiveness of the cooking utensils and the computer making them easily distinguishable patterns in those activities. *Get up/sit down* is quite ambiguous and hard to disentangle from other activities like *watch TV* or *eating* where the virtual human is also in a seated position resulting in lower precision. These results demonstrate that modern video architectures can effectively learn activity patterns from gaming-based synthetic data when training and evaluation distributions match, achieving accuracies comparable to real-world ADL benchmarks.

3.1.3.2 ADL Recognition in the Real-World

We evaluate cross-domain transfer by applying models trained on SIMS4ACTION to real videos from the Toyota Smarthome dataset [52] (Figure 6), as described in Section 3.1.1.1. Results are reported in Table 6.

Cross-domain performance reveals substantial degradation. I3D (Kinetics-initialized) drops from 89.1% in the Gaming→Gaming setting to 22.8% on real-world data, a 66.3 percentage point decline. Similarly, S3D experiences a 50.6 percentage point drop (from 84.7% to 34.1%). This divergence between in-domain and out-of-domain performance quantifies the severe synthetic-to-real domain gap for gaming-based generation.

This severe degradation is particularly striking as humans recognize these activities across domains effortlessly, revealing neural networks’ unintuitive sensitivity to visual distribution shift despite identical semantic content. The results show that while models learn activity patterns from synthetic data, they remain heavily dependent on appearance features that differ drastically between gaming and real environments.

Despite this substantial gap, all models clearly outperform the random baseline (10%), demonstrating that gaming-based synthetic data does transfer meaningful activity patterns. The cross-domain results establish both the promise and limitations of gaming-based generation: it provides accessible synthetic data collection but faces severe real-world deployment constraints, motivating the domain adaptation work in Section 3.2 and domain generation approaches in Section 3.3.

3.2 DOMAIN ADAPTIVE KNOWLEDGE DISTILLATION

This section is based on our publication in *CVPR ROSE Workshop, 2022* [223]. © 2022 IEEE.

A short note on the contribution within this thesis: Schneider et al. [223] present two major ideas: the usage of human pose-sequences as a supervision signal for self-supervised representation learning and the usage of optimal transport for domain adaptive knowledge distillation. Only the latter falls within the scope of this thesis. Hence, we introduce the representation learning method briefly as context in this section, but focus on the domain adaptation part.

Section 3.1 established Sims4Action as a testbed for synthetic-to-real activity recognition, revealing substantial challenges in cross-domain transfer. Models trained on game-based synthetic data experience accuracy drops exceeding 60% when evaluated on real-world Toyota Smarthome videos [52], despite strong within-domain performance. Standard neural network-based approaches [30, 288] pre-trained on supervised RGB datasets struggle when appearance changes drastically [207].

Having quantified the magnitude of this domain gap, the question becomes: how can we enable effective transfer from synthetic to real domains? We address this through two complementary approaches. First, we employ self-supervised representation learning that combines pose and video modalities to learn domain-invariant features, encouraging the feature extractor to focus on human body regions and motions across network layers. Second, we introduce an optimal transport-based domain adaptation technique that calibrates prediction distributions to match target domain characteristics. While both approaches can be combined, this section focuses on the domain adaptation method, which can be applied independently of any pre-trained feature extractor.

Our method addresses a fundamental challenge in unsupervised domain adaptation: aligning prediction distributions across domains without target labels. Existing approaches often match within-network feature distributions using adversarial losses [35, 83], but provide no guarantees on output prediction marginals. Our approach optimizes the classifier by directly shaping the output prediction distribution to a desired marginal.

The technique employs teacher-student knowledge distillation with optimal transport-based prediction calibration, which we refer to as Optimal Transport-based Domain Adaptive Knowledge Distillation (OT-DAKD). Given unlabeled target domain data, we assume a plausible marginal label distribution and enforce this distribution during batch-wise training. The teacher network produces predictions on target batches, which are then calibrated using the Sinkhorn-Knopp algorithm to match the desired marginal. A student network, initialized with the teacher weights, is trained to predict these calibrated soft labels, effectively learning to align its predictions with the expected distribution.

3.2.1 Problem Formulation

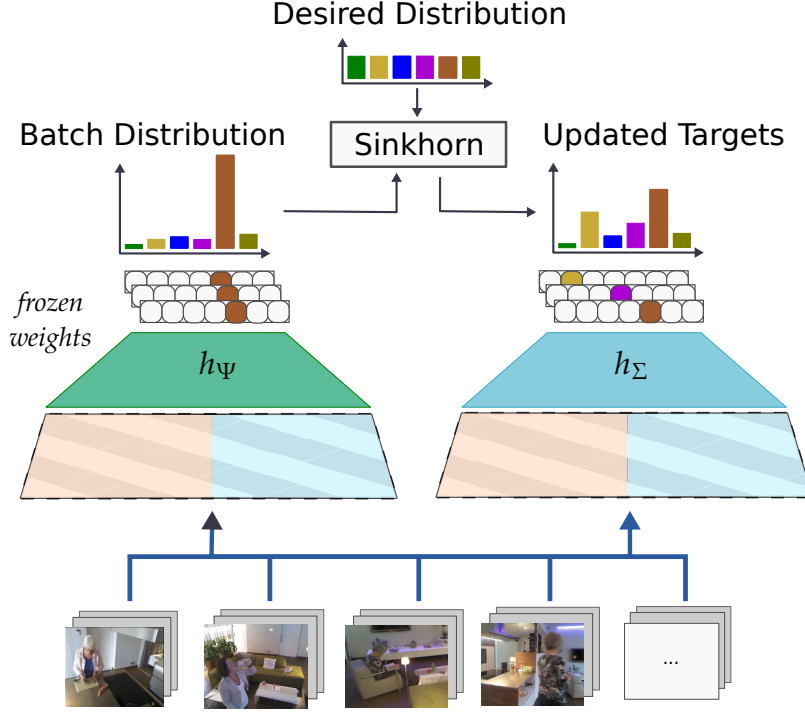


Figure 9: Overview of our domain adaptation method. We evaluate the batch distribution of label predictions and use the Sinkhorn-Knopp algorithm to calculate target predictions which conform to a desired distribution.

The domain adaptation problem for this section can be formulated as follows. Let $f_\Theta : \mathcal{X} \rightarrow \mathbb{R}^m$ denote a feature extractor with parameters Θ that maps input samples to representation vectors $y \in \mathbb{R}^m$, where m is the representation dimensionality. A classifier h_Ψ with parameters Ψ maps these representations to class predictions. We assume access to a labeled source domain dataset $\mathcal{D}_S = \{(\mathbf{x}_S^i, y_S^i)\}_{i=1}^{N_S}$ with samples $\mathbf{x}_S^i \in \mathcal{X}$ and labels $y_S^i \in \{1, \dots, K\}$ for K classes, and an unlabeled target domain dataset $\mathcal{D}_T = \{\mathbf{x}_T^j\}_{j=1}^{N_T}$ with samples $\mathbf{x}_T^j \in \mathcal{X}$. The goal of domain adaptation is to adapt the classifier to perform well on the target domain without requiring target labels.

In the experiments presented in this work, we employ a two-stream architecture combining video and pose modalities. The video stream uses S3D [288], the pose stream processes SkeleMotion representations [26] through a lightweight CNN encoder. Pre-training employs pose-based self-supervised contrastive clustering in a common feature space [223] to encourage domain-invariant representations.

While we employ this specific architecture in our experiments, the domain adaptation method presented in Section 3.2.2 is agnostic to these architectural choices. The approach can be applied to any feature extractor that produces meaningful representations for both source and target domains.

3.2.2 Optimal Transport for Domain Adaptation

Formally, let h_Ψ denote the teacher network with pre-trained weights Ψ . For a batch of B target domain samples $\mathbf{X} = \{\mathbf{x}^i\}_{i=1}^B$ from $\mathcal{D}_T$ without labels, the teacher produces logit predictions $\mathbf{z}_t^i \in \mathbb{R}^K$ for each sample $\mathbf{x}^i$, where K is the number of classes. The student network h_Σ , initialized with weights $\Sigma = \Psi$, produces corresponding predictions $\mathbf{z}_s^i \in \mathbb{R}^K$. Note that we do not require samples from the source domain during adaptation, making this a source-free domain adaptation approach. An overview of the method is shown in Figure 9.

SINKHORN-KNOPP CALIBRATION To generate reliable pseudo-labels, we calibrate the teacher’s predictions to match the known target class distribution $\mathbf{r} \in \mathbb{R}^K$ with $\sum_{k=1}^K r_k = 1$. We first compute soft predictions $\mathbf{p}_t^i = \text{softmax}(\mathbf{z}_t^i)$ and form a score matrix $\mathbf{Q} \in \mathbb{R}_{>0}^{K \times B}$ with temperature $\epsilon > 0$, where column i contains scores for sample i :

$$Q_{ki} = \exp\left(\frac{p_{t,k}^i}{\epsilon}\right).$$

Let the desired marginals be the target prior for rows and uniform columns:

$$\text{rows: } \mathbf{r} \in \mathbb{R}^K, \quad \text{cols: } \mathbf{c} = \frac{1}{B} \mathbf{1}_B \in \mathbb{R}^B.$$

The Sinkhorn-Knopp algorithm iteratively normalizes $\mathbf{Q}$ to match these marginals. Initialize $\mathbf{Q} \leftarrow \mathbf{Q} / \|\mathbf{Q}\|_1$. For $n = 1, \dots, N$:

$$\mathbf{u} \leftarrow \mathbf{Q} \mathbf{1}_B, \quad \mathbf{Q} \leftarrow \mathbf{Q} \cdot \text{diag}(\mathbf{r} \oslash \mathbf{u}),$$

$$\mathbf{Q} \leftarrow \mathbf{Q} \cdot \text{diag}(\mathbf{c} \oslash \mathbf{Q}^\top \mathbf{1}_K),$$

where $\oslash$ denotes element-wise division. After convergence, column-normalize $\mathbf{Q} \leftarrow \mathbf{Q} \cdot \text{diag}(\mathbf{1}_B \oslash \mathbf{Q}^\top \mathbf{1}_K)$ and apply softmax to obtain calibrated predictions for sample i :

$$\tilde{\mathbf{p}}_t^i = \text{softmax}(\mathbf{Q}_{:,i}) \in \mathbb{R}^K.$$

This formulation ensures the batch-level class distribution matches the target prior $\mathbf{r}$ while treating each sample equally, preventing over-prediction of majority classes.

The Sinkhorn-Knopp algorithm iteratively balances the prediction matrix through row and column normalizations. Intuitively, the row normalization ensures that the overall batch predictions match the desired class distribution $\mathbf{r}$, while the column normalization ensures each sample receives equal treatment rather than concentrating probability mass on easy samples. This alternating normalization converges to a doubly-stochastic matrix that satisfies both constraints simultaneously. The temperature parameter ϵ controls the smoothness of the resulting distribution: lower values produce sharper, more confident pseudo-labels, while higher values yield softer distributions that hedge uncertainty. By enforcing batch-level distributional constraints while maintaining per-sample fairness, the algorithm generates calibrated pseudo-labels that guide the student network toward predictions aligned with the target domain’s statistical properties.

TRAINING OBJECTIVE Pseudo-labels are extracted via $\hat{y}^i = \arg \max_k \tilde{p}_{t,k}^i$. The student is trained with cross-entropy loss:

$$\mathcal{L} = -\frac{1}{B} \sum_{i=1}^B \log(\text{softmax}(\mathbf{z}_s^i)_{\hat{y}^i}).$$

The student weights Σ are updated via backpropagation while teacher weights Ψ remain frozen. We use $N = 12$ iterations and $\epsilon = 0.05$.

The primary advantage of this approach is the ability to directly shape the overall output marginal distribution to match expected class frequencies, addressing prediction calibration at its source. Adversarial domain adaptation methods such as DANN [83] or TA³N [35] match feature distributions but provide no guarantees on prediction marginals. In contrast, our method explicitly enforces the desired label distribution through optimal transport, providing direct control over output predictions.

This distribution-level control is particularly valuable when class imbalances differ between domains, as is the case for Sims4Action→Toyota Smarthome transfer. In the fully unsupervised setting where the target distribution is unknown, the source dataset label distribution serves as a reasonable default assumption. However, the method can be extended to weakly supervised domain adaptation by providing an oracle-based distribution or by sampling a small subset of target domain samples to estimate the actual target distribution, which is especially beneficial when domain distributions are known to differ substantially.

3.2.3 Experiments

We evaluate our method primarily on the Sims4Action→Toyota Smarthome transfer introduced in Section 3.1, which provides the most challenging synthetic-to-real benchmark for ADL recognition. The Gaming→Real setting represents the core motivation for this work: enabling practical deployment of models trained on inexpensive game-based data.

We use Kinetics-400 [123] with Kinetics-Skeleton [290] pose sequences for unsupervised pre-training of a frozen P-HLVC feature extractor [223] and evaluate on the Sims4Action [207]→Toyota Smarthome [52] transfer as described in Section 3.1. We compare our domain adaptation method to state-of-the-art baselines including VideoDG [293], TA³N [35], Kinetics-400-pretrained S3D [288], and vanilla P-HLVC.

3.2.3.1 Results

Tables 7 and 8 present results for domain generalization and adaptation, following the protocol from Section 3.1. We report accuracy and mean Per-Class Accuracy (mPCA) for both full-video and mid-chunk predictions. Pre-training significantly improves generalization performance (Table 7). Best results are achieved by keeping the classification head shallow; models with three fully-connected layers or full end-to-end fine-tuning perform worse, likely due to overfitting and degradation of transferable backbone representations.

This suggests that when the pre-trained encoder already captures action-descriptive representations, linear probing with minimal trainable weights suffices for cross-domain transfer. This is especially the case for pose-motion pre-trained encoders, which focus on human body regions and movement patterns, making them well-suited for activity

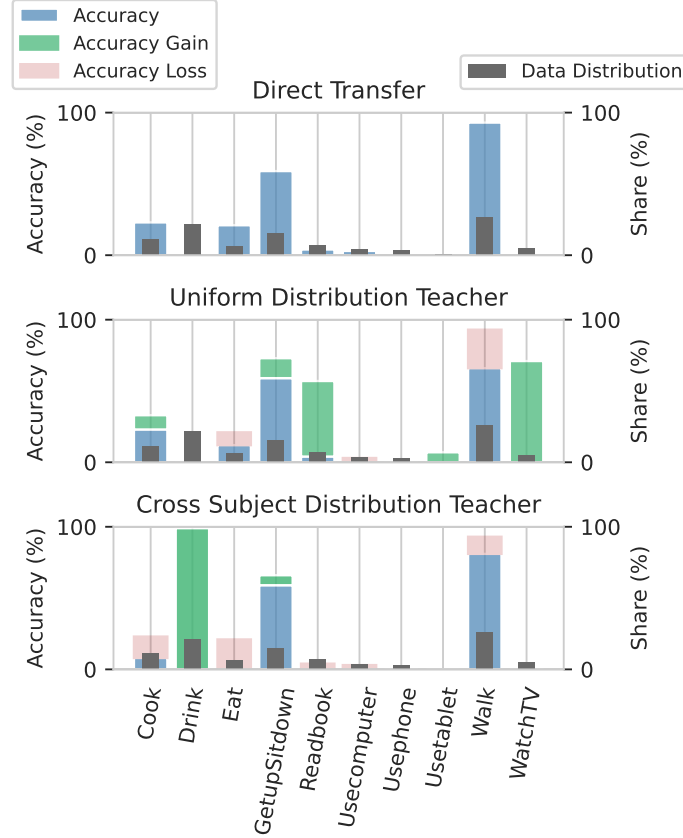


Figure 10: Comparison of direct transfer (top) with unsupervised distribution training assuming the source label distribution (uniform), or using the target cross-subject train label distribution.

recognition across visual domains. In contrast, when using more general pretraining approaches not specifically focused on human motion, deeper classification heads with additional trainable parameters become necessary but increase susceptibility to domain shift. The more weights require adaptation, the greater the risk of overfitting to source domain characteristics and forgetting transferable features. Conversely, shallow heads with fewer trainable weights are less sensitive to domain gaps, provided the encoder representations are already well-suited for the target task.

Among domain generalization baselines, pose-based representation learning [223] with Video+Pose achieves 40.4% accuracy and 29.0% mPCA, outperforming supervised Kinetics-400 pre-training (36.0% / 27.3%) on both metrics, demonstrating the value of domain-invariant pose-based representations.

Table 8 presents domain adaptation results with optimal transport-based calibration. In the unsupervised setting, our domain adaptation method achieves 40.6% accuracy and 31.3% mPCA, substantially outperforming TA³N’s 8.8% / 12.7%. We evaluate two calibration strategies: using the source dataset label distribution (unsupervised) or the target cross-subject train set distribution (weakly supervised). In the weakly supervised setting, the method achieves 53.4% accuracy, outperforming TA³N’s label-based approach (33.1%) by 20.3 percentage points. The source distribution calibration yields better mPCA (31.3% vs. 25.5%) as this metric treats all classes equally, while target distribution calibration maximizes standard accuracy by accounting for class imbalances.

Table 7: Domain generalization baseline results. Pre-training with pose-based representations improves transfer from synthetic to real domains.

Method	Pre-training	Full Video		Mid-Chunk	
		Acc	mPCA	Acc	mPCA
Domain Generalization (No Pre-Training)					
S3D [207]	-	18.5	13.4	20.0	12.4
P-HLVC (Video)	-	13.4	10.0	13.0	9.1
Domain Generalization (With Pre-Training)					
S3D [207]	K400	36.0	27.3	34.1	23.2
TA ³ N [35]	ImageNet	14.4	14.2	-	-
VideoDG [293]	ImageNet	21.1	25.7	-	-
APN [293]	ImageNet	17.2	22.1	-	-
P-HLVC (Video)	P-HLVC	38.8	21.3	35.6	19.8
P-HLVC (Video+Pose)	P-HLVC	40.4	29.0	38.3	28.0

K400: Kinetics-400, **P-HLVC**: Pose-based contrastive pre-training
Results show domain generalization from synthetic (Sims4Action) to real (Toyota Smarthome) data.
[223] report slightly reduced results for TA³N and VideoDG, for fairness we list the improved results from [162] here.

Figure 10 illustrates these distribution effects per-class. Direct transfer succeeds primarily on motion-distinctive activities (walking, get up/sit down) and those with distinctive contexts (cooking, eating). Uniform distribution calibration reduces walking’s over-prediction, redistributing performance across classes and improving mPCA (31.3%). However, this treats all classes equally regardless of target frequencies. Target distribution calibration aligns predictions with true class frequencies: frequent classes gain higher recall through stronger prediction signals, while infrequent classes receive less emphasis. This frequency alignment maximizes overall accuracy (53.4%) at the cost of balanced performance, explaining the mPCA versus accuracy trade-off observed in Table 8.

The optimal transport-based approach demonstrates that explicit prediction distribution control improves cross-domain transfer when target domain characteristics are known or can be estimated. However, this adaptation strategy requires some knowledge of the target domain, whether through distributional assumptions or weak supervision. An alternative question emerges: rather than adapting to specific target domains, can we improve cross-domain robustness by diversifying the synthetic training data itself? Section 3.3 explores this complementary approach through multimodal domain generation, where learned synthetic variation serves as an alternative pathway for addressing domain gaps without requiring target domain access.

Table 8: Domain adaptation results from Sims4Action to Toyota Smarthome.

Method	Target Supervision	Full Video		Mid-Chunk	
		Acc	mPCA	Acc	mPCA
Unsupervised Domain Adaptation		Using Source Label Distribution			
TA ³ N [35]		8.8	12.7	-	
P-HLVC + OT-DAKD		40.6	31.3	36.0	28.1
P-HLVC + OT-DAKD [†]		39.4	29.3	36.3	27.1
Weakly Supervised Domain Adaptation		Using Target Label Distribution			
TA ³ N [35]	Labels	33.1	13.4	-	
P-HLVC + OT-DAKD	Distribution	53.4	25.5	49.9	24.3
P-HLVC + OT-DAKD [†]		53.2	26.0	49.3	25.5

[†] Classifier using three linear layers instead of one.

3.3 MULTIMODAL DOMAIN GENERATION FOR DOMAIN-TRANSFER

This section is based on our paper published at IROS 2022 [162]. © 2022 IEEE.

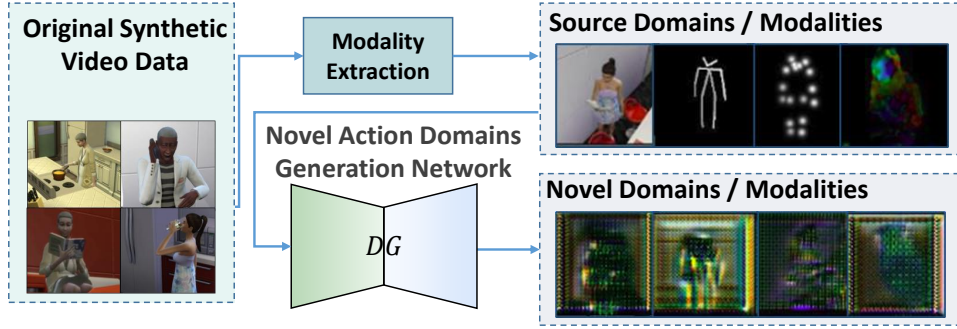
Can generative adversarial networks create additional training domains to improve cross-domain transfer? This section introduces MMGen, a framework that employs GANs to generate novel modalities from existing multi-modal data (RGB, optical flow, pose). The work demonstrates that synthetic domain augmentation reduces performance gaps when deploying models trained on synthetic data to real-world scenarios. Beyond validating the domain generation paradigm, this section provides a systematic cross-domain transfer analysis for multimodality.

Simulations are an excellent way of economical data collection, but SYNTHETIC → REAL domain shifts lead to severe accuracy degradation when recognizing Activities of Daily Living (ADLs), as quantified in Section 3.1.

Section 3.2 demonstrated that pose-based representations provide domain-invariant features that support cross-domain transfer, establishing the value of alternative modalities beyond RGB appearance. Building on these findings, this section explores whether systematic leveraging of multiple modalities—pose heatmaps, skeletal limbs, optical flow—can improve synthetic-to-real transfer through a complementary approach: rather than adapting predictions to target distributions, we generate diverse synthetic training domains that improve robustness across unknown target conditions.

We tackle this challenge and introduce an activity domain generation framework which creates novel ADL appearances (*novel* domains) from different existing activity modalities (*source* domains) inferred from video training data. Our framework computes human poses, heatmaps of body joints, and optical flow maps and uses them alongside the original RGB videos to learn the essence of source domains in order to generate completely new

Enhanced Synthetic Training Data: True and Newly Created Domains



Application: *real* domestic environments (evaluation on *real* videos)

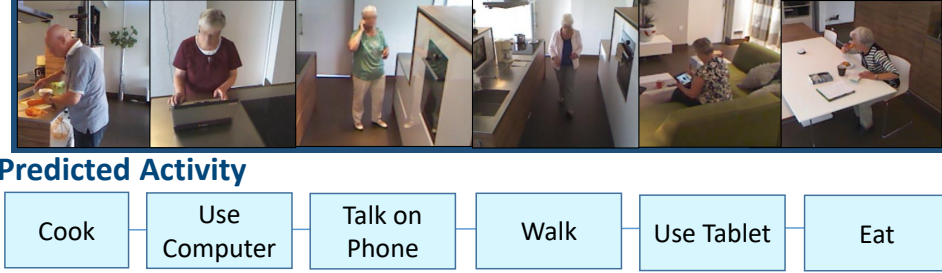


Figure 11: An overview of the proposed SYNTHETIC→REAL activity recognition pipeline improved through neural network-based generation of novel activity domains. In addition to synthetic RGB videos, the training data is enriched with modalities explicitly extracted from videos (*source* modalities) and action representations from the *novel* modalities generated with our model. Such multimodal diversification of the training samples significantly mitigates adverse effects of the SYNTHETIC→REAL domain shift.

ADL domains. The model is optimized by *maximizing* the distance between the existing source appearances and the generated novel appearances while ensuring that the semantics of an activity is preserved through an additional classification loss. Multimodality is central to this design, yet the framework requires only single RGB videos as input. All source modalities (pose heatmaps, limbs, optical flow) are automatically extracted from RGB, eliminating the need for multi-sensor hardware setups while retaining the benefits of multimodal representation learning.

The newly created activity domains are then integrated in the training of the ADL classification networks, resulting in models far less susceptible to changes in data distributions. Extensive experiments on the SYNTHETIC→REAL benchmark Sims4Action demonstrate the potential of the domain generation paradigm for cross-domain ADL recognition, setting new state-of-the-art results.

3.3.1 Motivation

When roboticists apply activity recognition models in practice, they will quickly discover the problem of domain shifts. In fact, a model is rarely deployed under conditions identical to the ones in the training set, as we face changes in illumination, camera type and -placement [111, 285]. One domain change vital in robotic ADL assistance is the transition from synthetic to real data, as simulations ease the burden of data collection and privacy

concerns in domestic environments [109, 111, 137, 207]. In the light of the ageing population, domain-invariant recognition of Activities of Daily Living (ADL) is an important element for household robot perception and human-tailored planning [41, 160].

Recently, the paradigm of novel domain generation has been proposed as a way of extensive data augmentation for image recognition, leading to more domain-invariant models, e.g., for digit classification [305]. At the same time, activity analysis allows a plethora of true modalities based on human movement. Apart from raw RGB videos, modalities such as optical flow or body poses can be automatically extracted and used as *source* domains for learning to generate more diverse and activity-specific *novel* domains. Despite its high potential for mitigating domain shifts, such domain generation has been overlooked in activity analysis and is therefore the main motivation of our work.

We aim to make a step towards ADL recognition less susceptible to changes in data distribution and introduce a generative framework enriching the training data through generation of previously unseen activity domains. In our approach, multimodal action representations derived, e.g., from body pose, joint heatmaps and optical flow, are used to learn creating novel activity domains by *maximizing* the distance between the existing modalities (*source* domains) and the generated appearances (*novel* domains) while ensuring that the semantics of an activity are preserved through an additional *Classification Loss*. The newly generated novel modalities are then mixed with the initial source modalities and constitute a diversified version of the training set (see overview in Figure 11).

Multimodality is a central concept in our framework and we believe that the novel domain generation paradigm suits activity analysis, as body pose and movement dynamics enable a wider range of existing source modalities, which in return encourages diversity of the generated novel domains. Nevertheless, our framework does not rely on multi-sensor setups since all source modalities are inferred from RGB videos only.

The problem of domain shift on the SYN→REAL benchmark is significant, as established in Section 3.1. The current state-of-the-art model is trained on the Sims4Action [207] dataset and experiences substantial performance degradation when evaluated on real data. We conduct extensive experiments on the SYNTHETIC→REAL ADL recognition benchmark [207] and make two observations. Firstly, multimodality itself is beneficial for cross-domain ADL recognition and using body pose related data representations leads to a significant performance boost. Secondly, our idea of generating novel ADL appearances for training data enhancement consistently improves recognition results. Our framework outperforms pre-existing results, demonstrating the potential of multimodal domain generation for human activity analysis without any pre-training on real data.

Note on our Terminology. Similarly to [276, 304], we define a *domain* as a joint distribution P_{XY} over a feature space $\mathcal{X}$ and a label space $\mathcal{Y}$. In our work, we extract multiple *modalities* from synthetic data and learn to generate novel *modalities*. Since each modality occupies a distinct feature subspace $\mathcal{X}'$ and exhibits a unique appearance, we treat each modality as a distinct domain for the purposes of domain generation. Note that this domain-as-modality usage (e.g., RGB domain, flow domain) is distinct from the broader synthetic-to-real domain shift (Sims4Action→Toyota Smarthome), which refers to the distribution shift between data sources. Throughout this section, when we refer to “novel domains” or “source domains,” we mean distinct modalities unless otherwise specified.

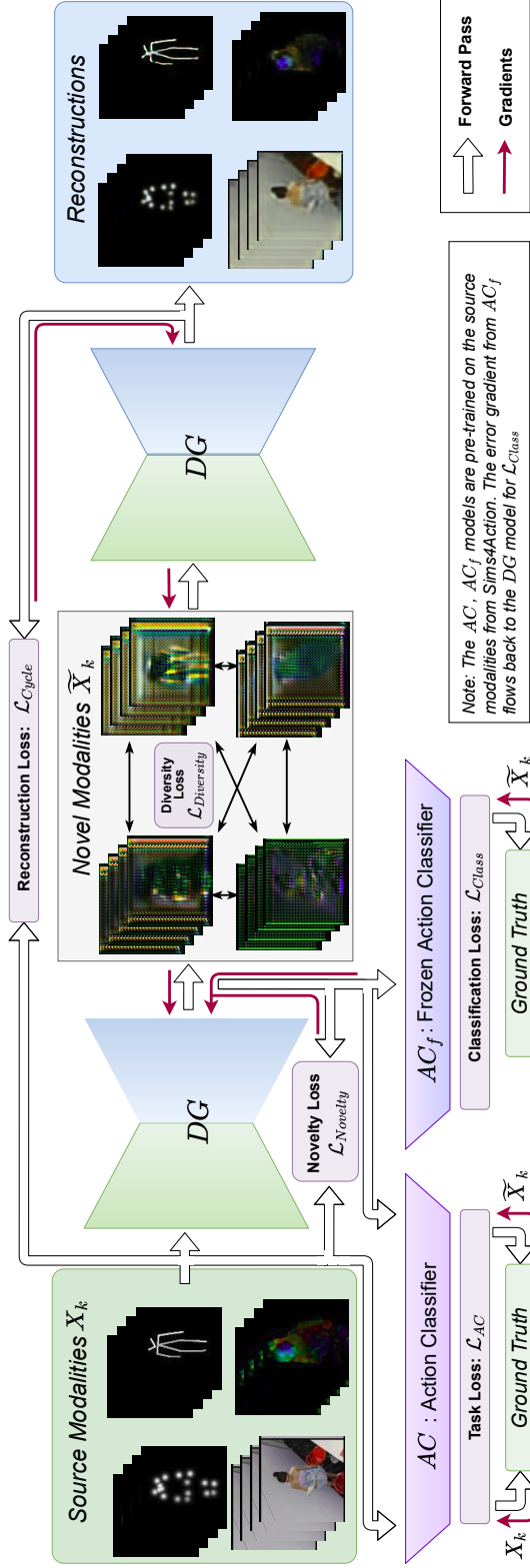


Figure 12: Architecture of our multimodal framework for generation of novel ADL modalities. First, multiple modalities related to the body pose and movement are extracted and used alongside the original RGB videos as our source modalities X_k . Activity examples in these source modalities are then fed into the *Domain Generator (DG)* as single images in order to generate novel modalities. The distribution divergence between each source-novel domain pair is estimated via the *Novelty Loss*, whereas the *Diversity Loss* computes the divergence for each novel-novel domain pair. The source domains and the generated novel domains are both used as inputs to the *Action Classifier*, which makes separate predictions for both. The same generator is used to reconstruct the original input and the *Reconstruction Loss* is computed. The error signals for training our model via backpropagation are depicted as red arrows.

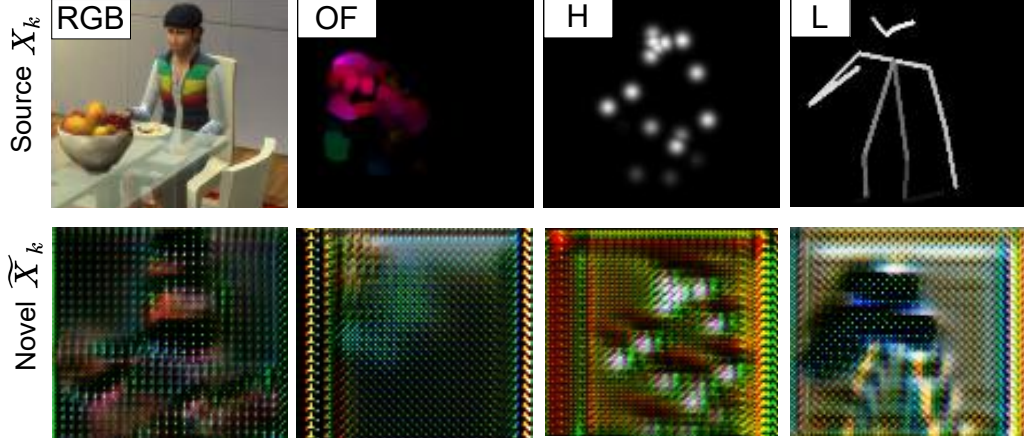


Figure 13: Examples of the *source* X_k (top row) and *novel* modalities $\tilde{X}_k$ (bottom row). H: Heatmaps, L: Limbs, OF: Optical Flow. We use all 8 modalities $X_k \cup \tilde{X}_k$ for training and evaluate our models on REAL data.

3.3.2 Multimodal Framework for Activity Domain Generation

In this section, we introduce a multimodal framework for better domain generalization in activity recognition, aiming to lighten the impact of appearance changes when moving from synthetic ADL training data to real-world robotic applications. We follow the domain generation paradigm [305] recently proposed for image recognition, extending it to the scope of human activity analysis (Figure 12).

Conceptually, our framework produces *novel* modalities $\tilde{X}_k$ by learning to transform given *source* modalities X_k and comprises four main building blocks: 1) A modality extraction module used to compute multiple source modalities from the original RGB video 2) a pre-trained domain classifier DC , 3) an action classifier AC , 4) and a domain generator DG . As we are using each extracted modality as a source *domain* we refer to the source and novel domains as source and novel *modalities*.

The task we address is SYNTHETIC→REAL domain generalization. We train an action classifier AC on samples from SYNTHETIC domains $x_s \in X_k \cup \tilde{X}_k$ with action labels $y \in Y$. In domain generalization, training and test data originate from distinct probability distributions, in our case $x_s \sim p_{synthetic}$ and $x_r \sim p_{real}$, and test samples x_r neither have labels, nor are seen during training. Our goal is to classify each instance x_r in the REAL target test domain X_r , which has a shared action label set Y with the training set. For this, we utilize the synthetic Sims4Action [207] dataset for training and the real Toyota Smarthome [52] and ETRI-Activity3D-LivingLab (ETRI) [112] as two separate test sets.

3.3.2.1 Extracting Source Modalities

The nature of activity recognition and video data in general allows us to leverage a wide range of modalities, such as body pose and movement dynamics, which would not be applicable in conventional image classification. We utilize four source modalities $X_k, k \in \{0, 1, 2, 3\}$ which are extracted directly from the training data (*i.e.*, RGB videos).

The source modalities consist of 1) heatmaps of the body joint locations, 2) limbs connecting the joints as lines, 3) dense optical flow extracted between each two frames, and

4) raw RGB images (see top row of Figure 13). The heatmaps and limbs are extracted using the AlphaPose [73, 139, 289] pose detector, which infers 17 joint locations. The heatmaps modality $h(x, y)$ at pixel (x, y) is obtained by applying 2D gaussian maps, centered at each joint location (x_i, y_i) and weighted by its confidence c_i as seen in Equation 1.

$$h(x, y) = \exp\left(\frac{-((x - x_i)^2 + (y - y_i)^2)}{2\sigma^2}\right) \cdot c_i \quad (1)$$

The limbs domain is composed by connecting the joints with white lines and weighting each line by the smaller confidence of its endpoints. The optical flow is estimated using the Gunner-Farneback method [74]. We refer to these modalities as the four *source* modalities X_k (see top row in Figure 13).

3.3.2.2 Domain Classifier

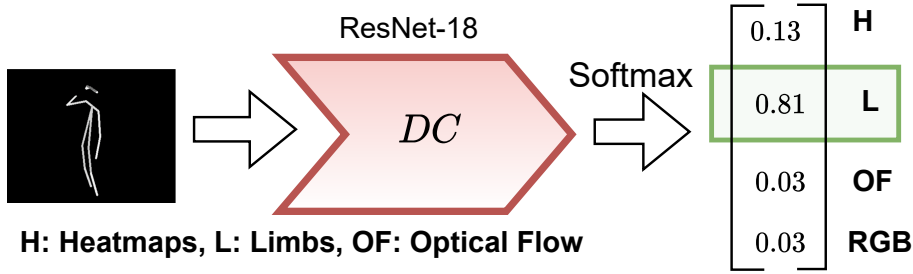


Figure 14: Domain Classifier (DC): Overview of inference and training. The DC model is pre-trained end-to-end on Sims4Action.

The domain classifier DC is trained on the synthetic Sims4Action [207] to classify the source modalities X_k with labels $k \in \{0, 1, 2, 3\}$ as shown in Figure 14. Afterwards, its weights are frozen and it is utilized for training the domain generator DG . The frozen DC is used to obtain embeddings from the source X_k and novel modalities $\tilde{X}_k$ as seen in Figure 15. The Sinkhorn distance [46] between the embeddings is utilized as a distribution divergence metric and it is used to compute the *Novelty Loss* for each source-novel modality pair $(X_k, \tilde{X}_k)$ and the *Diversity Loss* for all novel-novel pairs $(\tilde{X}_k, \tilde{X}_l)_{k \neq l}$ (see Equations 2, 3). The error gradient is propagated back to the domain generator DG and conditions it to produce novel modalities, which are both diverse and different from the source modalities.

3.3.2.3 Domain Generator

The goal of the domain generator DG is to extend and diversify the synthetic training data from Sims4Action [207]. The DG model is trained on the four source modalities X_k to generate four novel modalities $DG(X_k) = \tilde{X}_k$, which should be as diverse as possible, while remaining semantically and structurally consistent. The diversity of the new modalities is enforced by the *Novelty* and *Diversity* loss terms $\mathcal{L}_{Novelty}$ and $\mathcal{L}_{Diversity}$. The *Novelty Loss* maximizes the distribution divergence between the source and novel modalities, while the *Diversity Loss* maximizes the distribution divergence between each pair of generated novel modalities. The divergence measure we use is the Sinkhorn distance [46] $d(\cdot, \cdot)$ between the embeddings obtained by the domain classifier and is computed as illustrated

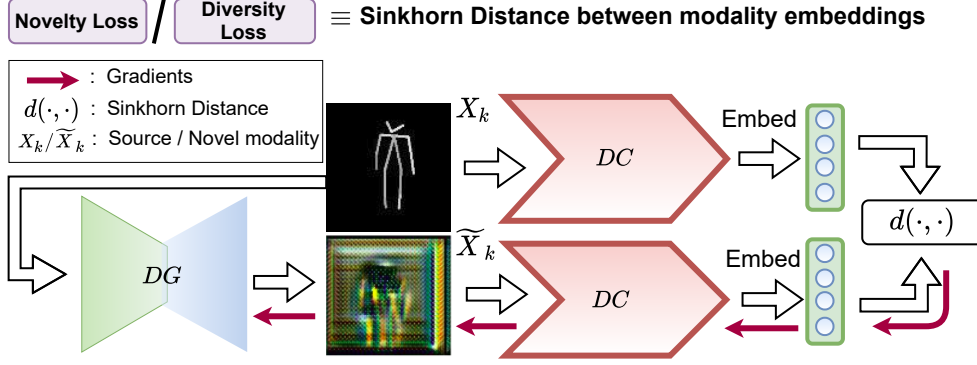


Figure 15: Domain Classifier: Computation of the *Novelty Loss* for all source-novel modality pairs $(X_k, \tilde{X}_k)$. Note that the *Diversity Loss* is computed the same way with all novel-novel pairs $(\tilde{X}_k, \tilde{X}_l)_{k \neq l}$.

in Figure 15. To ensure that the new modalities are dissimilar to the source modalities and are also dissimilar to each other these two loss terms are maximized w.r.t. DG as shown in Equations 2 and 3.

$$\mathcal{L}_{Novelty} = \max_{DG} \sum_{k=0}^3 d(\tilde{X}_k, X_k) \quad (2)$$

$$\mathcal{L}_{Diversity} = \max_{DG} \sum_{k=0}^3 \sum_{l=0}^3 d(\tilde{X}_k, \tilde{X}_l) \quad (3)$$

where $k, l \in \{0, 1, 2, 3\}$, $k \neq l$, and $\tilde{X}_k = DG(X_k)$.

Furthermore, the domain generator is conditioned to preserve the semantic and structural consistency of the actions in the novel modalities by minimizing the action *Classification Loss* $\mathcal{L}_{Class}$ and the *Reconstruction Loss* $\mathcal{L}_{Cycle}$.

$$\mathcal{L}_{Class} = \min_{DG} \sum_{k=0}^3 \mathcal{L}_{CE}(AC_f(\tilde{X}_k), Y_k) \quad (4)$$

$$\mathcal{L}_{Cycle} = \min_{DG} \sum_{k=0}^3 \|DG(DG(X_k)), X_k\|_1 \quad (5)$$

where Y_k is the ground-truth action label, $\mathcal{L}_{CE}$ is the cross-entropy loss, and AC_f is the frozen action classifier pre-trained on Sims4Action’s [207] source modalities X_k (explained in Section 3.3.2.6). Note that we train on mini-batches. For this reason, $X_k, \tilde{X}_k, Y_k$ refer to individual batches of size B in all equations, rather than the whole dataset.

3.3.2.4 Action Classifier

The action classifier AC is firstly pre-trained only on the source modalities X_k and then trained further on both X_k and $\tilde{X}_k$ to assign the correct activity label. This encourages the model to learn representations in both the source and novel modalities and results in a larger and a more heterogeneous training set. The novel modalities increase the diversity

of the training samples as their distribution differences are maximized by $\mathcal{L}_{Novelty}$ and $\mathcal{L}_{Diversity}$, which leads to a more versatile training dataset. The generation of $\tilde{X}_k$ can hence be viewed as a method for data augmentation.

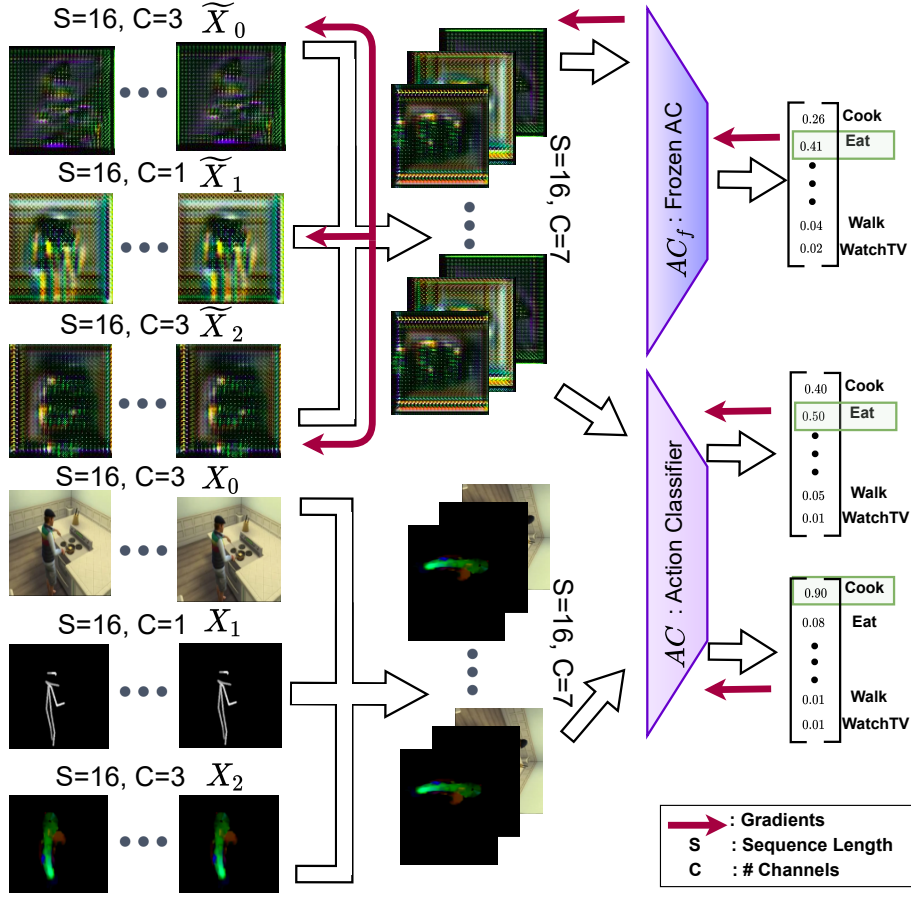


Figure 16: Example of AC training with three modalities ($k=2$). The input sequences are concatenated along the channel dimension C . The action classifier AC is trained on both the source X_k and novel $\tilde{X}_k$ modalities for action recognition and makes separate predictions for both. Note that the input to the AC model is sequence-based ($S=16$) as opposed to the image-based DC and DG ($S=1$). Its frozen copy AC_f is pre-trained on the source modalities from Sims4Action and is only used for the computation of $\mathcal{L}_{Class}$ in Equation 4. The error gradient from AC_f flows back to the novel modalities as its prediction is used to train DG with $\mathcal{L}_{Class}$ (Eq. 4).

3.3.2.5 Model Architectures

As our action classifier AC , we utilize a 3D-CNN model to exploit the temporal features in the videos, which are important for action recognition [100, 255, 265]. 3D-CNN models such as the Inflated 3D-ConvNet (I3D) [30] have shown remarkable performance in recognition for ADL [30, 52]. Our action classification 3D-CNN model is implemented with the Separable 3D architecture (S3D) [288]. The S3D model leverages separable convolutions and replaces most of the 3D convolutions of I3D with cheaper 2D convolutions to reduce the complexity while maintaining the performance of I3D. For the domain classifier DC we employ ResNet18 [102] to learn to classify all source modalities (see Figure 14). Lastly, the domain generator DG , which also operates on single images, is modeled after Zhou

et al. [305] and consists of two down-sampling conv-layers, two residual blocks [102] with instance normalization [261] and two transposed conv-layers to up-sample back to the input's size.

3.3.2.6 Training Procedure

All of our models are trained *only* on Sims4Action [207]. The AC model is evaluated in the end on the real Toyota Smarthome [52] and ETRI [112] datasets to test its capability for domain generalization. The rest of the models are used only for the computation of loss terms - AC_f , DC , or for enhancing AC's training data - DG . The domain classifier DC is trained on Sims4Action to distinguish the source modalities X_k as seen in Figure 14. Then the DC model is frozen and used solely for the computation of $\mathcal{L}_{Novelty}$ and $\mathcal{L}_{Diversity}$ as in Figure 15 for the rest of the training.

Before we jointly train the other two models AC and DG , we pre-train an action classifier on all source modalities X_k . We then produce two copies of the pre-trained action classifier - the first copy AC_f is frozen and used only to enforce the semantic consistency of the domain generator in $\mathcal{L}_{Class}$ (see Equation 4). The second copy AC is further fine-tuned on both the source X_k and novel modalities $\tilde{X}_k$ with the cross-entropy loss. Only the AC model is evaluated in the end on the real Toyota Smarthome and ETRI datasets.

LOSS COMPUTATION In each iteration, a sequence of images is sampled from random chunks of the training videos from all source modalities X_k . The modality images from the sequence are concatenated along the channel dimension and are used to train the AC model. Each individual image from the sequence is transformed by DG into the novel modalities $\tilde{X}_k$, which are reshaped into image sequences (see Figure 16 top-left). The novel modality sequences are again concatenated along the channels and fed to AC and AC_f . The weights of AC_f are not updated, but its error is propagated back to the domain generator to compute $\mathcal{L}_{Class}$ as seen in the red arrows in Figure 16. The *Novelty* and *Diversity* loss terms are computed with the help of the frozen DC model and the Sinkhorn distance [46] (see Figure 15). Finally the *Reconstruction Loss* is computed by iterating over all source and novel modality images and applying Equation 5. The final loss function $\mathcal{L}_{DG}$ for the DG model is:

$$\mathcal{L}_{DG} = \min_{DG} \lambda_c \mathcal{L}_{Class} + \lambda_r \mathcal{L}_{Cycle} - \lambda_d (\mathcal{L}_{Novelty} + \mathcal{L}_{Diversity}) \quad (6)$$

where $\lambda_c, \lambda_r, \lambda_d$ are balancing parameters for the loss terms.

The AC model is trained on both the source and novel modality sequences. Hence, the *Task Loss* $\mathcal{L}_{AC}$ is:

$$\mathcal{L}_{AC} = \min_{AC} \sum_{k=0}^3 \alpha \mathcal{L}_{CE}(AC(X_k), Y_k) + (1 - \alpha) \mathcal{L}_{CE}(AC(\tilde{X}_k), Y_k) \quad (7)$$

where α balances training on source X_k and novel modalities $\tilde{X}_k$ and the other terms follow the notation of Equation 4.

3.3.2.7 Evaluation Procedure

During test-time, we use the action classifier AC to perform action recognition on the real test datasets. We extract the pose and optical flow modalities from the real test samples

so that they match the input format of the AC network. The model is evaluated using balanced accuracy (mean per-class accuracy) as the primary metric, since real-world datasets have unbalanced class distributions. We also report unbalanced accuracy for consistency with prior work [207], though this metric is biased towards overrepresented categories.

3.3.3 Experimental Results

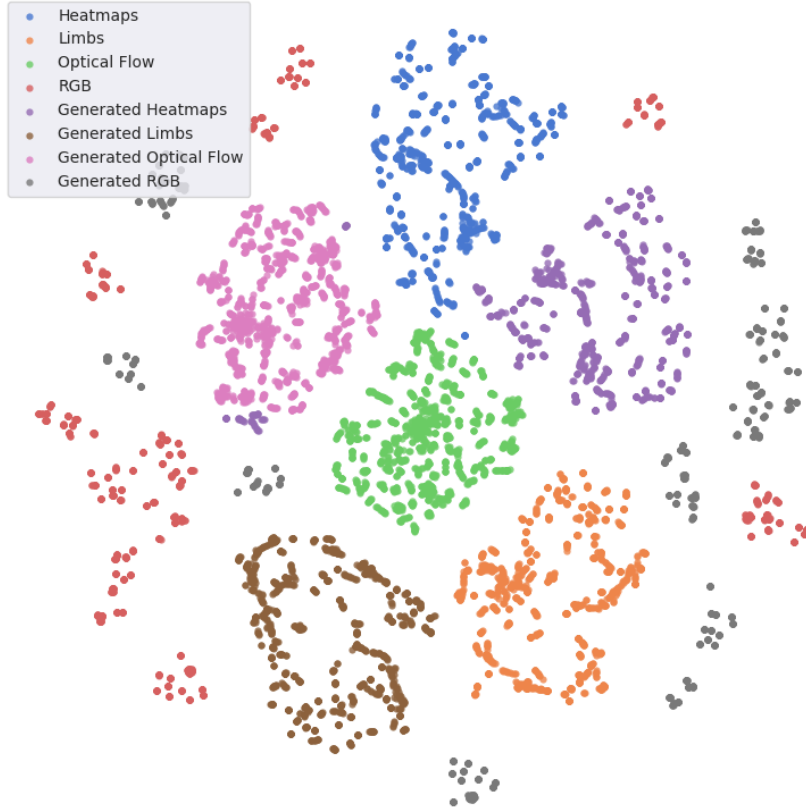


Figure 17: t-SNE visualization of the domain generator’s bottleneck layer embeddings. Each modality (both source and novel) forms distinct clusters, confirming that the generator has learned to create diverse, non-overlapping modality distributions.

3.3.3.1 Benchmark Details

We focus on domain generalization between simulated and real data using Sims4Action [207] as a training dataset. For testing, we use real data, which originates from the Toyota Smarthome [52] and ETRI [112] datasets. Sims4Action provides ten hours of synthetic video material recorded from the computer game Sims 4, covering 10 basic human actions of daily living which have direct correspondences in the two real datasets.

Toyota Smarthome [52] contains video material of 18 subjects performing 31 unscripted activities of daily living within a single apartment, and ETRI [112] is composed of 50

Table 9: Comparison of our model to state-of-the-art results on the SYNTHETIC→REAL benchmark [207] and to three additional domain generalization methods - TA³N [35], APN [293], and VideoDG [293].

Approach	Pre-training	Toyota SmartHome		Etri Activity	
		Balanced Accuracy	Unbalanced Accuracy	Balanced Accuracy	Unbalanced Accuracy
Random Choice	None	10.00	10.00	10.00	10.00
Currently reported state-of-the-art results on the SYNTHETIC→REAL benchmark					
S3D [30, 207]	None	12.40	19.95	11.71	13.86
S3D [30, 207]	Kinetics-400 [123]	23.25	22.75	23.45	28.57
Other domain generalization methods on the SYNTHETIC→REAL benchmark					
TA ³ N [35]	ImageNet [59]	14.19	14.44	25.11	35.12
APN [293]	ImageNet [59]	22.09	17.19	27.97	37.67
VideoDG [293]	ImageNet [59]	25.71	21.12	25.55	41.09
Our model with domain generation (best modality combination)					
Ours	None	27.73	32.20	29.05	41.88

subjects performing 55 actions recorded from various perspectives representative of robot placements in home environments. However, we use only the 10 action correspondences to Sims4Action from both real datasets for our evaluation. We use the official cross-subject test split from Toyota SmartHome [47, 52], in which 7 of the subjects are reserved for testing. The same protocol is adopted in the state-of-the-art results [207] on the SYNTHETIC→REAL benchmark. For ETRI, we use the unsupervised domain adaptation protocol from [125] and use the whole dataset for testing.

As our evaluation data was collected in real households [52, 112], the number of samples per activity class is unbalanced and follows a Zipf-like distribution rather than a uniform one. We therefore focus on the *balanced accuracy* (mean per-class accuracy) as our main evaluation metric and additionally report the *unbalanced accuracy* (correct prediction rate among the complete test set). Note, that while we also evaluate the unbalanced accuracy to be consistent with [207], the metric is highly biased towards overrepresented categories in the test set and should be taken with caution especially in real-life datasets, where the categories are not evenly distributed.

3.3.3.2 Qualitative Analysis

First, we inspect the domain generation results produced by the domain generator *DG* provided in Figure 13. The model has learned to produce novel appearances that differ substantially from photorealistic rendering (Figure 13, bottom row). This demonstrates a key insight: visual realism is not a prerequisite for cross-domain utility. Despite their unconventional appearance, these generated modalities improve recognition on real-world data (Table 10), establishing that distributional diversity—rather than appearance fidelity—drives domain generalization. The activity semantics are preserved through the classification loss $\mathcal{L}_{Class}$, ensuring that motion patterns remain recognizable even as vi-

sual appearance diverges from source modalities. Note that the movement typical for the activity is more obvious in video results than in still images.

To better understand the learned representations, we visualize the output of the domain generator’s bottleneck layer with the t-distributed stochastic neighbor embeddings (tSNE) [153] in Figure 17. The embeddings are produced by sampling 10 random frames from each video from the Sims4Action dataset for each of the source modalities. Each image sample is fed to the domain generator and its embedding is used for the t-SNE visualization. The DG model is then applied again to the novel modalities $\tilde{X}_k$ to obtain their embeddings. The plot in Figure 17 shows that each modality forms an individual cluster of points. This confirms that the generator has learned features which help differentiate between all 8 modalities $X_k \cup \tilde{X}_k$. Furthermore, the modalities span a wider and non-overlapping distribution in the latent space, i.e. the training dataset has been diversified by adding the novel modalities $\tilde{X}_k$, which leads to better domain generalization.

3.3.3.3 Quantitative Results

In Table 10, we compare the ACs trained only on the source modalities X_k to the ones which are trained including the generated novel modalities $\tilde{X}_k$. All modalities $X_k \cup \tilde{X}_k$ are produced from Sims4Action [207] only. Additionally, we consider all 15 modality combinations and evaluate all models on the real Toyota Smarthome [52] and ETRI [112] datasets via the balanced and unbalanced accuracy metrics.

For most combinations, the extension of the training dataset with novel modalities increases the balanced accuracy significantly, in some cases by up to 13 percentage points (ETRI: L). This is especially prevalent in multi-modal settings joined over early fusion, and we hypothesize that the generation of additional data alleviates overfitting problems which arise when training a larger, multi-modal model. Optical flow (OF) profits the least from this technique when combined with other modalities; in some cases, it performs worse than the baseline. This might be explained by OF’s low accuracy on the SYNTHETIC→SYNTHETIC benchmark (44%), i.e. OF is a weak modality on Sims4Action. However, performance losses in these cases are small in comparison to the significant gains which are achieved by the other combinations. We do list unbalanced accuracy as well, which provides similar results. Note that the improvement depends on the modality combination and test dataset. However, the average improvement for the balanced accuracy for Toyota Smarthome/ETRI is +3.1%/+2.8% and +4.7%/+8.0% for unbalanced accuracy.

These consistent improvements across both test datasets demonstrate that learned domain generation improves cross-domain robustness without requiring target domain access or real-world pretraining. The gains validate the domain generation paradigm for activity recognition, establishing that synthetic modality augmentation provides a complementary approach to target-specific adaptation methods (Section 3.2). While adaptation techniques calibrate predictions to known target distributions, domain generation diversifies training data to improve robustness across unknown target domains, making it particularly suitable when target characteristics are unavailable during training.

Beyond validating the domain generation paradigm, these experiments provide a systematic multimodal analysis of synthetic-to-real transfer for activity recognition. By evaluating all 15 combinations of source modalities (RGB, heatmaps, limbs, optical flow) both with and without novel domain augmentation, we establish which representations support cross-domain transfer most effectively. Pose-based representations (heatmaps, limbs)

achieve stronger cross-domain transfer than appearance-based modalities, with limbs reaching 27.7% balanced accuracy on Toyota Smarthome and 29.1% on ETRI. The multi-modal analysis reveals that early fusion benefits substantially from domain augmentation, with particularly large improvements observed for limbs on ETRI (+13.0%). This suggests that generated domains particularly alleviate overfitting in larger multimodal models, where additional training variation proves most beneficial.

3.3.3.4 *Comparison to State-of-the-art and Other Approaches*

In Table 9 we compare the best results achieved by our approach (*i.e.*, the best modality variants from Table 10 marked in bold), and compare them with current state-of-the-art results on the Sims4Action SYNTHETIC→REAL benchmark [207]. We also compare our method to three other domain generalization methods: TA³N [35], APN [293], and VideoDG [293]. To this end, we train all three models end-to-end on Sims4Action [207] and follow the same evaluation protocols for Toyota Smarthome [52] and ETRI [112], which we used for our approach in Section 3.3.3.3.

Despite not making use of real-world pretraining, we improve upon the previous state-of-the-art on the Sims4Action→Real benchmark [207], which achieved 23.25% balanced accuracy on Toyota Smarthome using S3D with Kinetics-400 pretraining. Our approach reaches 27.73% balanced accuracy and 29.05% on ETRI, establishing new state-of-the-art results through synthetic domain augmentation alone. Our method also consistently outperforms the other domain generalization methods TA³N [35], APN [293], and VideoDG [293]. APN and VideoDG also outperform the previous state-of-the-art on some metrics, which supports the claim that a domain generalization strategy for the SYNTHETIC→REAL benchmark is beneficial.

		Balanced Accuracy		Unbalanced Accuracy		SYNTHETIC→REAL		Balanced Accuracy		Unbalanced Accuracy		Bal. Acc.	Unbal. Acc.
		Toyota Smartthome [52]				ETRI [112]				Sims4Action [207]			
Test Set	Train Set	Only X_k	Ours: $X_k \cup \tilde{X}_k$	Only X_k	Ours: $X_k \cup \tilde{X}_k$	Only X_k	Ours: $X_k \cup \tilde{X}_k$	Only X_k	Ours: $X_k \cup \tilde{X}_k$	Only X_k	Ours: $X_k \cup \tilde{X}_k$	Only X_k	
	Input												
Individual													
	RGB	13.7	13.3 (-0.4)	18.5	17.6 (-0.9)	11.7	15.0 (+3.3)	13.9	15.6 (+1.7)	61.8	59.4		
	Heatmaps (H)	20.2	25.8 (+5.6)	20.0	23.7 (+3.7)	15.2	17.9 (+2.7)	22.2	40.9 (+18.7)	71.4	70.4		
	Limbs (L)	22.0	27.7 (+5.7)	21.9	21.7 (-0.2)	16.1	29.1 (+13.0)	16.2	38.7 (+22.5)	75.1	74.4		
	Optical Flow (OF)	21.3	22.6 (+1.3)	31.5	32.2 (+0.7)	11.6	14.7 (+3.1)	13.5	17.5 (+4.0)	44.5	43.7		
Early Fusion by Channel Concatenation													
	RGB + H	10.3	20.7 (+10.4)	6.1	22.8 (+16.7)	7.3	12.9 (+5.6)	7.6	14.8 (+7.2)	76.8	77.2		
	RGB + L	13.5	18.5 (+5.0)	8.2	18.6 (+10.4)	13.4	15.7 (+2.3)	15.9	23.5 (+7.6)	81.5	81.0		
	RGB + OF	19.4	17.0 (-2.4)	31.0	30.4 (-0.6)	14.6	14.1 (-0.5)	20.9	41.9 (+21.0)	70.4	72.3		
	H + L	15.7	25.3 (+9.6)	15.7	20.2 (+4.5)	15.0	17.8 (+2.8)	16.5	17.9 (+1.4)	57.8	50.1		
	H + OF	16.7	23.7 (+7.0)	20.0	24.6 (+4.6)	12.6	14.2 (+1.6)	15.7	17.9 (+2.2)	71.1	64.7		
	L + OF	19.7	19.8 (+0.1)	24.8	22.1 (-2.7)	11.4	12.5 (+1.1)	13.4	16.7 (+3.3)	72.7	71.2		
	RGB + H + L	15.3	13.9 (-1.4)	10.0	23.2 (+13.2)	5.8	12.2 (+6.4)	5.8	13.8 (+8.0)	73.7	73.1		
	RGB + H + OF	11.6	13.4 (+1.8)	13.0	29.5 (+16.5)	12.8	14.2 (+1.4)	12.5	22.3 (+9.8)	70.7	73.1		
	RGB + L + OF	12.0	14.1 (+2.1)	15.0	13.5 (-1.5)	11.5	10.1 (-1.4)	15.9	17.8 (+1.9)	83.0	79.8		
	H + L + OF	25.1	20.7 (-4.4)	36.8	22.9 (-13.9)	15.2	12.5 (-2.7)	13.9	15.6 (+1.7)	77.1	72.9		
	RGB + H + L + OF	12.1	18.6 (+6.5)	7.0	27.4 (+20.4)	10.3	13.0 (+2.7)	9.1	18.9 (+9.8)	66.8	65.1		

Table 10: Evaluation results for all 15 modality combinations on the SYNTHETIC→REAL benchmark for the real Toyota Smarthome [52] and ETRI [112] datasets and on the SYNTHETIC→SYNTHETIC benchmark for Sims4Action [207]. Our method of generating novel modalities $\tilde{X}_k$ for the training data leads to a significant improvement of the vast majority of the models on the SYNTHETIC→REAL benchmark. The performance on Sims4Action is listed to illustrate the domain gap. The performance boost by adding novel modalities $\tilde{X}_k$ to the training data is indicated in brackets for each modality combination.

3.4 CHAPTER CONCLUSION

Systematic evaluation requires controlled benchmarks. Sims4Action pairs gaming-based synthetic data from The Sims 4 with real-world evaluation on Toyota Smarthome under standardized protocols, supporting systematic comparisons of transfer methods and serving as a domain adaptation testbed.

Mitigation relied on two complementary strategies. Adaptation methods explicitly align synthetic and real distributions: OT-DAKD uses optimal-transport-based prediction calibration with pose-based representations, assuming a known target distribution and improving Sims4Action performance. Generalization methods improve robustness without target-domain access: MMGen evaluates multiple modality combinations and shows that distributional diversity drives cross-domain transfer.

Despite this progress, the limited controllability of gaming-based generation motivates more involved simulator-based approaches. The next chapter examines parametric body models, which offer two advantages over gaming platforms: (1) controlled rendering with systematic variation in pose, appearance, and viewpoint, and (2) applicability for biomechanical simulation coupling visual observations with physiological signals such as muscle activations, enabling research on motion beyond what cameras capture.

3.4.1 Contributions to Cross-Domain Synthetic Data Transfer

Contribution 1: To evaluate synthetic-to-real transfer characteristics of deep neural networks for the recognition of activities of daily living, we establish *Sims4Action*, pairing life simulation game-based synthetic data from The Sims 4 with real-world evaluation on Toyota Smarthome. Sims4Action quantifies domain gaps and provides standardized protocols for cross-domain methods, serving as a domain-adaptation and generalization testbed.

Contribution 2: To bridge synthetic→real gaps identified by Sims4Action, we introduce *Optimal Transport-based Domain Adaptive Knowledge Distillation (OT-DAKD)*, combining pose-aligned features with knowledge distillation. OT-DAKD shows that target marginal constrained domain adaptation can mitigate appearance and motion gaps between synthetic training and real evaluation.

Contribution 3: To improve cross-domain robustness without target domain knowledge, we introduce *MMGen*, a framework for systematic multimodal domain generation. Building on findings from Section 3.2 that pose-based representations facilitate transfer, MMGen shows that modality diversity can drive cross-domain robustness. By generating novel training modalities through adversarial optimization while preserving semantic content, MMGen provides an effective alternative when target characteristics are unknown.

SYNTHETIC DATA FROM 3D HUMAN BODY MODELS

The previous chapter addressed synthetic-to-real domain gaps through benchmarking, adaptation, and generalization. This chapter explores a different approach, i. e. the use of parametric human body models for controllable synthetic data generation. The SMPL (Skinned Multi-Person Linear) model [150] and its extensions provide a unified representation which connects visual observations to biomechanical understanding. These models enable a bidirectional mapping: neural networks estimate SMPL parameters from video, whereas the same parameters regenerate synthetic humans which display identical motion. This chapter examines two questions: (1) *How can parametric body models enable large-scale synthetic data generation, and does pre-training on such data improve real-world action recognition?* (2) *Can parametric body models combined with biomechanical simulation software enable muscle activation datasets at scale, and which architectures effectively learn pose-to-muscle activation mappings?* Both works build on motion sequences from the AMASS dataset [155] and thereby demonstrate different applications of parametric body models, across visual and biomechanical simulation. Section 4.1, based on our ICRA 2024 publication [219], addresses the first question through SynthAct, a Unity-based renderer which generates large-scale multi-view synthetic data. Section 4.2, based on our NeurIPS 2024 publication [221], addresses the second question through Muscles in Time (MinT), which combines parametric body models with OpenSim biomechanical simulation for muscle activation data generation.

Parametric human body models present a structured representation of human shape and pose through learned statistical models. The SMPL model [150] and its extensions (SMPL-H [211], SMPL-X [186]) represent human bodies through shape parameters (body proportions), pose parameters (joint rotations), and surface mesh vertices. This parametric formulation allows for a bidirectional mapping between visual observations and synthetic data generation. Modern neural networks estimate SMPL parameters from RGB or depth video, whereas rendering engines reconstruct photorealistic synthetic humans from these parameters. The same motion sequence can be observed through multiple modalities, e. g. visual appearance, skeletal pose, or depth maps, or it can be transferred to biomechanical models for physiological analysis. This dual capability turns SMPL-based models into a basis for scalable and controllable synthetic data generation.

The works presented in this chapter demonstrate two different applications of parametric body models for synthetic data generation. SynthAct makes use of body model rendering for large-scale multi-view synthetic data creation. By combining motion capture sequences from AMASS [155] with SMPL-based body rendering in Unity, the framework

generates synchronized multi-camera perspectives without requiring expensive multi-view recording infrastructure. This approach addresses a major challenge in multi-view learning: real-world multi-camera datasets remain scarce due to recording complexity, whereas synthetic multi-view data generation becomes straightforward through virtual camera placement. Muscles in Time extends body model synthesis to biomechanical simulation and works with the same AMASS motion sequences. SMPL's consistent surface correspondence enables mapping 3D points across frames. This allows for the transfer of motion data to OpenSim musculoskeletal models [58] for muscle activation simulation.

A unique advantage of body model synthesis emerges from using body motion as a cross-dataset anchor. The same motion sequence can link RGB video (from SynthAct), muscle activations (from MinT), depth maps, skeletal pose, and additional modalities. This creates potential for multi-modal learning without simultaneous capture of all modalities. While we focus on these two approaches individually, this shared basis opens possibilities for future research which combines visual, skeletal, and physiological modalities.

4.1 MOTION CAPTURE-BASED SYNTHETIC DATA GENERATION

This section is based on our paper published at ICRA 2024 [219]. © 2024 IEEE.

Can purpose-built synthetic data achieve state-of-the-art performance on real-world action recognition benchmarks? This section presents SynthAct, a versatile Unity-based synthetic generator, and AMARV, an 800,000-video dataset that demonstrates synthetic data can not only transfer to real-world scenarios but enable real-world performance improvements while addressing privacy concerns inherent in large-scale data collection.

In the field of computer vision and machine learning, the ethical implications of data collection are increasingly important. Large-scale datasets like Kinetics [30], often scraped from the web, have advanced the field of human action recognition but come with drawbacks. These datasets frequently include videos collected without the knowledge or consent of the individuals involved, a practice raising privacy and consent concerns. Particularly in domestic or human-robot interaction scenarios, as robotic systems become more common in household settings, the need for privacy preserving data collection becomes more urgent. Focusing exclusively on body pose sequences in the data collection process represents a significant advancement in this direction which mitigates privacy concerns. While pose-based action recognition methods make use of this paradigm, they cannot maintain the advantages of video based human action recognition.

Creating a synthetic dataset for human behavior is a challenging task due to the wide range of variables such as body types, appearances, clothing styles, and backgrounds, both indoor and outdoor, as well as different activities and motion styles. Existing databases often focus on specific use-cases [151, 207, 266] or offer a narrow range of categories [44, 56]. To address these limitations, we introduce a synthetic data generator for human action recognition. Developed using the Unity engine, SynthAct leverages body shape and movement models from SMPL-H [211] to create a diverse range of human appearances.

For large-scale pretraining, we leverage SynthAct to introduce the **Archive Of Motion Capture As Rendered Videos (AMARV)**, which is based on the AMASS [155] motion capture dataset. By incorporating state-of-the-art 3D pose estimation techniques, SynthAct can be applied to arbitrary human action datasets without the need for labeled ground

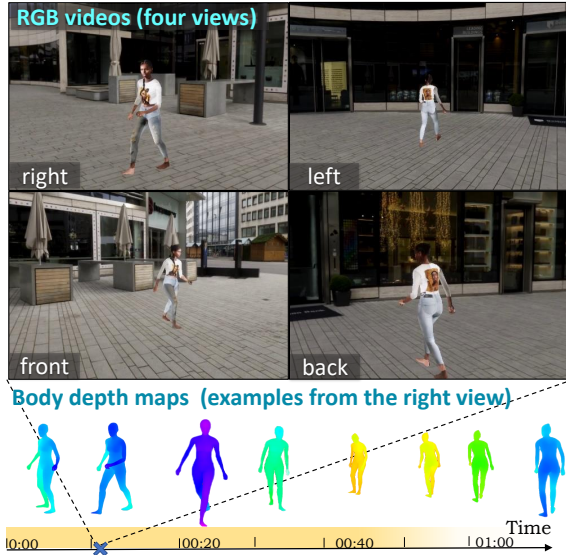


Figure 18: The four camera views of SynthAct (top) and corresponding depth maps (below).

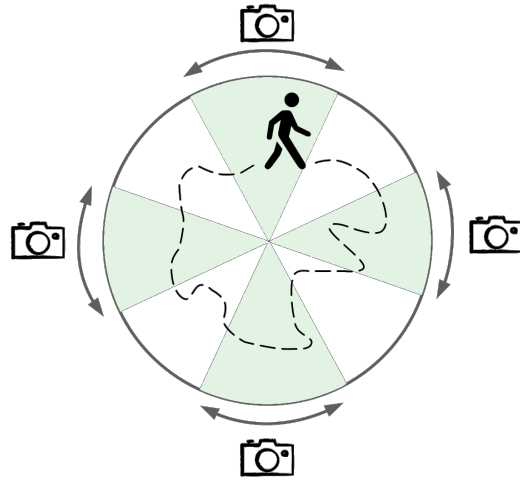


Figure 19: Random camera positions.

truth poses. To demonstrate the applicability of SynthAct to in-the-wild scenarios, we create the Synthetic Smarthome dataset as a counterpart to Toyota Smarthome [52], avoiding the use of ground truth poses.

4.1.1 Implementation

An overview of our Unity-based data generation pipeline is provided in Figure 20. In our experiments we use the publicly available AMASS database [155] in combination with BABEL labels [195] or pose sequences extracted with 4DHumans [89] from Toyota Smarthome [52] as motion source. Diverse clothing styles and textures are either adapted from Multi-Garment Net [16] or generated through stable diffusion [209] and adapted with Blender [43], and virtual humans are placed into heterogeneous 3D environments in form of license free HDRI images obtained from Poly Haven [191].

Motion diversity Using the AMASS dataset [155] as a source of action animation sequences for our data generator provides high quality motion capture animations. We limit ourselves to the 10,892 motion capture sequences which are provided with dense annotations by the BABEL language label dataset [195]. For Synthetic Smarthome we make use of 4DHumans, a 3D pose and shape estimation network. By estimating 3D pose from 2D video data, 4DHumans induces significant jitter, especially on the z-axis of the root node as well as false positive recognitions. We make use of simple thresholding and SLERP interpolation smoothing to restrict to higher quality motions.

Body model and animation AMASS provides its motions in form of the SMPL-H body model [211]. SMPL-H is a parametric model with Shape Blend Shapes, which enable the depiction of people with varying body proportions. In our dataset we specifically try to address discriminating biases and therefore actively cover a wide range of body shapes by randomly varying blend shape parameters within manually chosen bounds in order to improve the generalization of action recognition models.

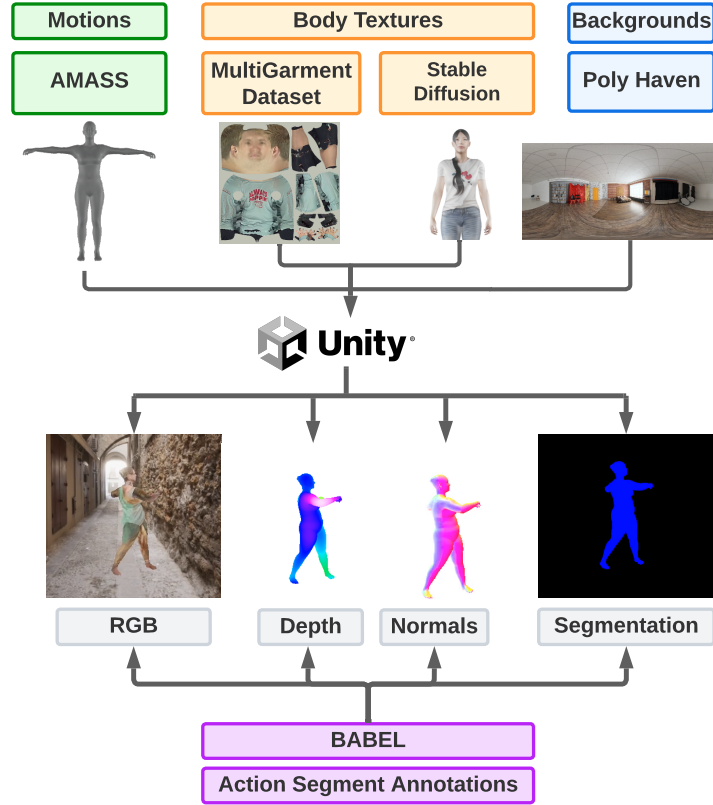


Figure 20: Overview of our data generation pipeline using multiple publicly available data sources.

Body and background textures Our main source of clothing textures is the data published by Multi-Garment Net [16] which includes 98 body clothing textures. A downside of this data is that it mainly depicts young to middle aged males, mostly Caucasian and always with exactly the same face. In order to address this, we generate front and back views of realistic body textures with stable diffusion [209] for eight younger and eight older female characters with varying ethnic backgrounds. Part of our dataset consists of scrambled textures; we crop individual subregions of full body textures like left and right arm or leg and randomly mix them with each other, resulting in a total of 1240 different textures. We made sure to balance male and female source textures for this process. To set up multiple different background scenes for the motions, we make use of a total of 201 HDRI images from Poly Haven [191] as skybox backgrounds which we selected to represent a roughly equal amount of 103 indoor scenes and 98 outdoor scenes.

Varying lighting Variations in lighting are known to have a major impact on the quality of vision based recognition systems [128]. In our scenes, part of the lighting comes from HDRI images, the intensity can be adjusted to create strongly or weakly lit environments. However, the background light is not sufficient to properly illuminate the person. Therefore, four light sources were integrated at the positions of the cameras to obtain similar lighting conditions from all angles. In order to create diverse and realistic lighting settings, we randomly parameterize the foreground and background lighting intensity within manually selected bounds.

Table 11: Supervised training on AMARV. Table 12: Synthetic data approaches fine-tuned on NTU RGB+D.

Model	Init	Accuracy		Bal. Acc.	Method & Training	NTU 60	
		Top-1	Top-5			CS	CV
<i>Val-as-test-260</i>							
X3D-S		41.35	73.10	12.1	HPM _{RGB} + Traj [147]	S	75.8 83.2
MViTv2-S	MF-A	41.3	73.4	12.7	SURR[266]	R	89.0 93.1
					SURR[266]	S+R	89.6 94.1
<i>Val-as-test-120</i>							
X3D-S		42.2	71.7	21.8	X3D-S	A → R	88.54 97.2
MViTv2-S	MF-A	42.3	72.4	23.2	MViTv2-S	MF-A → R	93.8 98.6
S = Synthetic data R = Real data A → R = Supervised pre-training on AMARV then fine-tuning on real data only MF-A → R = Masked feature pre-training [281] on AMARV then fine-tuning on real data only							
<i>Val-as-test-60</i>							
X3D-S		44.8	75.4	28.8			
MViTv2-S	MF-A	45.1	76.3	28.5			

Multi view recordings One of the features of our synthetic data generator is the use of four simultaneously recording cameras (front, left, right and back camera; Figure 18). While keeping the actor in the image center, each camera is allowed to randomly position itself for each individual clip within $\pm 30^\circ$ in a 2D plane, resulting in a large range of possible camera angles, see also Figure 19. It is possible to recover the exact angle of a camera from recorded metadata, allowing for a finer differentiation of multiple views. We evaluate the covered space of each animation before rendering it and then arrange the cameras around the center of that area with a minimal distance, ensuring that an actor coming up close to the camera will never walk past it.

Multi-modal data output SynthAct captures depth information, 2D and 3D bounding boxes, body keypoints, pixel normals, and segmentation maps for each person in the scene in addition to the ground truth labels of the animations. Furthermore, meta-information such as camera and light source pose, light source intensity values and background information are saved frame wise.

Raw depth data requires significant amounts of storage up to a point which is infeasible for a dataset like AMARV. To reduce the required storage space and enable us to use common video compression algorithms, we apply the slightly adapted approach of Sonoda et al. [241] to convert depth values to RGB by mapping them to hues. The mapping induces a resolution of about 1530 different values. To minimize the loss of accuracy during conversion, we analyze the motion range in the z-axis and perform z-cropping, meaning that we map the closest and furthest point in the scene to the available range. By doing so, we achieve a varying precision for our videos ranging from sub-millimeter precision for motions with little z-movement to 10 mm precision for motions with 15 m of z-movement. Both values are within the expected precision of a modern real world depth estimation camera.

Table 13: Cross-view-subject evaluation on NTU RGB+D, comparison with SURREACT.

Training data		0°	45°	90°
SURREACT [266]	R	86.9	74.5	53.6
X3D-S	R	86.4	77.8	60.4
SURREACT [266]	S → R	84.1	77.5	66.2
X3D-S	A → R	89.9	81.8	68.0
MViTv2-S	MF-A → R	94.4	84.5	65.1
SURREACT [266]	S + R	89.7	82.0	69.0

4.1.2 Experiments

We provide a pipeline for generating synthetic videos from in the wild human action data in order to partially or fully replace it during training. In Section 4.1.2.1, we list results on pre-training on our large scale dataset AMARV. Fine-tuning on NTU RGB+D for both, RGB and depth based action recognition is evaluated in Section 4.1.2.2 and Section 4.1.2.3. In Section 4.1.3 we make use of Synthetic Smarthome and supplement as well as replace real world data with the intention to maintain performance and in Section 4.1.3.1 we analyze the domain gap between data generated by SynthAct and real world datasets.

Architectures and Training We make use of multiple neural network architectures for our experiments. For end-to-end pre-training on AMARV in Section 4.1.2.1 as well as for downstream experiments in Section 4.1.2.2, we select a commonly used convolutional neural network, X3D-S, as well as a commonly used visual transformer architecture, MViTv2-S. For each network we choose the small configuration which allows for training on AMARV in a reasonable amount of time without access to large scale hyperscalers. In Section 4.1.3 we also use MViTv2-S, but since we focus on synthetic data in the downstream dataset we initialize the model with Kinetics pre-trained weights. For feature extraction based experiments in 4.1.3.1 as well as depth based action recognition on AMARV and NTU RGB+D, we make use of the pre-trained OMNIVORE[87] network which is based on Swin-B. All our experiments can be reproduced with four NVIDIA A100 GPUs or comparable. Unless noted otherwise, fine tuning is performed for 100 epochs with batch size 128 for X3D-S and 48 for MViTv2-S.

Datasets For our experiments we make use of Toyota Smarthome [52] and AMASS [155] in combination with BABEL [195] as motion source and for experiments. NTU RGB+D [228] is used as downstream dataset.

4.1.2.1 AMARV Pre-Training

The BABEL annotations from Punnakal et al. [195] provide more than one label for many AMASS segments which does imply a multi-label problem but the authors decided to ignore the multi-label property and instead chose to sample segments multiple times, once for each possible label. We follow them in this decision. This inhibits a perfect Top-1 score on AMARV, since a clip will be evaluated multiple times with different labels. Following [195], we list Top-5 accuracy to mitigate this problem. BABEL does not provide

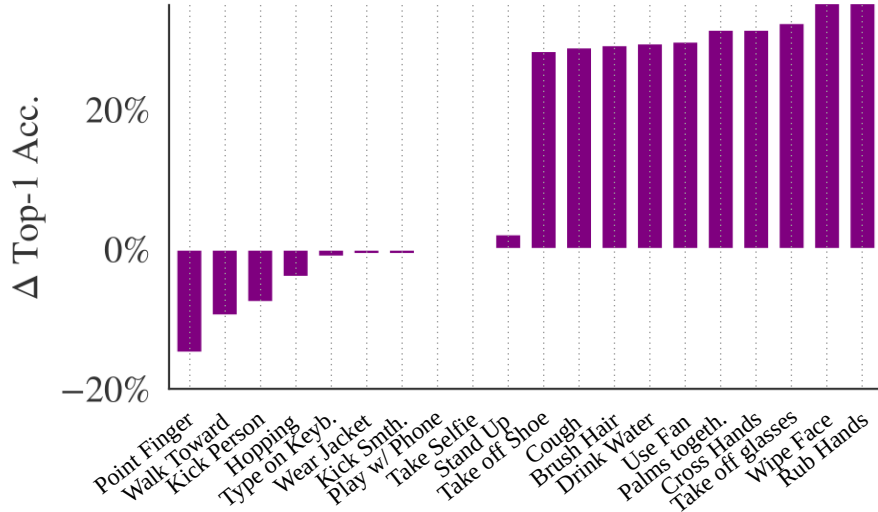


Figure 21: Gain of pre-training on AMARV over OMNIVORE baseline on NTU RGB+D [228] using depth only. We plot the gain in per-class Top-1 accuracy for the top ten and bottom ten classes. Pre-training on AMARV improves the Top-1 accuracy on 52 out of the 60 total classes.

the test set labels, for this reason we propose to publish results by using the official BABEL validation set for testing and refer to this protocol by *val-as-test*.

X3D-S is trained from scratch for 300 epochs, for MViT2-S we apply the self-supervised masked feature approach from [281] for 230 epochs which we refer to as MF-A, followed by 70 epochs of supervised training.

We list our results in Table 11, for all as well as the most frequent 120 or 60 classes. Despite their different architectures, both networks perform very similarly on AMARV.

4.1.2.2 Fine-tuning on NTU RGB+D

We consider AMARV as a dataset for large scale synthetic human action video pre-training and therefore evaluate models by fine-tuning performance on downstream datasets. We list supervised fine-tuning of X3D-S pre-trained on AMARV ($A \rightarrow R$) or MViT2-S with masked feature pre-training (MF-A $\rightarrow R$) in Table 12.

On NTU RGB+D we can compare to the experiments from Varol et al. [266] who publish full fine-tuning results based on mixed synthetic SURREACT and real data training ($S + R$). X3D-S pre-trained on AMARV outperforms [266] on the cross-view split and performs almost equal on the cross-subject split. MViT2-S significantly improves over the results of [266] on both NTU RGB+D splits. Note that SURREACT [266] makes use of mixed synthetic and real data for training while we use synthetic data of AMARV *only for pretraining* since our synthetic data has no direct correspondence to NTU.

Cross-view-subject comparison. Varol et al. [266] specifically designed SURREACT to improve cross-view performance. They extended the cross-subject evaluation protocol on NTU RGB+D by only allowing training on front view videos while testing individually on the 0° , 45° and 90° views. Our results are listed in Table 13, we do not use synthetic data on NTU RGB+D but only make use of our synthetic multi-view pretraining dataset. X3D-S with synthetic pre-training on AMARV ($A \rightarrow R$) not only outperforms the results of [266] on the comparable synthetic pre-training based setting ($S \rightarrow R$) but even performs

Table 14: NTU RGB+D [228] and AMARV using depth only. Results demonstrate that the models pre-trained on AMARV are more generalizable and show superior fine-tuning performance on NTU.

Method	NTU Cross-Subj.		NTU Cross-View		AMARV	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
TSN [277]	73.00		68.32			
Dhiman et al. [61]	68.30		72.40			
HPM _{3D} [147]	71.50		70.50			
DDIs [204]	84.00		82.06			
Wang et al. [275]	87.73		87.37			
Swin-B (OMN., frozen)	46.89	78.50	45.24	78.11		
Swin-B (AMA., frozen)	62.01	91.20	63.70	92.00		
Swin-B (OMNIVORE)	92.80	99.42	92.84	99.61	22.3	53.44
Swin-B (AMARV)	92.89	99.41	93.38	99.65	37.22	72.85

on par with mixed synthetic and real fine-tuning performed by [266] (S + R). MViT2-S with self supervised pre-training on AMARV significantly outperforms all other methods on the front view but is not as good for the 90° view.

4.1.2.3 Depth-based Fine-Tuning

In this section, we describe our depth-based experiments on NTU RGB+D. We use OMNIVORE-B [87], which is based on Swin-B [149], as our baseline model. OMNIVORE is pre-trained jointly on three datasets with different visual modalities: ImageNet-1K (images), Kinetics-400 (videos) and SUN RGB-D (images and depth maps, but no humans). We use the disparity maps as input for the Swin-B model. Specifically, we first reconstruct depth maps from the depth videos coded in RGB format, and take the inverse depths, i.e., disparity maps, as input for the model.

We use standard image augmentation methods, such as RandAugment [45], Mixup [297], and Random Erasing [303], following [87]. First, we compare the Swin-B model (initialized with OMNIVORE-B weights) on AMARV with full fine tuning, further referred to as Swin-B AMARV or frozen backbone fine-tuning. Results in Table 14 show that full fine-tuning (AMARV) is necessary due to the large domain gap between AMARV and pre-training datasets of OMNIVORE, fine tuning of the patch embedding as well as the classifier alone (OMNIVORE) does not provide good performance.

In order to assess the generalization capability, we carry out experiments on the NTU RGB+D [228] dataset using the masked depth maps provided by the authors. We first freeze the main portion while keeping the patch embedding layer and classifier of the two models, i.e., Swin-B (AMARV) and Swin-B (OMNIVORE), trainable. Results in the middle compartment of Table 14 show that the model fine-tuned on AMARV is more generalizable, demonstrating the effectiveness of the proposed dataset. Additionally, we compute the per-class Top-1 accuracy of these two models, and plot the gain for the top ten and bottom ten classes in Figure 21. Furthermore, we fine-tune these two models fully on the NTU

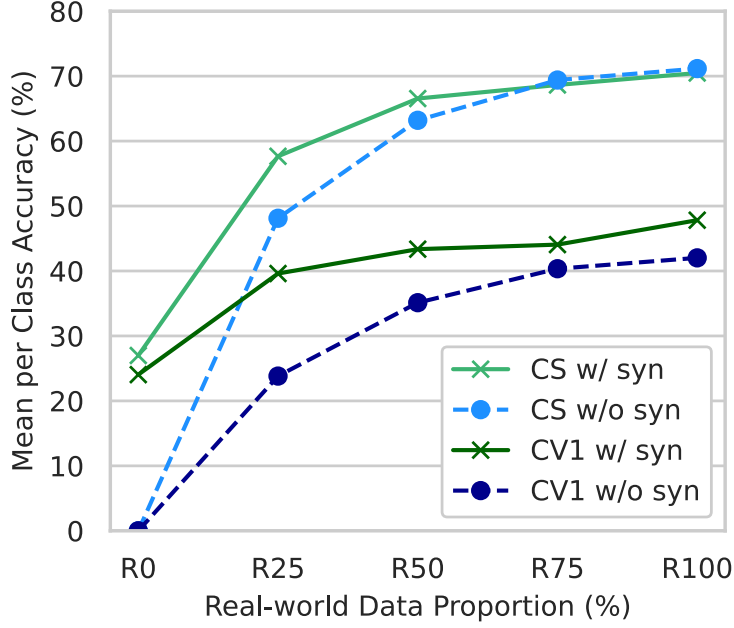


Figure 22: Comparison between synthetically supplemented training with real-world-only experiments considering varying shares of real world data on Toyota Smarthome.

RGB+D dataset, and show the results in the bottom compartment of Table 14. While both models outperform all existing depth-based methods in all settings, the model pre-trained on AMARV shows even superior results over the OMNIVORE baseline, especially in the cross view setting.

4.1.3 Synthetic data in Downstream Datasets

This experiment assesses if synthetic data can minimize the need for real-world data in downstream fine-tuning, particularly in privacy-sensitive, indoor scenarios like Toyota Smarthome. Using our SynthAct generator, we present results on the Synthetic Smarthome dataset which provides a counterpart to Toyota Smarthome. The Synthetic Smarthome dataset was generated without the use of ground truth poses by employing 4DHumans [89] as pose estimation on Toyota Smarthome to extract poses and motions and running these motions through SynthAct.

Figure 22 shows performance retention on real world data reduction across two training splits: *cross-subject*, with a larger multi-view training set, and *cross-view one*, only offering a single camera perspective for training. We observe that supplementing real world data with synthetic counterparts is beneficial for smaller real world data portions (25% and 50%) but provides no gain if real world data portions are large (75% and 100%). However, we can maintain or improve performance with 25% real world data while using 75% synthetic data from the Synthetic Smarthome dataset. Our generator’s multi-view video output from single view input poses notably boosts performance in the sparse cross-view setting, allowing almost 75% reduction in real-world data (remaining with only 469 real world samples) while mostly maintaining real world performance. This is especially relevant for privacy-sensitive scenarios like household monitoring where it is preferable

Table 15: Toyota Smarthome comparison on cross-subject and cross-view 1 splits.

Methods	Modality		mPCA (%)	
	Pose	RGB	CS	CV1
P-I3D[51]	✓	✓	54.2	35.1
Separable STA [52]	✓	✓	54.2	35.2
VPN [54]	✓	✓	60.8	43.8
VPN+SSTA-PRS [291]	✓	✓	65.2	-
MMNet [22]	✓	✓	70.1	37.4
VPN++ [53]	✓	✓	71.0	-
LSTM [154]	✓	×	42.5	13.4
MS-AAGCN [232]	✓	×	56.5	-
2s-AGCN [231]	✓	×	57.1	22.1
5C-AGCN [291]	✓	×	62.1	22.8
DT [274]	×	✓	41.9	20.9
I3D [30]	×	✓	53.4	34.9
AssembleNet++ [212]	×	✓	63.6	-
VPN++ [53]	×	✓	69.0	-
Ours (R)	×	✓	71.2	42.2
Ours (S + R)	×	✓	70.5	47.8

Table 16: Similarity of synthetic datasets (including our proposed AMARV dataset) and three real datasets with OMNIVORE features.

Real	Synthetic	MMD	KL-Div	JS-Div
NTU RGB+D [228]	Surreact	24.89	393.89	67.05
	Mixamo	51.44	502.72	103.96
	AMARV	33.81	300.68	74.97
UCF-101 [242]	Surreact	33.28	501.58	118.98
	Mixamo	38.49	518.28	141.23
	AMARV	20.40	353.08	104.98
HMDB-51 [131]	Surreact	32.79	539.80	88.85
	Mixamo	37.99	414.55	86.91
	AMARV	18.27	195.61	50.52

to collect as little real world data as possible. While we are not able to solely rely on synthetic data, we can significantly reduce the need for real world recordings.

Table 15 compares our results with existing state-of-the-art approaches for Toyota SmartHome. Our baseline RGB-only model trained on real data achieves 71.2% mPCA on the cross-subject split and 42.2% on the more challenging cross-view 1 split. The addition of synthetically generated data (S + R) allows us to significantly improve on the cross-view split, achieving 47.8% mPCA; an improvement of over 5 percentage points. Notably, this RGB-only approach with synthetic augmentation outperforms even multi-modal methods that leverage both RGB and pose information during inference, such as VPN [54] (43.8% on CV1) and other state-of-the-art multi-modal approaches, demonstrating that our synthetic multi-view data provides more effective cross-view generalization than additional modalities.

4.1.3.1 Domain Analysis

In this section we examine the domain gap between the data generated by SynthAct and real world datasets. To quantify the domain gap, we employ a pre-trained feature extractor (OMNIVORE [87]) and analyze the similarity of extracted features through k-nearest neighbor (k-NN) experiments and embed analysis.

We use the entire HMDB-51 dataset and a subset of AMARV containing the same action classes as HMDB-51 and perform k-NN classification using OMNIVORE features.

For comparison, we also test on the Mixamo synthetic dataset [44]. Results in Table 16 demonstrate that the domain gap between AMARV and HMDB-51 is smaller than that between Mixamo and HMDB-51. AMARV features generalize better to real-world data than Mixamo features, indicating that the SynthAct generator produces more realistic synthetic data.

4.2 BIOMECHANICAL SYNTHETIC DATA GENERATION

This section is based on our paper published at NeurIPS 2024 [221].

Building on the same AMASS motion sequences used to generate AMARV in Section 4.1, this section bridges computer vision and biomechanics by enriching these motions with muscle activation simulations from validated musculoskeletal models. The resulting Muscles in Time (MinT) dataset enables viewing the same human motion sequences from a biological perspective, coupling 402 muscle strand time series with the RGB and depth information generated by SynthAct.

As explored in the introduction, neural networks for human motion understanding face a challenge analogous to Plato’s cave: like the prisoners seeing only shadows, they rely on indirect representations rather than direct, biologically grounded data. In Plato’s allegory, prisoners in a cave see only shadows cast on the wall, not the true objects. Similarly, neural networks trained on accessible data, such as RGB and depth-based video recordings or motion capture, only perceive surface-level appearance of motion in contrast to the inner mechanics of the human body.

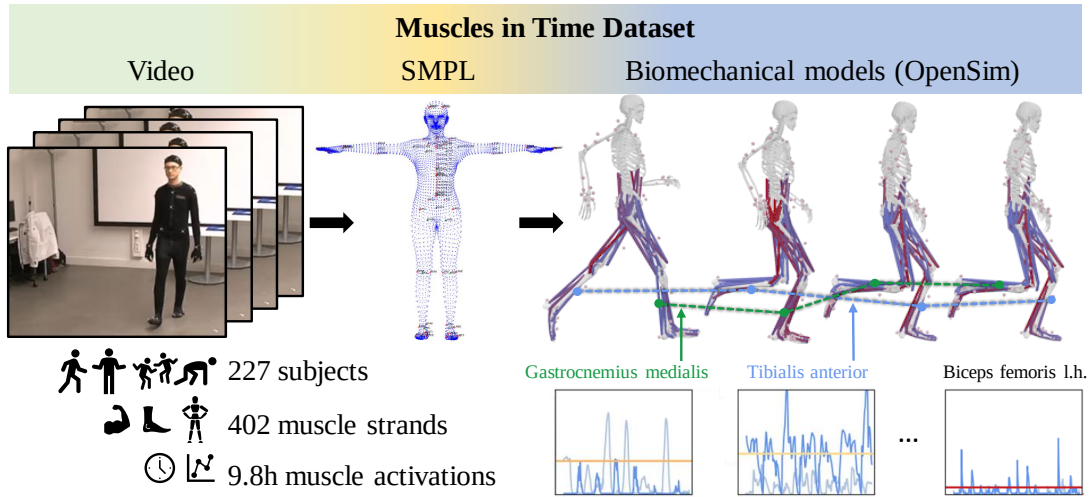


Figure 23: Simulation pipeline of the Muscles in Time dataset. The SMPL representation is extracted from videos, then, the SMPL represented motions are mapped to bio-mechanically validated human body models to simulate fine-grained muscle activation, connecting computer vision with biomechanical research. Bottom right: two activation sequences for exemplary muscles. Images from [58, 158]

This reliance on external observations provides an incomplete understanding of the true complexities of human motion. Just as the prisoners lack a direct view of the objects casting the shadows, current models lack exposure to the internal workings of the human body, such as the muscle activations driving motion. This gap limits their ability to develop in-depth understanding of physical exertion or motion difficulty.

The field has progressed from camera-based motion capture and activity prediction to pose-based recognition systems that account for the body and its motion over time. These advances, while significant, still overlook the interplay of muscle activations driving pose sequences and patterns that underlie physical performance.

Collecting electromyographic (EMG) data or more commonly used surface electromyographic (sEMG) data, as a measure of muscle activation, presents challenges. It is resource intensive, requiring specialized equipment, controlled environments, and is an invasive procedure. Existing EMG and sEMG datasets are small, limited in scope, and not representative of the variety of human motions. These limitations hinder the development of neural networks that can generalize across different types of motion and subjects.

While acknowledging the contributions of EMG and sEMG datasets, we identify an opportunity to supplement this domain with a synthetic dataset that overcomes some limitations of real-world data collection. The strength lies in its scale and detail of muscle activation data, a feat not achievable through conventional methods alone.

Every dataset, simulated or real, has domain-specific fidelity and relevance. Real-world recordings offer authenticity that underpins our understanding of human biomechanics, capturing nuances such as subject-specific EMG measurements that vary over the course of a day. Simulated datasets, like ours, offer a complementary perspective by providing comprehensive and systematically controlled data at unprecedented scale for the understanding of muscle activation patterns through a scalable data acquisition pipeline.

In this work, we present a comprehensive large-scale dataset incorporating muscle activation information. We enrich existing motion capture datasets with muscle activation simulations from biomechanical models of the human body. Our pipeline uses simple pose and shape sequences with estimated weight and mass to simulate muscle activations for individual movements. Using this, we generate the muscles' activation that fit the provided human motions. Figure 23 provides an overview of our pipeline.

We showcase the utility of muscle activations as an additional data type for human motion understanding and gather insights by visualizing our data. Our dataset, with its magnitude and detail, describes muscle activation across a wide array of movements. By enhancing the state of available data to other researchers, we expand the potential for scientific investigation and innovation in the study of human motion.

4.2.1 *The Muscles In Time dataset*

To develop the *Muscles in Time (MinT)* framework, we harnessed the AMASS dataset, which consolidates marker-based motion capture (mocap) sequences into a uniform representation using the MoSh++ method, resulting in Skinned Multi-Person Linear Model (SMPL) parametric representations for body pose and shape. AMASS amalgamates mocap data from multiple sources, including the KIT Whole-Body Human Motion Database [159], BMLrub, and BMLmovi [86], comprising over 11,000 motion captures from more than 300

subjects. This extensive collection enables the analysis of a broad array of human movements, providing a rich basis for studying diverse motion patterns.

The SMPL model serves as a pivotal link, translating mocap data from AMASS into mesh representations in a format compatible with the OpenSim [58] platform. OpenSim is instrumental in constructing intricate biomechanical models that simulate the musculoskeletal system's physical and mechanical properties with high fidelity, allowing for an in-depth and accurate analysis of human motion. These models are intricate, requiring precise definitions of joints, masses, inertia, and muscle parameters, such as maximum isometric force, which act as the force-generating actuators.

In this work, we abstain from developing new biomechanical models due to the complexity and special expertise required. Instead, we utilize established, pre-validated models, specifically the lower body model by Lai *et al.* [135] and the thoracolumbar region body model by Bruno *et al.* [23], see Figure 24. These models simulate muscle activations for a large number of individual muscle strands across various muscle groups.

Tailoring body model parameters to an individual's specific anatomical properties results in similar difficulties as with the creation of new body models, therefore model parameters are most commonly used as specified in the validated original models [23, 135], in the OpenSim community. We follow this approach, providing simulation results for standard models rather than subject-specific human bodies.

To integrate human motion data from AMASS with OpenSim, we map virtual mocap markers to the SMPL-H body mesh's surface vertices, following the method proposed by Bittner *et al.* [17]. This results in a carefully curated selection of 67 strategically placed vertices that accurately represent marker positions on the body mesh, visualized on the left of Figure 24. We deliberately and consistently exclude soft tissue dynamics from the SMPL-H mesh generation to maintain consistent marker positions during motion.

Despite OpenSim's automatic scaling capabilities, manual adjustments of marker positions are sometimes necessary to accurately reconcile differences between simulated and real-world data. These adjustments are made on a subject-specific basis, rendering our pipeline semi-automatic. The manually adjusted marker positions are thoroughly documented, validated and shared to ensure the reproducibility of our simulations.

AMASS lacks data on ground reaction or contact forces, which are crucial for realistic motion simulation. To address this, we integrate the OpenSimAD [70] implementation

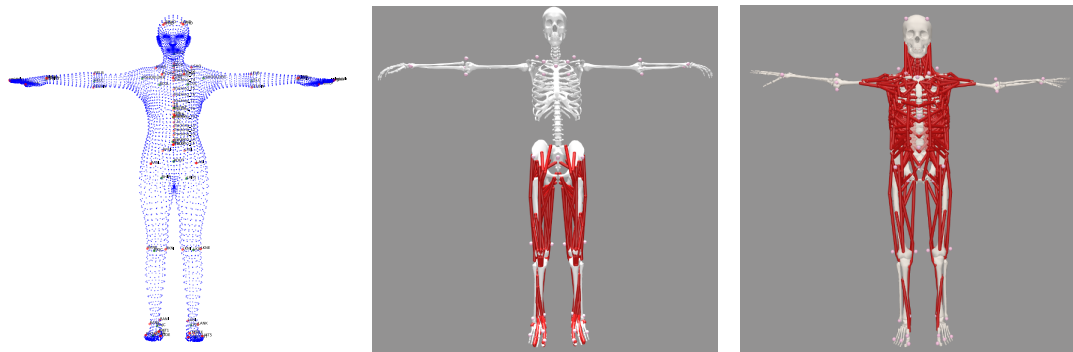


Figure 24: The AMASS body model with specific indices mapped onto the OpenSim lower body model by Lai *et al.* [135] (middle) and model of the thoracolumbar region by Bruno *et al.* [23] (right). Best viewed by zooming in.

used in the OpenCap [260] project, which calculates ground reaction forces based on kinematic data and the musculoskeletal model. We employ a tailored parameter setup to optimize the trajectory problem, balancing computational load and accuracy.

Kinematic data is analyzed using OpenSim’s *Inverse Kinematics* method. Muscle activations for the lower body are derived from a trajectory optimization problem described in [260]. The estimated ground reaction forces serve as inputs for the *Static Optimization* method, which calculates muscle activations for the thoracolumbar region.

Due to the computational demands of the trajectory optimization problem, we process the data in segments, ensuring manageable computation times without compromising the continuity of the motion capture sequences. We implement overlapping buffers to mitigate inaccuracies during segment processing, discarding data that fails to meet our stringent error tolerance criteria to maintain data quality.

The Muscles in Time (MinT) dataset represents a significant contribution to the field of biomechanical and computer vision simulation. By integrating and refining existing methodologies, we present a robust pipeline that facilitates the accurate simulation of human muscles in motion by combining established biomechanical models with high quality mocap data.

4.2.1.1 Dataset Composition

Due to missing information on external forces based on object interactions, inaccurate motion capture recordings or non-converging simulations, the *MinT* dataset covers a subset of its originating datasets in AMASS and does not follow their statistics.

Anthropometrics While the motion capture recordings in AMASS provide gender labels, information about subjects’ height and weight is approximated from the SMPL body model. Body weight is calculated by volume resulting from average shape parameters, which follows the approach of Bittner *et al.* [17]. The weight is relevant for the calculation of ground reaction forces and the distribution of weight in the model, affecting the muscle activation in different parts of the body.

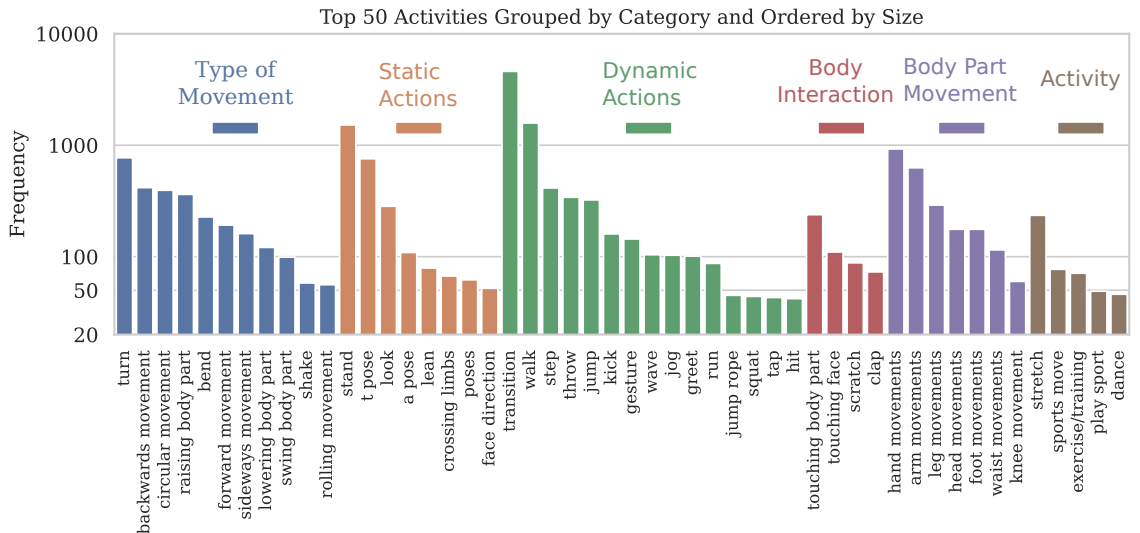


Figure 25: Prevalence of different motions in the MinT dataset.

Figure 26 shows the distribution of weight across all subjects, indicating significant diversity in body composition. Underweight subjects are slightly underrepresented in the dataset, while subjects in the obese range are well represented.

Composition of Subdatasets Within AMASS, *MinT* is limited to the subdatasets EyesJapan, BMLrub, KIT, BMLmovi, and TotalCapture. The short sequences in BMLmovi typically depict single activities, while the longer ones for example in EyesJapan capture a more diverse range of motions within a single sequence. Since we compute activation information for shorter segments and rejoin them afterwards, longer sequences are more prone to gaps in the analysis due to individually failing segment computations.

Motion Diversity Figure 25 displays the frequencies of grouped activities on a logarithmic scale. The action labels are based on the BABEL dataset, a large annotation dataset which is coupled with AMASS. Most interesting are dynamic actions, since expected mus-

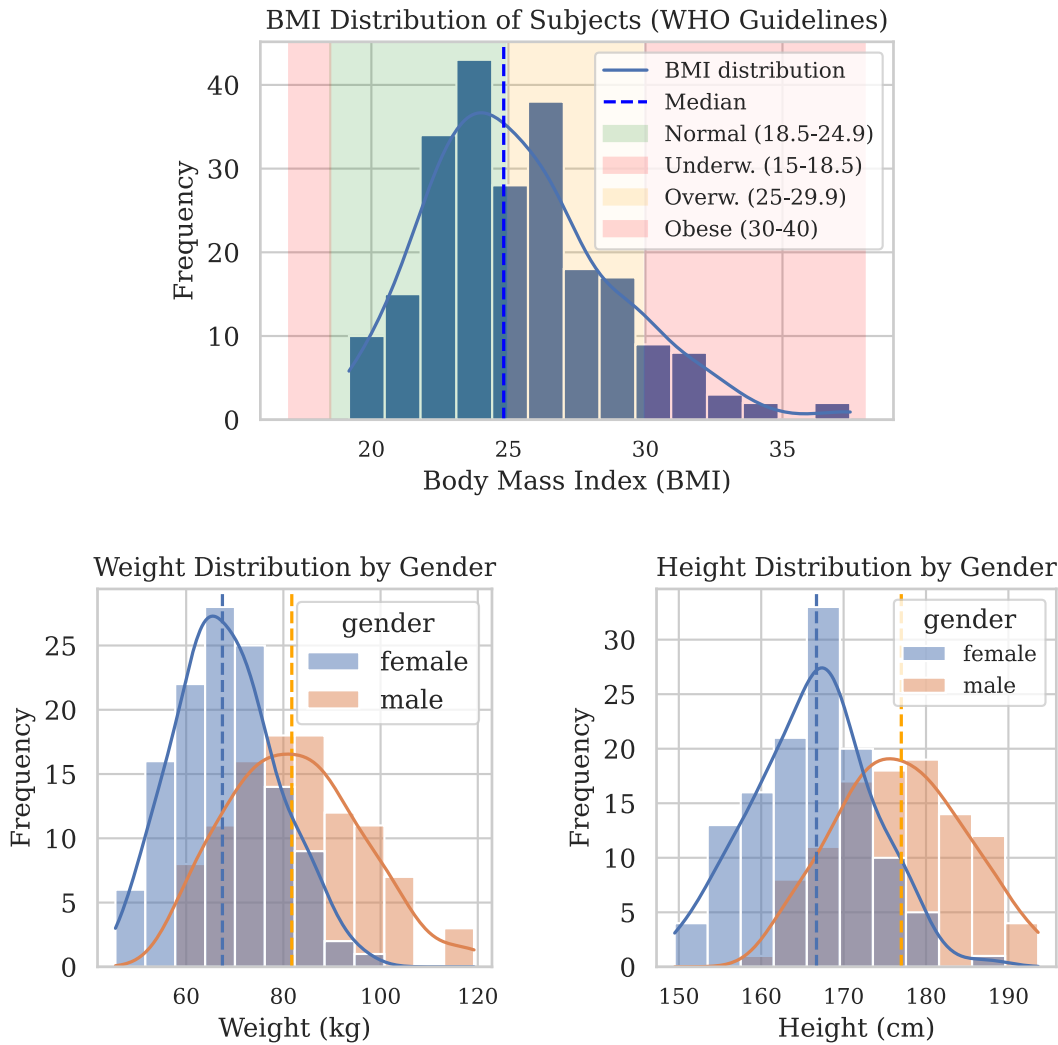


Figure 26: Approximated weight and height distribution of the analysed subjects in the MinT dataset.

cle activations for simple dynamic actions are well documented and we present a short qualitative analysis based on such actions in Section 4.2.1.2.

4.2.1.2 Data Analysis, Validation, and Visualization

In Figure 27 (left) we explore the interrelation between different activities by investigating our simulated muscle activation time-series. To this end, we extract features from the temporal muscle activation sequences using tsfresh [40], a commonly used framework in time series analysis that extracts a feature vector based on time series characteristics such as mean, skewness, standard deviation *etc.* We chose distinct and descriptive groups of activities from the BMLmovi subset such as jumping, kicking, stepping and walking, the resulting features were normalized and clustered using FINCH [217] and visualized with h-NNE [218]. It can be observed that activities do not only cluster together based on variations within the same category (e.g., different types of jumps, including jumping jacks), but also align closely across different categories, when they share similar motion patterns (e.g. sideways movements). This underlines the descriptive information contained in our simulated muscle activation sequences for characterizing activities.

To provide better insight into the quality of simulation and the realism of the muscle activation sequences, we provide an overview of key muscle activation characteristics in the appendix under Appendix A.4.

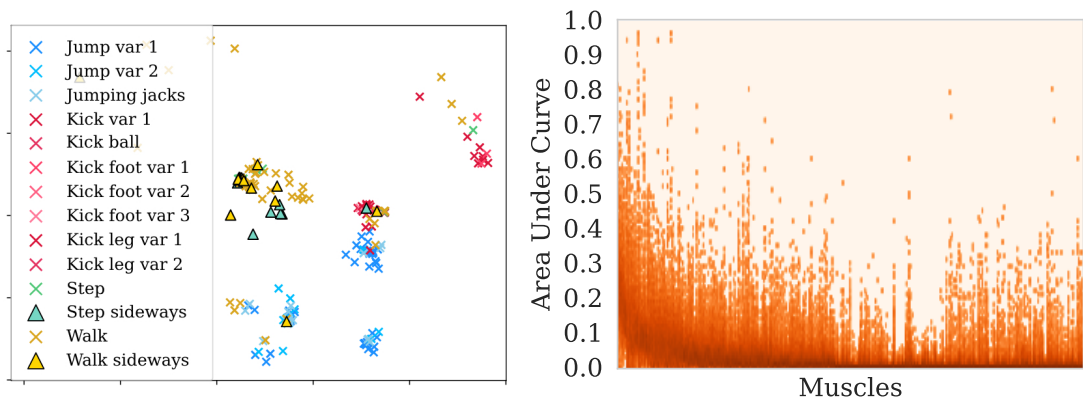


Figure 27: **Left:** Clustering of multiple activities within the BMLmovi dataset by muscle activation features. **Right:** Column-wise color coded histograms of areas under muscle activation curves for 402 muscle strands, sorted by histogram medians. Log-normalized color map.

Table 17: Human motion-to-muscle activation prediction results for the lower- and upper body model.

Act	VQ-VAE [298]			FConv [84]			LSTM [108]			Mamba2 [50]			Transformer [268]		
	R↓	S↓	P↑	R↓	S↓	P↑	R↓	S↓	P↑	R↓	S↓	P↑	R↓	S↓	P↑
Lower body model															
all	0.058	59.7	0.40	0.052	66.0	0.49	0.052	57.8	0.48	0.051	55.4	0.49	0.048	45.1	0.54
jump	0.062	66.7	0.52	0.053	68.1	0.66	0.052	62.2	0.67	0.051	60.8	0.68	0.051	52.3	0.71
kick	0.069	69.1	0.38	0.057	74.9	0.55	0.058	66.5	0.55	0.059	67.6	0.55	0.053	54.8	0.62
stand	0.056	60.0	0.42	0.049	64.4	0.51	0.050	58.2	0.51	0.049	55.1	0.52	0.046	45.0	0.58
walk	0.053	57.7	0.66	0.046	61.7	0.73	0.045	53.7	0.73	0.045	50.4	0.74	0.044	42.4	0.77
jog	0.059	64.8	0.58	0.052	69.1	0.66	0.050	61.5	0.68	0.047	58.2	0.69	0.046	51.1	0.71
dance	0.070	71.4	0.40	0.064	76.0	0.59	0.063	71.5	0.57	0.063	70.2	0.57	0.057	58.5	0.65
Upper body model															
all	0.041	115.3	0.32	0.034	114.8	0.47	0.035	111.1	0.48	0.034	112.2	0.50	0.033	107.7	0.55
jump	0.064	118.1	0.38	0.052	119.6	0.54	0.054	115.4	0.56	0.053	117.2	0.58	0.052	112.7	0.63
kick	0.058	122.2	0.35	0.048	121.5	0.55	0.048	118.1	0.57	0.048	119.4	0.58	0.044	114.8	0.65
stand	0.039	117.6	0.34	0.031	118.2	0.48	0.031	114.2	0.49	0.030	114.9	0.51	0.028	110.5	0.55
walk	0.028	110.2	0.43	0.021	109.8	0.55	0.022	105.6	0.57	0.020	106.8	0.59	0.019	102.6	0.63
jog	0.040	117.1	0.52	0.034	118.5	0.64	0.032	113.9	0.66	0.031	115.4	0.66	0.029	110.8	0.71
dance	0.046	127.3	0.29	0.041	129.5	0.48	0.044	126.7	0.48	0.039	128.2	0.49	0.036	121.8	0.59

Best values marked in boldface. Second best values marked with underline.

R: RMSE S: SMAPE P: PCC

4.2.2 Motion to Muscle Activation Estimation Benchmark

While OpenSim provides a means for simulating muscle activations, it is both highly compute intensive as well as sensitive regarding hyper parameters. These properties limit it to be used by experts in an offline manner and prevent usage in everyday applications. In this section we explore the usage of *MinT* as a training dataset for the estimation of muscle activation based on pose motion. Such networks provide muscle activation estimation in an instant and can easily be deployed for various downstream tasks. Given pose motion sequences, we use the preprocessing step defined by [96] which adjusts skeletal structure to a uniform format and normalizes positions and enriches the resulting data points with additional features. This procedure maps each input to a 263-dim descriptor, resulting in samples of the form $x = [x_1, \dots, x_T]$, $x_t \in \mathbb{R}^{263}$. For training our models we segment the resulting data into clips of 1.4 second sampled at 20 frames per second, resulting in $T = 28$ input frames. Given a network $f_\Theta : \mathbb{R}^{T \times d} \mapsto \mathbb{R}^{T \times m}$ we predict $f(x) = y$ with $y = [y_1, \dots, y_T]$, $m = 402$ being the number of individual muscle strain activations simulated in our dataset, consisting of 80 lower body muscle strains from [70] and 322 muscle strains for the upper thoracolumbar region body model [23]. Evaluation is performed by calculating Root Mean Squared Error (RMSE), Pearson Correlation Coefficient (PCC), and Symmetric Mean Absolute Percentage Error (SMAPE). RMSE is commonly used but highly susceptible to data scaling, resulting in significantly lower error values for downsampled data. In practice, EMG signals vary strongly between subjects, scaling of signals is therefore a common preprocessing step. PCC is a good indicator for muscle activation series similarity, since it is scale and offset invariant. SMAPE allows for considering fixed offsets as error while being less sensitive to scaling in comparison to RMSE. PCC and SMAPE are calculated for each muscle strain individually and averaged. For our benchmark we use the train, val and test splits defined by the BABEL dataset [195]. Evaluation results are reported separately for muscles of the upper and lower body model.

4.2.3 Experiments

We evaluate five different architectures on *MinT*. Since we make use of human motion as input for our predictor, we adapted a common architecture for motion-to-motion prediction from [298] to the task of motion-to-muscle activation prediction by simply exchanging its prediction head. We further evaluate a Long Short-Term Memory (LSTM) [108], a fully convolutional network (FConv) [84], a Mamba2 Mixer model [50] and a simple transformer architecture [268] with 16 transformer layers, results for the lower and upper body model are listed in Table 17. All models are trained from scratch for 300k iterations with a batch size of 256 unless noted otherwise.

The evaluated transformer architecture showed the best results as compared to the adapted VQ-VAE model, LSTM, FConv and Mamba in all metrics on all evaluated motion types. The results of the experiment also show the importance of reporting PCC and SMAPE, since the differences in RMSE are marginal, while PCC and SMAPE both show significant improvements. We suspect this to be the case, since many muscles in the human body are mostly relatively inactive unless required for specific motions. For a simple analysis of this effect, we calculated the integral for each individual ground truth muscle activation sequence in all our validation set chunks and created 402 color coded

histograms that are sorted by median and vertically displayed side by side on the right hand side of Figure 27 (one column in the image is a single muscle activation integral area frequency histogram). A wide range of muscles are rarely activated, resulting in the majority of activation sequences displaying integral areas significantly below 0.1 or 0.05. This property is challenging for RMSE and SMAPE, average RMSE shows a small error, since most activations are close to zero and SMAPE reports a high percentage error, since a deviation from a close to zero value is more likely to result in a high percentage deviation. For similar reasons, the upper body model displays lower RMSE and higher SMAPE, the upper body model contains a larger number of small and rarely activated muscles in contrast to the lower body model.

We recommend that future users of our dataset actively evaluate on either the full range of provided muscle activations or choosing one of these muscle strand subsets depending on their specific application. Please also see the appendix for additional dataset statistics and consider the follow-up work of [148].

4.2.3.1 Qualitative Results

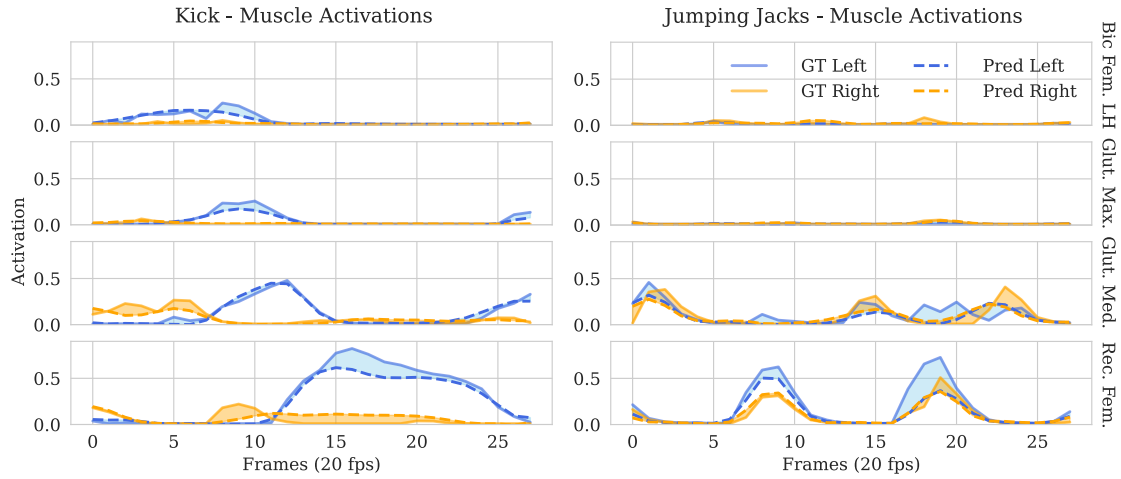


Figure 28: Example lower body muscle activations (split in left and right muscle strands) for the actions *kick* and *jumping jacks*. It is clearly visible that the kick is performed with the left leg. During *jumping jacks*, *gluteus medius* and *rectus femoris* are activated alternately for both legs.

In Figure 28 we list two examples from our dataset, one displaying the action *kick*, the other displaying the action *jumping jacks*, predictions are calculated with the 8-layer transformer architecture. The figure displays four key muscles essential for lower body locomotion; *biceps femoris long head* (knee flexion and hip extension), *gluteus maximus* (hip extension and external rotation), *gluteus medius* (abduction and medial rotation of the hip), and *rectus femoris* (hip flexion and knee extension), each for the left and right body half. The kick is clearly executed with the left leg with *rectus femoris* providing the force for the swing in the second half of the motion and the other muscles of the left leg preparing it in the first half. During *jumping jacks*, *gluteus medius* and *rectus femoris* are activated alternately for both legs. Predicted muscle activations closely follow the ground truth from our dataset, with slight underestimation at the activation peaks.

4.3 CHAPTER CONCLUSION

This chapter addressed two questions about parametric body models for synthetic data generation: (1) *How can parametric body models enable large-scale synthetic data generation, and does pre-training on such data improve real-world action recognition?* (2) *Can parametric body models combined with biomechanical simulation software enable muscle activation datasets at scale, and which architectures effectively learn pose-to-muscle activation mappings?*

Section 4.1 addressed the first question and shows, that parametric body models enable large-scale synthetic data generation through AMARV (Archive Of Motion capture As Rendered Videos), an 800,000-segment dataset which renders SMPL bodies from AMASS motion sequences across four synchronized camera perspectives. This approach generates multi-view synthetic data at scales difficult to gather through real-world multi-camera recording. Critically, experiments demonstrate that models pre-trained on AMARV transfer effectively to real-world action recognition benchmarks in cross-view testing. This shows, that synthetic multi-view data competes with real-world performance and at the same time addresses privacy concerns which are inherent in web-scraped video datasets.

Section 4.2 addressed the second question, establishing that parametric body models combined with OpenSim biomechanical simulation software enable muscle activation datasets at unprecedented scales. MinT generates 10 hours of 402-muscle activation data across 227 subjects. Such scales are unfeasible through real-world EMG measurement, as equipment requirements, subject burden, and cost stand in the way. Experiments on pose-to-muscle activation prediction demonstrate, that deep learning models can effectively learn these mappings from synthetic biomechanical data. Detailed architecture comparisons reveal, which network structures best capture pose-to-activation relationships. This shows, that the combination of parametric models and biomechanical simulation creates training data at scales which suffice to train deep learning models from scratch on physiological tasks.

Together, SynthAct and MinT establish parametric human body models as a versatile foundation for synthetic data generation. The SMPL representation yields controllable synthesis across multiple modalities, i. e. RGB video, depth maps, skeletal pose, and muscle activations, all from the same underlying motion sequences. Both works incorporate this to build generation pipelines. For this reason, the shared SMPL-based motion representation creates potential for cross-dataset multi-modal linkage: body motion serves as an anchor which connects visual, skeletal, and physiological modalities without simultaneous capture. We do not present such experiments here, yet we identify this as a promising direction for future research building on the contributions of this chapter, and pick it up again in Section 7.1.1.

4.3.1 Contributions to SMPL-based Synthetic Data Generation

Contribution 1: *SynthAct* presents a Unity-based rendering pipeline which generates large-scale multi-view synthetic data from motion capture sequences. By rendering SMPL bodies from AMASS motion sequences across four synchronized camera perspectives, the framework produces AMARV, a video dataset which enables multi-view pre-training without expensive multi-camera recording infrastructure. The approach demonstrates effective transfer to real-world action recognition benchmarks and at the same time addresses privacy concerns which are inherent in web-scraped datasets. Thereby, parametric body model rendering becomes a practical approach for scalable synthetic data generation.

Contribution 2: *Muscles in Time (MinT)* bridges computer vision and biomechanics, as it enriches motion capture sequences with simulated muscle activation patterns. By using SMPL's consistent surface correspondence to transfer motion data to OpenSim musculoskeletal models, MinT couples AMASS motion sequences with 402-strand muscle activation simulations. This biomechanical enrichment enables neural networks to learn relationships between visual observations and physiological processes underlying human movement. This shows, that parametric body models can serve as a basis for both, visual and physiological synthetic data generation. The shared SMPL-based motion representation provides potential for cross-dataset multi-modal linkage: body motion serves as an anchor which connects visual, skeletal, and physiological modalities.

Chapters 3 and 4 addressed cross-domain challenges and parametric synthesis methods. The practical deployment of activity recognition systems, however, requires us to address specialized application scenarios. The following chapter explores two such applications: fall detection benchmarking with generative video synthesis, and privacy-preserving activity recognition through avatar-based anonymization with differential privacy guarantees. Both applications demonstrate, how synthetic data generation enables solutions to real-world deployment challenges which extend beyond performance optimization.

SYNTHETIC DATA FOR REAL-WORLD PROBLEMS

The preceding chapters established methods for bridging domain gaps and synthetic data generation. In this chapter we take a *problem first* perspective and then implement synthetic data based solutions in an attempt to meaningfully contribute to and advance a field unrelated to the synthetic data research domain. We explore two separate scenarios in fall detection and privacy preserving vision: (1) *Can we advance the field of human fall detection by generating synthetic data that dominates lab based real-world datasets on out-of-distribution generalization?* (2) *Can we advance privacy preserving action recognition by combining synthetic avatar anonymization with selective differential privacy to provide comprehensive privacy protection for both subjects and environmental background context?* Section 5.1, based on our ECCV 2026 paper [220], addresses the first question through OmniFall, a unified fall detection benchmark comprising our synthetic video dataset based on the Wan 2.2 video diffusion model. Section 5.2, based on our ECCV 2024 workshop paper [222], addresses the second question through MaskDP (Masked Differential Privacy), a hybrid framework which pairs avatar-based anonymization with differential privacy and achieves privacy for both subjects and their personal living environments.

When does synthetic data solve real-world problems that real data collection cannot address? In specialized deployment scenarios, real data collection faces major barriers: staged laboratory data fails to capture authentic real-world variation, whereas the collection of sensitive data from vulnerable populations violates privacy regulations and ethical guidelines. These barriers are not merely practical inconveniences but fundamental limitations that constrain what models can achieve and where they can be deployed.

Fall detection for elderly care exemplifies both challenges: staged datasets feature healthy actors in controlled environments with safety equipment. This creates unknown generalization gaps to genuine accidents, whereas the collection of diverse real fall data from vulnerable elderly populations raises ethical concerns. Privacy-preserving activity recognition in home monitoring similarly faces barriers: video data contains identifiable information about individuals and their living environments, while standard privacy-preserving techniques introduce utility penalties that prevent practical deployment. Both scenarios require solutions beyond what traditional real-data approaches can provide.

5.1 OMNIFALL

This section is based on our paper accepted at ECCV 2026 [220]¹.

Falls remain a leading cause of fatal and non-fatal injury in older adults, claiming about 684,000 lives globally and sending more than 37 million people to medical care each year [282]. Preventing death hinges not only on detecting such falls, but even more so on recognizing the sustained *fallen state*—when a person remains on the ground unnoticed and unable to summon help. “Long-lie” episodes increase one-year mortality significantly [79, 251], underscoring the importance of detection of the fallen.

Computer vision-based fall detection offers potential for automated monitoring in care environments, yet faces critical limitations: (1) systems focus on brief fall events rather than persistent fallen states; (2) models trained on small, homogeneous datasets generalize poorly across environments; (3) collecting fall data from vulnerable elderly populations raises ethical and practical challenges; and (4) no standardized benchmark exists for evaluating performance on genuine accidents or assessing cross-domain generalization.

Existing fall detection datasets [2, 11, 33, 94, 134, 163, 256, 302] share common limitations: collection in single controlled environments, visually homogeneous staging conditions, and incompatible annotation taxonomies that prevent cross-dataset comparison. This lack of diversity produces systems that perform well within their training domain but fail under novel conditions. Moreover, the lack of real-world evaluation protocols prevents assessment of staged-data models’ generalizability to actual accidents.

We introduce *OmniFall*, a unified benchmark comprising three complementary components: *OmniFall-Staged (OF-Sta)* (staged real-world data), *OmniFall-In-the-Wild (OF-ItW)* (genuine accident videos), and *OmniFall-Synthetic (OF-Syn)* (large-scale synthetic data with controlled diversity). The benchmark provides 33 hours of densely annotated single-view footage (82 hours, including multi-view recordings), enabling systematic cross-domain evaluation previously impossible due to annotation heterogeneity.

Why unify existing staged datasets? Individual datasets offer limited environmental and subject diversity, leading to poor cross-dataset generalization. *OmniFall-Staged (OF-Sta)* addresses this by unifying eight public datasets [2, 11, 33, 94, 163, 256, 302] under standardized annotations, providing 14 hours of unique recordings (42 hours multi-view) across 101 subjects and 29 camera views. This aggregation enables cross-dataset, cross-subject, and cross-view evaluation protocols while substantially increasing training diversity compared to any single dataset alone.

Why the need for in-the-wild evaluation? Staged datasets feature healthy actors in controlled settings with safety equipment, creating unknown generalization gaps to real-world deployment. *OmniFall-In-the-Wild (OF-ItW)* establishes the first standardized benchmark for staged-to-wild evaluation, curating 2.6 hours of genuine accident videos from uncontrolled environments with natural camera motion, varied lighting, and authentic falls. This out-of-distribution test set enables systematic assessment of how models trained on staged or synthetic data perform on real accidents.

Why add synthetic data? Collecting diverse fall data from elderly populations and children raises ethical concerns and practical barriers. *OmniFall-Synthetic (OF-Syn)* con-

¹ Reproduced with permission from Springer Nature.

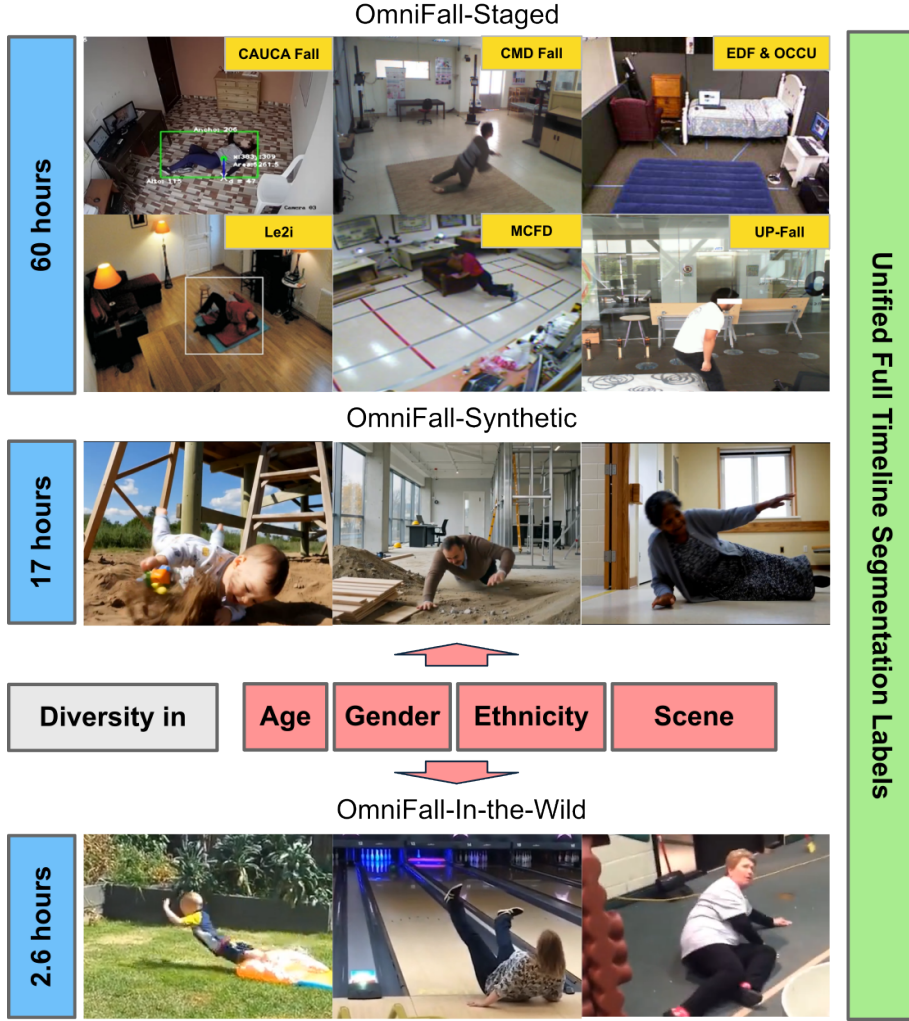


Figure 29: OmniFall comprises OmniFall-Staged, OF-Synthetic, and OF-In-the-Wild with cross-dataset compatible labels.

tributes 17 h of synthetic video, featuring controlled diversity across age, ethnicity, body type, gender, and environments. Each video depicts a unique person-environment combination, providing substantially greater diversity than staged collections while addressing privacy concerns. Significantly, our experiments reveal that training on OF-Syn alone surpasses training on the complete OF-Staged superset for real-world generalization on OF-ItW (Fall F1: 64.2% vs 60.5-61.2%), representing the first large-scale demonstration that synthetic video data can outperform real staged data for fall detection.

Our contributions are threefold. **First**, OF-Staged and OF-In-the-Wild provide a staged-to-wild cross-domain benchmark with over 60 hours of manual frame-level annotations (Figure 30). **Second**, OF-Synthetic adds synthetic-to-wild evaluation and diverse training data with controlled demographic variation. **Third**, comprehensive fine-tuning experiments demonstrate that the staged-to-wild domain gap is larger than the synthetic-to-wild gap, with synthetic data outperforming real staged data for out-of-distribution generalization—suggesting that synthetic video data may surpass existing data collections for real-world human fall detection.

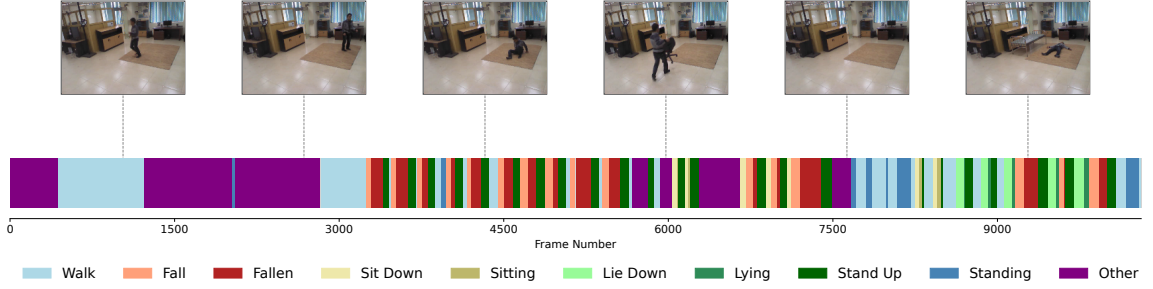


Figure 30: Our annotations on an example video of CMDFall [256].

5.1.1 Dataset

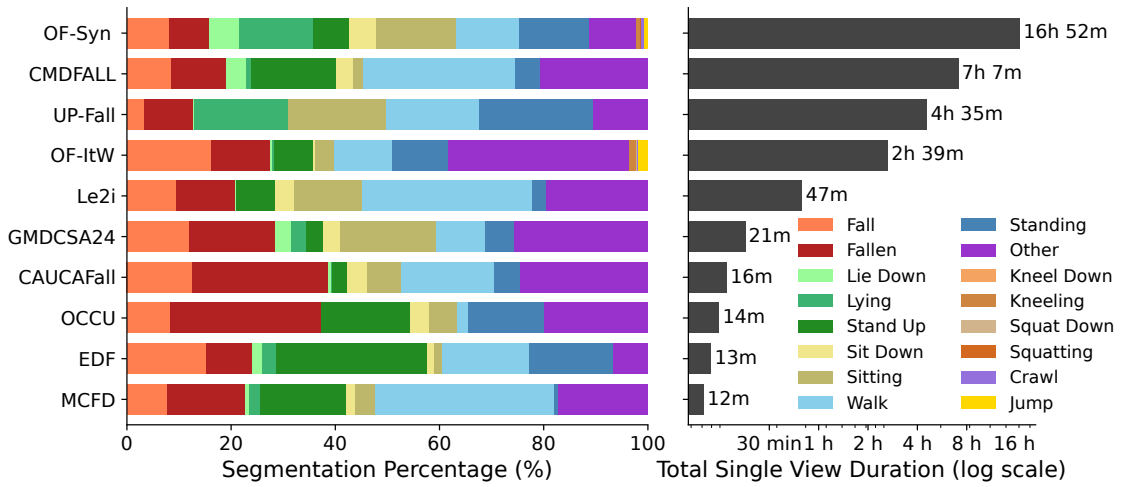


Figure 31: Label distribution and single-view duration per dataset: eight OF-Sta constituents, OF-ItW, and OF-Syn (33h total).

OVERVIEW. **OmniFall** comprises three complementary components (Table 18, Figure 29): (1) **OF-Staged** (eight public datasets) provides controlled scenes, multiple views, and scripted routines; (2) **OF-In-the-Wild (OF-ItW)** curates authentic falls from OOPS [68]; (3) **OF-Synthetic (OF-Syn)** contributes 12k videos generated with Wan 2.2, covering diverse demographics and environments under controlled camera setups. Together they total ~33 h of single-view footage and enable *in-distribution* training on staged data and *out-of-distribution* (OOD) testing on wild and synthetic.

DESIGN PRINCIPLES AND CONTRIBUTIONS. Our goal is *comparability across domains*. To that end we: (i) *Standardize labels* by manually re-annotating all videos to a unified **16-class** timeline taxonomy (four transient actions: *fall*, *sit_down*, *lie_down*, *stand_up*; five locomotion/transition actions: *walk*, *kneel_down*, *squat_down*, *crawl*, *jump*; six static states: *fallen*, *sitting*, *lying*, *standing*, *kneeling*, *squatting*; plus *other*). (ii) *Densify annotations* per frame, so methods can be evaluated on both classification and timeline segmentation. (iii) *Unify splits* by providing cross-subject (CS) and cross-view (CV) partitions for each

Table 18: Datasets in the OmniFall benchmark. The first eight rows constitute OmniFall-Staged (OF-Sta), followed by OmniFall-In-the-Wild (OF-ItW) and OmniFall-Synthetic (OF-Syn).

Dataset	Rec	Rec Dur.	Total Dur.	Views	Syn.	Res.	FPS	Sub
CMDFall [256]	55	7h25m	27h59m	7	✓	640×480	20	50
UP Fall [163]	561	4h37m	9h15m	2	✓	640×480	18	17
Le2i [33]	222	51m	51m	6	✗	320×240	25	9
GMDCSA [2]	160	21m	21m	3	✗	1280×720	30	4
EDF [302]	5	14m	28m	2	✓	320×240	30	5
OCCU [302]	5	14m	28m	2	✗	320×240	30	5
MCFD [11]	24	18m	2h25m	8	✓	720×480	≤30	1
CAUCA [94]	100	16m	16m	1		720×480	23	10
OF-ItW	818	2h39m	2h39m	>818	✗	variable	30	–
OF-Syn	12k	16h53m	16h53m	12k	✗	1280×720	16	–

Rec: Recordings, **Dur:** Duration, **Syn:** synchronized views, **Sub:** subjects.

Recording: single event captured by multiple synchronized camera views.

staged dataset, mirroring original protocols when available and filling gaps otherwise. (iv) *Enable OOD assessment* with a held-out OF-ItW split and controllable OF-Syn splits (random, cross-age, cross-ethnicity, cross-BMI).

ANNOTATIONS AND PROBLEM SETUP.

We cast fall understanding as both *video classification* and *timeline segmentation*.

The taxonomy distinguishes transient actions (*fall, lie down, sit down, kneel down, squat down, stand up, jump, crawl* and *walk*), static states (*fallen, lying, kneeling, squatting, sitting, standing*), facilitating detection of critical fallen states even when fall events are missed and disambiguation from non-critical activities such as lying down. As open-world fallback class we specifically include the class *other* to make sure any timeline can be completely and densely labelled with one of these classes.

The unified 16-class taxonomy enables detecting medically critical *fallen* states even when the fall event is missed, and disambiguating visually similar actions (e.g., *lie_down* vs. *fall*). All data is released with per-frame labels in this taxonomy, plus dataset-specific splits to standardize protocols.

5.1.1.1 OF-Staged (aggregation and splitting).

Many fall detection datasets have been introduced in recent years [2, 5, 11, 12, 33, 37, 42, 94, 134, 163, 256, 302], reflecting the growing interest in automated fall detection. As part of our related work, we conducted a systematic search to identify publicly available video-based datasets to characterize the landscape of the fall detection field. We found that most existing datasets feature staged falls performed by healthy individuals in controlled settings, with no standardized benchmarks for evaluating how models trained on such data perform on real-world falls. In the following, we first review the datasets included

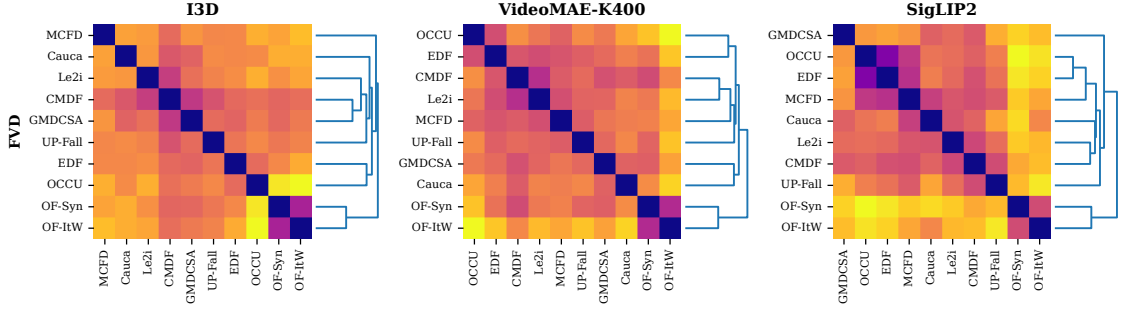


Figure 32: Pairwise Fréchet Video Distances calculated on features from I3D, VideoMAE-K400 and SigLip2, visualised as heat-maps. Darker colors indicate higher similarity (smaller distances) between datasets.

in our OmniFall benchmark, then discuss additional datasets we identified but did not include.

5.1.1.2 Action and fall classification

FALL DETECTION DATASETS IN OMNIFALL-STAGED. To collect existing public fall detection datasets from prior work, we conducted a systematic search in accordance with PRISMA guidelines (5)-(9) [181]. We queried PubMed and Google Scholar using the keywords: [fall] and [detection] and ([survey] or [review]). The search was carried out on 31 August 2023, and we limited the publication period to [2015-2023]. We removed duplicates, including pre-prints followed by their peer-reviewed versions. Subsequently, we conducted an initial manual screening of titles and abstracts to ensure that the selected surveys were relevant. After this initial screening, full texts were retrieved and reviewed for eligibility based on specific inclusion criteria: (1) Surveys with English full texts. (2) Surveys that have undergone peer-review. (3) Surveys listing public RGB-based video fall detection datasets. Of the 98 surveys retrieved, only 9 met all inclusion criteria. The public datasets (1)-(8) extracted from these surveys are listed in Table 18.

(1) The **CMDFall** dataset [256] is a multimodal, multiview collection acquired from seven synchronized Kinect sensors (providing color and depth data) together with two accelerometers. It comprises recordings of 50 subjects performing eight distinct fall types alongside twelve non-fall activities, supporting cross-view and multimodal analysis. (2) **UP Fall** [163] offers a multimodal setup from 17 subjects in a laboratory environment. It includes vision data from a pair of synchronized cameras (640×480 at 18 fps), inertial measurements from accelerometers and gyroscopes worn at different body locations, and context sensors, covering six activities of daily living (ADL) and five fall types. (3) The **Le2i** dataset [33] contains 143 staged falls and 79 ADL actions recorded from a single fixed camera at 640×480 pixels and 25 fps across four environments and six views, stressing various real-world issues such as occlusions and lighting variations. (4) **GMDCSA** [2] delivers high temporal resolution data, capturing 81 fall and 79 ADL clips recorded with a stationary laptop camera at 1280×720 pixels and 30fps in three distinct home settings. Its design emphasizes intra-class variability, with subjects altering clothing and backgrounds between clips. (5) The **EDF** and (6) **OCCU** datasets [302] were acquired in a controlled apartment environment using two Kinect sensors at 320×240 pixels and 30 fps. The EDF

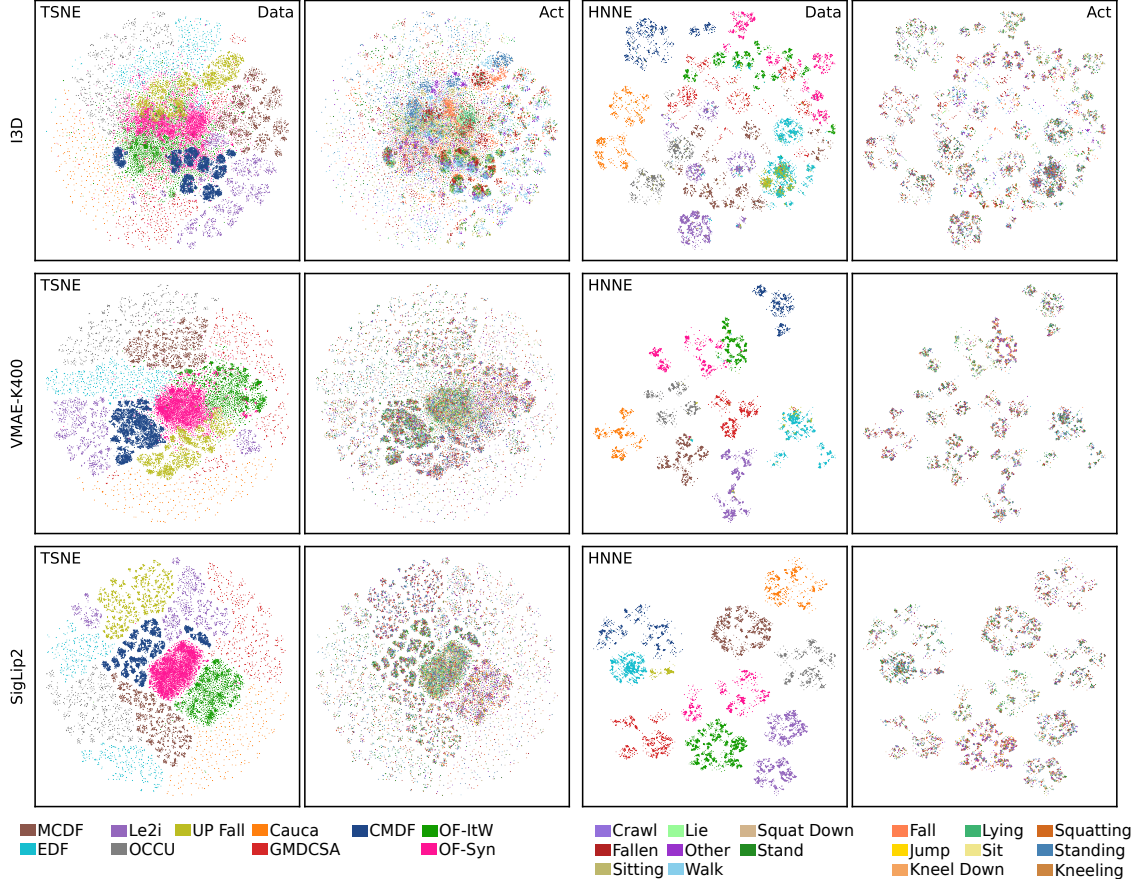


Figure 33: 2D projections, colour-coded by dataset and activity. Rows correspond to feature type, columns show T-SNE and H-NNE clustering

set focuses on 40 non-occluded falls by five subjects performing falls from eight directions, as well as 30 non-fall actions. In comparison, OCCU presents a more challenging scenario in which 30 falls become partially occluded near the end, accompanied by a suite of 80 ADL actions. (7) **The Multiple Camera Fall (MCFD)** dataset [11] leverages eight synchronized IP cameras installed in a fixed indoor setting to capture 22 fall scenarios, interleaved with pre-fall ADL sequences, facilitating 3D reconstructions either explicitly or implicitly. (8) The **CAUCA Fall** dataset [94] provides recordings from 10 subjects in an apartment setting under varying lighting conditions. Captured at 720×480 resolution and 23 fps, it includes both color and infrared modes (in low-light situations) with detailed, frame-wise annotations that report not only fall versus non-fall labels but also spatial metrics, such as distance and angle with respect to the camera.

OTHER DATASETS. We identified several additional fall detection datasets, but excluded them from OF-Staged due to specific limitations. UR Fall [134] was not included due to its limited size. MUVIM [60] offers multimodal data, but all videos are captured from a top-down view. FPDS [156] consists of static images, we only considered full video datasets. While High Quality FSD [12], FallFree [5], SIMPLE [42], and ACT4² [37] align well with our objectives, we were unable to include them due to missing usage licenses or the lack of explicit permission from the authors.

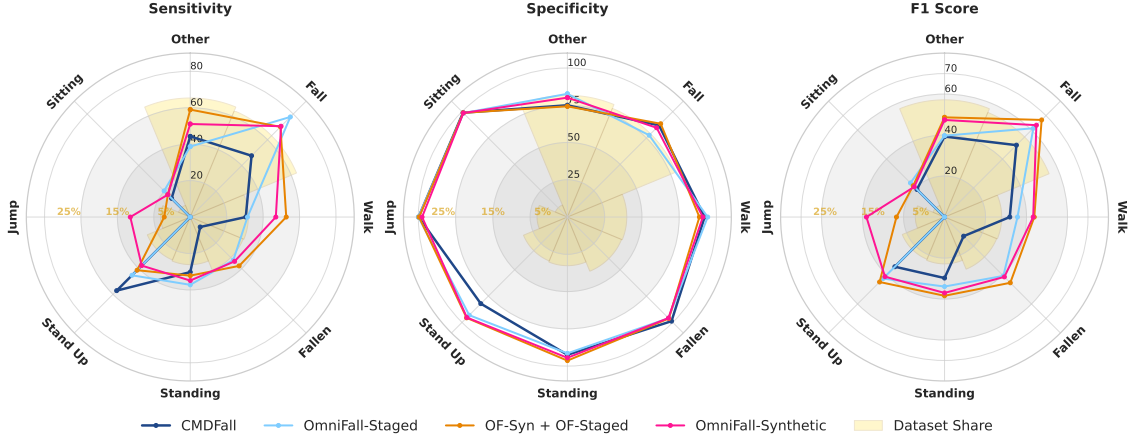


Figure 34: Class-wise *sensitivity*, *specificity* and *F1-score* on OF-In-the-Wild for the eight most frequent classes, cross-domain generalization.

OF-Staged spans 14 h of recordings across eight datasets (Figure 31), with synchronized views excluded from duration accounting but preserved for CV evaluation. We provide CS/CV splits per dataset and a *combined* split (OF-Staged) that supports training once and testing across domains.

CROSS-SUBJECT SPLITS. We implement dataset-specific subject splits, following original protocols where available: CMDFall (following published odd/even protocol: train: odd-IDs 1-49, val: IDs 2/8/18/42/48, test: remaining 20 even-IDs); CaucaFall (train: S1-S7, val: S10, test: S8-S9, with S6/S8 being nighttime recordings); EDF (train: Jianjun/Jinhui/Peter, val: Songsong, test: Zhong); OCCU (train: Jiayan/Wangde/Yangyi, val: Zhangbo, test: Zhouhui); GMDCSA (train: subjects 1/3, val: subject 2, test: subject 4); Le2i (train: subjects 0/1/3/4/6/8, val: subject 5, test: subjects 2/7); UP Fall (train: IDs 1/3/5-10/12-14/17, val: ID 11, test: IDs 2/4/15/16). MCFD’s single subject is used only for training in cross-subject evaluation.

CROSS-VIEW SPLITS. For multi-view datasets, we create camera-based splits: CMDFall (train: camera 1, val: camera 5, test: cameras 2/3/4/6); EDF and OCCU (each using train/val: view 1, test: view 2); GMDCSA (train: subjects 2/3 location with identical room/sensor, val: subject 1 location, test: subject 4 location); Le2i (train: Coffee-room and Lecture-room views, val: Office view, test: Home-room views); MCFD (train: camera 1, val: camera 3, test: cameras 2/4/5/6/7/8); UP Fall (train/val: camera 2, test: camera 1). Single-view CaucaFall participates only in training for cross-view evaluation.

5.1.1.3 What is missing.

The eight datasets in OF-Staged represent over a decade of research effort, spanning 2012 (Le2i [33]) to the present. Collectively, they provide 14 hours of unique recordings across 101 subjects, 29 camera views, and diverse environments; a breadth unattainable from any single source. The substantial financial investment and researcher time required to stage, record, and annotate these corpora reflects the community’s sustained commitment to advancing automated fall detection. OF-Staged aggregates this collective effort

Table 19: **VMAE-K400** [253] trained on CMDFall CS and CV.

Dataset	Split	16-class			Fall Δ			Fallen Δ			Fall $\cup$ Fallen Δ		
		BAcc	Acc	F1	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
CMDFall [256]	CS	83.3	85.4	83.3	91.2	99.3	93.0	92.0	98.7	90.9	95.6	98.9	96.0
UP-Fall [163]	CS	55.6	52.9	41.7	12.3	98.8	21.1	50.9	96.6	62.4	31.6	93.5	45.1
	CV	60.8	57.2	43.8	4.3	99.4	8.1	64.0	95.5	69.9	33.1	93.4	46.2
Le2i [33]	CS	65.9	71.9	52.7	77.3	97.8	79.1	85.7	99.5	90.0	86.0	98.1	89.2
	CV	59.1	72.3	50.2	86.5	95.5	80.0	96.8	94.4	80.0	92.6	88.8	81.3
GMDCSA24 [2]	CS	59.5	54.8	50.4	64.7	89.5	61.1	52.9	100	69.2	61.8	88.1	67.7
	CV	72.0	60.2	59.3	64.7	85.5	56.4	52.9	100	69.2	67.6	86.4	70.8
EDF [302]	CS	52.6	58.6	43.7	6.2	96.4	9.5	75.0	96.3	76.9	50.0	93.5	60.0
	CV	31.2	44.9	30.2	24.4	96.2	34.4	40.0	96.9	49.0	33.3	92.7	44.2
OCCU [302]	CS	44.2	38.6	29.4	0.0	100	0.0	6.2	97.6	10.5	3.1	97.1	5.7
	CV	42.0	41.4	29.9	14.3	99.5	24.5	40.0	96.0	50.0	26.8	94.3	38.9
CaucaFall [94]	CS	62.4	74.5	61.3	66.7	100	80.0	100	94.7	90.0	88.9	96.6	91.4
MCFD [11]	CV	41.7	54.9	38.5	40.6	92.4	47.2	83.1	84.6	59.9	67.0	75.3	61.9
OF-Syn	Ra	32.4	51.9	31.0	33.0	96.5	40.9	10.0	97.4	14.2	27.3	94.1	35.7
OF-ItW		16.0	35.3	14.7	51.1	86.9	52.3	10.0	98.7	16.6	42.9	85.7	50.3

Balanced Accuracy, Sensitivity, Specificity.

into a unified superset with standardized annotations, enabling cross-dataset evaluation and providing training diversity that substantially exceeds individual datasets. However, several fundamental gaps remain.

These datasets comprise *staged* recordings of healthy participants in controlled environments. While valuable for method development, these corpora do not establish how models transfer to *in-the-wild* falls, and they are annotated with heterogeneous label sets and protocols, hampering cross-dataset comparison. Three gaps stand out: (1) *Staging bias*: lab scenes, fixed cameras, and scripted routines dominate. (2) *Annotation heterogeneity*: labels vary from binary fall/no-fall to coarse ADL categories; few offer dense, consistent timelines. (3) *Lack of standardized OOD evaluation*: there is no widely-used protocol to quantify staged→wild generalization.

These gaps motivate the two additional OmniFall components. **OF-In-the-Wild (OF-ItW)** addresses the lack of standardized OOD evaluation by curating genuine accident videos from uncontrolled environments, establishing a test-only benchmark for measuring staged→wild generalization. **OF-Synthetic (OF-Syn)** mitigates ethical and practical barriers to collecting diverse fall data from vulnerable populations by generating synthetic videos with controlled demographic variation (age, ethnicity, body type, environment),

Table 20: **VMAE-K400** [253] trained on OF-Staged (merge of all staged datasets) CS and CV and tested on those splits respectively.

Dataset	Spl	16-class			Fall Δ			Fallen Δ			Fall $\cup$ Fallen Δ		
		BAcc	Acc	F1	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
CMDFall [256]	CS	82.9	85.4	82.9	<u>91.8</u>	<u>99.1</u>	93.0	92.6	98.9	92.0	<u>95.8</u>	98.9	<u>96.2</u>
	CV	71.6	80.9	73.9	94.1	95.7	85.2	91.4	98.7	90.8	96.0	94.6	90.6
UP-Fall [163]	CS	<u>92.5</u>	92.7	83.0	<u>95.6</u>	<u>98.8</u>	<u>95.6</u>	87.7	97.3	88.9	92.1	94.9	92.7
	CV	56.5	89.1	55.1	97.2	95.8	91.3	99.1	93.2	86.8	99.4	86.2	90.2
Le2i [33]	CS	<u>81.4</u>	89.7	<u>72.9</u>	100	100	100	100	100	100	100	100	100
	CV	63.1	79.4	64.8	97.3	99.2	96.0	100	96.4	87.3	100	95.3	93.2
GMDCSA24 [2]	CS	<u>64.5</u>	<u>73.1</u>	<u>66.8</u>	88.2	93.4	81.1	88.2	97.4	88.2	91.2	89.8	87.3
	CV	79.4	79.6	76.2	94.1	94.7	86.5	70.6	98.7	80.0	85.3	93.2	86.6
EDF [302]	CS	56.6	75.8	57.2	82.4	95.5	77.8	100	95.3	89.4	100	92.2	91.6
	CV	45.8	63.0	43.6	67.4	93.8	68.2	83.3	94.6	74.6	87.7	91.7	84.2
OCCU [302]	CS	86.0	90.1	84.6	100	97.6	94.1	100	97.6	94.1	100	94.2	94.1
	CV	59.4	73.6	50.1	88.1	94.4	82.2	95.0	97.0	90.5	92.7	89.8	87.4
CaucaFall [94]	CS	79.5	80.9	77.5	<u>88.9</u>	100	<u>94.1</u>	100	100	100	<u>94.4</u>	100	<u>97.1</u>
MCFD [11]	CV	47.3	67.4	48.9	77.2	92.7	74.1	70.0	94.9	69.3	81.7	88.3	79.4
OF-Syn	CS [†]	32.0	50.9	31.2	46.2	97.3	55.1	32.7	94.1	32.2	43.1	90.9	47.3
	CV [†]	31.2	50.3	29.6	70.8	91.0	58.0	32.7	91.6	28.2	61.0	82.0	51.1
OF-ItW	CS [‡]	20.5	41.4	18.5	77.6	77.6	61.2	33.4	95.4	39.4	74.1	73.3	65.7
	CV [‡]	20.7	41.5	18.2	76.4	77.6	60.5	37.6	95.0	42.6	74.7	72.8	65.7

Balanced Accuracy, Sensitivity, Specificity.

[†] Trained on OF-Sta CS or CV, test on OF-Syn Ra. [‡] Trained on OF-Sta CS or CV, test on OF-ItW.

providing appearance diversity while circumventing privacy concerns inherent in real-world elderly fall data collection.

5.1.1.4 OF-In-the-Wild (curation and labeling).

Starting from OOPS [68], we filter clips tagged as falls, then *manually verify and densely annotate* 818 videos (2.65h; annotation procedure detailed in Appendix A.3), resulting in 1.3k segments with real-world camera motion, occlusion, and clutter. We provide a 10/20/70 split; in this work we treat OF-ItW as *test-only* to measure staged→wild generalization and to leave room for future domain-adaptation research.

Table 21: VMAE-K400 [253] trained on OF-Syn. Bal. Accuracy, Sensitivity, Specificity.

Dataset	Split	16-class			Fall Δ			Fallen Δ			Fall $\cup$ Fallen Δ		
		BAcc	Acc	F1	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
CMDFall [256]	CS	45.8	<u>50.3</u>	34.9	61.0	96.4	66.1	27.6	96.0	34.6	52.4	93.3	60.5
	CV	53.7	57.9	41.0	68.5	97.3	73.9	38.5	96.2	45.8	62.0	94.8	69.9
UP-Fall [163]	CS	48.0	52.9	31.3	<u>63.2</u>	99.3	<u>76.2</u>	2.6	99.5	5.0	33.8	99.0	50.0
	CV	49.6	56.3	29.9	43.1	99.9	60.1	5.3	99.4	9.8	26.0	99.6	41.1
Le2i [33]	CS	53.1	<u>57.1</u>	39.5	59.1	100	74.3	9.5	98.9	16.0	39.5	100	56.7
	CV	40.6	52.8	30.1	48.6	100	65.5	0.0	99.2	0.0	29.4	100	45.5
GMDCSA24 [2]	CS	64.5	60.2	50.3	52.9	97.4	64.3	23.5	100	38.1	41.2	98.3	57.1
	CV	73.4	61.3	48.4	76.5	97.4	81.2	11.8	100	21.1	47.1	98.3	62.7
EDF [302]	CS	33.7	43.8	24.2	35.3	92.8	38.7	47.6	93.5	52.6	47.4	85.6	52.2
	CV	31.0	41.3	22.7	32.6	94.3	40.6	60.0	87.9	48.0	64.4	86.7	65.3
OCCU [302]	CS	54.1	53.5	30.5	37.5	100	54.5	0.0	100	0.0	18.8	100	31.6
	CV	51.2	52.7	31.0	28.6	100	44.4	0.0	100	0.0	14.6	100	25.5
CaucaFall [94]	CS	60.1	55.3	40.9	77.8	94.7	77.8	22.2	100	36.4	50.0	93.1	62.1
MCFD [11]	CV	31.1	40.5	22.9	41.5	93.1	48.5	7.1	98.5	12.3	29.4	90.9	39.8
OF-Syn	Ra	<u>59.0</u>	81.4	<u>60.5</u>	96.2	95.1	81.6	67.3	97.0	66.4	90.1	92.6	81.1
OF-ItW		<u>20.6</u>	<u>46.7</u>	<u>21.2</u>	72.0	<u>84.6</u>	<u>64.2</u>	<u>35.0</u>	96.1	42.3	69.5	81.8	67.9

Balanced Accuracy, Sensitivity, Specificity.

5.1.1.5 OF-Synthetic (controlled diversity).

OF-Syn contributes 12k videos (17 h) generated with Wan 2.2 [271] at 1280×720 and 15 fps, each depicting a unique individual and scene. Prompts vary age, gender presentation, body type (BMI band), ethnicity and skin tone, environment category, and camera placement. While clip length and temporal complexity is low, the set offers *factorial control* over appearance and viewpoint. Domain-gap analyses (Section 5.1.2.1) show OF-Syn sits closer to OF-ItW in feature space than many individual staged datasets, supporting its use for sim→real studies. Detailed generation methodology, including prompt construction and demographic diversity design, is provided in the appendix A.2.

5.1.2 Experiments

Our evaluation comprises three complementary experiments using different methodologies. For domain shift analysis (Section 5.1.2.1), we extract frozen motion oriented features from I3D [30] and VideoMAE-K400 [253] and more appearance-based features from SigLIP2 [258] to analyze dataset similarities and feature clustering. For action and fall clas-

Table 22: **VMAE-K400** [253] trained on OF-Sta(CS)+OF-Syn(Ra). Bal. Accuracy, Sensitivity, Specificity.

Dataset	Split	16-class			Fall Δ			Fallen Δ			Fall $\cup$ Fallen Δ		
		BAcc	Acc	F1	Se	Sp	F1	Se	Sp	F1	Se	Sp	F1
CMDFall [256]	CS	<u>83.0</u>	<u>85.4</u>	<u>83.3</u>	<u>92.6</u>	<u>98.9</u>	<u>92.6</u>	<u>90.9</u>	<u>98.9</u>	<u>91.0</u>	<u>96.3</u>	<u>98.8</u>	<u>96.3</u>
UP-Fall [163]	CS	<u>92.6</u>	<u>92.0</u>	<u>82.4</u>	<u>96.5</u>	<u>99.3</u>	<u>96.9</u>	<u>79.8</u>	<u>98.0</u>	<u>85.4</u>	<u>88.2</u>	<u>96.3</u>	<u>91.4</u>
Le2i [33]	CS	<u>83.2</u>	<u>89.7</u>	<u>82.4</u>	<u>100</u>	<u>99.4</u>	<u>97.8</u>	<u>100</u>	<u>100</u>	<u>100</u>	<u>100</u>	<u>99.4</u>	<u>98.9</u>
GMDCSA24 [2]	CS	<u>80.3</u>	<u>76.3</u>	<u>75.0</u>	64.7	<u>94.7</u>	68.8	<u>82.4</u>	<u>97.4</u>	<u>84.8</u>	<u>76.5</u>	<u>91.5</u>	<u>80.0</u>
EDF [302]	CS	<u>55.1</u>	<u>74.2</u>	<u>54.2</u>	<u>82.4</u>	<u>95.5</u>	<u>77.8</u>	<u>100</u>	<u>95.3</u>	<u>89.4</u>	<u>100</u>	<u>92.2</u>	<u>91.6</u>
OCCU [302]	CS	<u>84.4</u>	<u>91.1</u>	<u>81.9</u>	<u>100</u>	<u>97.6</u>	<u>94.1</u>	<u>100</u>	<u>98.8</u>	<u>97.0</u>	<u>100</u>	<u>95.7</u>	<u>95.5</u>
CaucaFall [94]	CS	<u>76.8</u>	<u>80.9</u>	<u>76.2</u>	<u>100</u>	<u>100</u>	<u>100</u>	<u>100</u>	<u>97.4</u>	<u>94.7</u>	<u>100</u>	<u>96.6</u>	<u>97.3</u>
OF-Syn	Ra	<u>60.5</u>	<u>81.2</u>	<u>62.4</u>	<u>96.2</u>	<u>95.8</u>	<u>83.6</u>	<u>68.7</u>	<u>96.7</u>	<u>66.0</u>	<u>89.5</u>	<u>92.7</u>	<u>81.0</u>
OF-ItW		<u>21.4</u>	<u>47.6</u>	<u>21.9</u>	<u>72.9</u>	<u>85.4</u>	<u>65.6</u>	<u>33.1</u>	<u>96.3</u>	<u>41.0</u>	<u>68.2</u>	<u>82.4</u>	<u>67.5</u>

sification (Section 5.1.1.2), we fine-tune VideoMAE-K400 end-to-end on our benchmark datasets. Timeline segmentation follows established protocols using I3D and SigLIP2 features (Section 5.1.2.2). Unless reported differently, we use cross-subject splits for CMDFall and OF-Sta and random split for OF-Syn and OF-ItW by default.

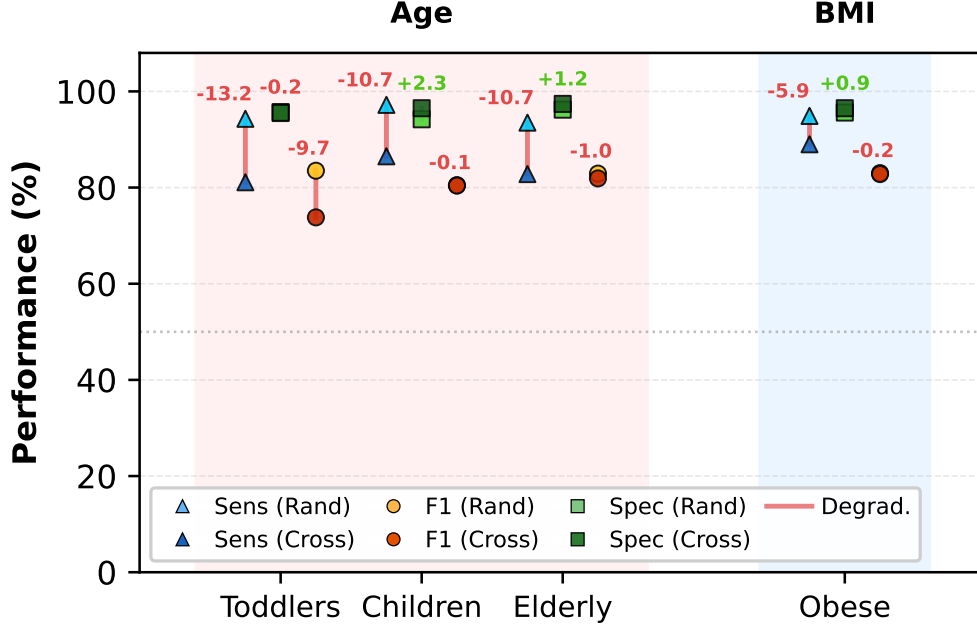
5.1.2.1 Domain shift analysis

Figure 32 shows pairwise Fréchet Video Distances (FVD) between datasets on different feature types as well as hierarchical clustering results on the underlying distance matrix.

The versatile CMDFall dataset demonstrates high similarity (darker colors) to most other staged datasets, highlighting its versatility as a training source. OF-Syn and OF-ItW display very strong similarity, presumably attributable to both datasets containing a large number of outside activities, which is completely missing in any staged dataset. Datasets in OF-Staged show substantial dissimilarity to OF-In-the-Wild on VMAE-K400, corresponding to the OOD performance gaps in Section 5.1.1.2.

FEATURE-SPACE VISUALIZATION. To analyze domain shift and action structure recoverability, we visualize frozen embeddings from three backbones (I3D, VideoMAE-K400, and SigLIP2) using t-SNE [153] and h-NNE [218]. Per domain, we randomly sample 5,000 representatives (five-frame means for temporal smoothing) for projection. In Figure 33, first two columns show t-SNE colored by *domain* and *action*; the last two columns show H-NNE with the same colorings and its induced hierarchical clustering view.

Domain gap. Across all backbones, domains form largely disjoint clusters with minimal overlap; this separation is most pronounced for frame-space SigLIP2 and clearly visible for VideoMAE and I3D as well. The result confirms a strong dataset shift within staged, in-the-wild, and synthetic videos.

Figure 35: OF-Syn demographic splits; *sensitivity, specificity, F1*

h-NNE provides both a 2D embedding and a multi-scale partitioning. This stabilizes cluster contours, exposes local $\rightarrow$ global organization, and yields hierarchical cluster labels. Overall, the plots jointly (i) make the cross-domain gap explicit and (ii) indicate how much action structure is already present in frozen features—motivating both domain adaptation and temporal modeling on our benchmark.

EXPERIMENTAL SETUP We evaluate four fine-tuning configurations for VMAE-K400 [253] to understand generalization effects of training data sources. First, we train on the largest and most comprehensive pre-existing fall dataset **CMDFall** alone (7h), using separate models for cross-subject (CS) and cross-view (CV) splits. Second, we train on our **OF-Staged** superset (14h, 8 datasets), again using CS and CV models to evaluate both generalization protocols. Third, we train on **OF-Synthetic** using its random (Ra) split (17h). Fourth, we combine **OF-Staged (CS)** with **OF-Synthetic (Ra)**, assessing whether synthetic data complements real staged data.

For metrics, we report accuracy and macro-F1 for the 16-class task, alongside sensitivity (Se), specificity (Sp), and F1-score for the critical binary subtasks *Fall* and *Fallen*. These fall-specific metrics are most relevant for deployment, as they measure detection of medically critical events. Macro-F1 scores tend to be lower due to the natural class imbalance across 16 categories (Figure 34 illustrates this effect). Tables 19–22 provide detailed results. Across these tables, we mark best and second-best results with green and orange.

TRAINING ON CMDFALL Table 19 presents results when training exclusively on CMDFall, the most comprehensive individual staged dataset. The model achieves strong in-distribution performance on cross-subject (83.3% acc, Fall F1: 93.0%) and cross-view (80.9% acc, Fall F1: 85.2%) splits. Generalization to OF-ItW remains limited (Fall F1: 52.3%),

Table 23: ASQuery [82] timeline segmentation results.

Dataset	Feat.	F1@10	F1@25	F1@50	Edit	Acc
OF-Staged	I3D	80.11	78.34	69.84	84.13	78.74
	SL2	82.13	80.64	72.70	81.90	80.19
→ OF-ItW	I3D	44.62	40.28	26.37	43.11	59.29
	SL2	39.15	33.95	20.26	36.42	43.32
OF-Synthetic	I3D	78.74	77.77	72.17	78.28	77.22
	SL2	84.17	83.32	78.11	83.71	82.71
→ OF-ItW	I3D	33.74	24.88	14.17	32.65	25.23
	SL2	53.92	40.56	27.60	49.97	59.39
OF-Sta	I3D	77.68	75.64	66.51	77.32	76.50
	SL2	82.48	80.99	73.24	82.51	79.88
∪ OF-Syn	I3D	77.46	76.65	70.62	76.46	75.42
	SL2	83.10	82.24	76.90	82.41	81.15
→ OF-ItW	I3D	42.53	33.99	19.52	39.43	42.37
	SL2	52.34	45.66	28.61	48.07	59.83

suggesting that even the largest staged dataset provides insufficient diversity for robust real-world deployment.

TRAINING ON OF-STAGED Expanding training to the unified OF-Staged superset substantially improves out-of-distribution performance. Despite similar data volume to CMD-Fall alone (61.5h vs 49.8h multi-view), the diversity of eight datasets with varied environments, subjects, and viewpoints improves OF-ItW results to 60.5-61.2% Fall F1 while maintaining strong in-distribution accuracy (CMDFall CS: 85.4%; Table 20). This demonstrates that dataset diversity, rather than simply data quantity, drives better real-world generalization.

TRAINING ON OF-SYNTHETIC Table 21 reveals a significant finding: training exclusively on synthetic OF-Syn achieves the best single-source OF-In-the-Wild performance (Fall F1: 64.2%), surpassing the OF-Staged superset of real staged data (60.5-61.2%). OF-Syn’s controlled demographic diversity appears to bridge the domain gap to in-the-wild settings more effectively than staged data, despite being fully synthetic. This aligns with the domain shift analysis (Section 5.1.2.1), where OF-Syn exhibits closer feature-space similarity to OF-ItW than individual staged datasets.

TRAINING ON OF-STAGED+OF-SYNTHETIC Combined training on OF-Staged(CS) and OF-Synthetic(Ra) yields the best overall performance (Table 22), reaching 65.6% Fall F1 on

OF-In-the-Wild while maintaining 85.4% accuracy on CMDFall CS. The improvement over OF-Syn alone is marginal (+1.4 F1 points), suggesting that synthetic data with controlled diversity already captures most of the variation needed for real-world generalization.

RADAR-PLOT VISUALIZATIONS We present class-wise results on the eight most frequent classes in Figure 34. As expected, F1-Score and Sensitivity positively correlate with class frequency, while specificity slightly negatively correlates. All our dataset contributions provide models that strongly dominate over CMDFall training. OF-Syn training improves over OF-Staged training with OF Staged+Syn slightly improving on *Fall* and *Fallen* scores.

CROSS-ATTRIBUTE SPLITS We provide a short evaluation of cross-age and cross-BMI training on OF-Syn in Figure 35 and report metrics for the class *Fall*. Here we train on the age groups *young adult* and *middle aged* and evaluate on *toddlers*, *children*, and *elderly*. We observe a marked drop in sensitivity, and for toddlers, also in F1 score. Slight negative effects are also visible for cross-BMI training.

5.1.2.2 Timeline segmentation

Thanks to its dense labels and its multi-domain coverage, OmniFall supports OOD timeline segmentation studies.

To demonstrate, we train ASQuery [82] at 10 fps using two frozen feature families: I3D [30] (motion-centric) and SigLIP2 (SL2; appearance-centric). We consider three training sets—**OF-Staged**, **OF-Synthetic**, and their union **OF-Sta $\cup$ OF-Syn**—and evaluate on in-distribution staged test splits and out-of-distribution **OF-ItW**. We report F1@IoU (10/25/50), normalized Edit, and frame accuracy (Table 23).

FEATURE COMPARISON. Across all training and test regimes, **SL2 outperforms I3D**. On OF-Sta (in-distribution), SL2 attains F1@50 of 72.70–78.11 % vs. I3D’s 66.51–72.17 %. The gap persists on OF-ItW: 27.60–28.61 % (SL2) vs. 14.17–26.37 % (I3D). This suggests appearance-centric embeddings are less sensitive to the synthetic→real shift.

TRAINING DATA COMPARISON. For **SL2**, training on **OF-Synthetic** alone achieves F1@50 of 27.60 % on OF-ItW, surpassing **OF-Staged** (20.26 %) and approaching **OF-Sta $\cup$ OF-Syn** (28.61 %). For **I3D**, the pattern reverses: **OF-Staged** yields 26.37 % vs. **OF-Synthetic** at 14.17 % on OF-ItW. This indicates *synthetic data supplies appearance diversity that benefits frame-appearance features*, while *staged data better supports motion dynamics captured by I3D*. **OF-Sta $\cup$ OF-Syn** on SL2 provides best overall robustness.

Overall, OmniFall exposes a challenging staged→wild transfer: models that score highly in controlled settings lose precision in OF-ItW, where appearance, motion, and camera factors vary widely. Coupling dense framewise timelines with complementary domains (staged motion structure, synthetic appearance diversity, real wild variability) under a unified protocol, OmniFall provides a rigorous, reproducible benchmark, testing timeline segmentation methods that must balance *boundary precision* and *domain robustness*.

5.2 AVATAR-BASED ANONYMIZATION WITH PRIVACY GUARANTEES

This section is based on our paper presented at the 2nd International Workshop for Privacy-Preserving Computer Vision at ECCV 2024 [222]. Significant parts of this work, including the concept of selective application of differential privacy were developed during a research internship at Sony Europe B.V., Hoofddorp, Schlieren/Switzerland Branch.

How can synthetic data generation be used for anonymizing activity recognition without risking information leakage from background or context areas? This section introduces MaskDP (Masked Differential Privacy), a hybrid privacy framework combining avatar-based anonymization with differential privacy. By decomposing videos into synthetic avatar regions (assumed non-identifying through visual appearance) and remaining sensitive spatio-temporal regions, MaskDP applies differential privacy exclusively to designated sensitive areas, achieving formal (ϵ, δ) -differential privacy guarantees for those regions while significantly reducing the utility penalty typically associated with standard differential privacy. This approach does not provide end-to-end differential privacy for complete videos; instead, it offers selective formal guarantees for explicitly marked sensitive regions, while relying on synthetic avatar anonymization for subject protection.

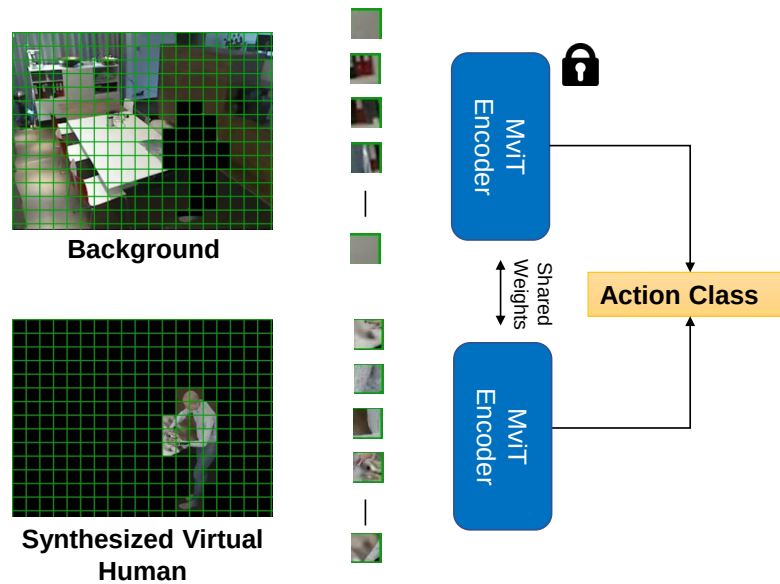


Figure 36: In our experiments we apply MaskDP on anonymized datasets containing non-sensitive synthetic humans. Only real-world data is considered private and needs protection.

In the past two decades, human activity recognition [30, 62, 121, 274] in videos has gained significant attention for its wide applications in video retrieval, behavior analysis, surveillance, and video understanding tasks. Despite notable progress in this field, concerns about data privacy have escalated as videos often contain sensitive information, such as human faces and other personal attributes. Safeguarding individuals' privacy rights in video understanding is not only a legal and ethical obligation but also a necessity

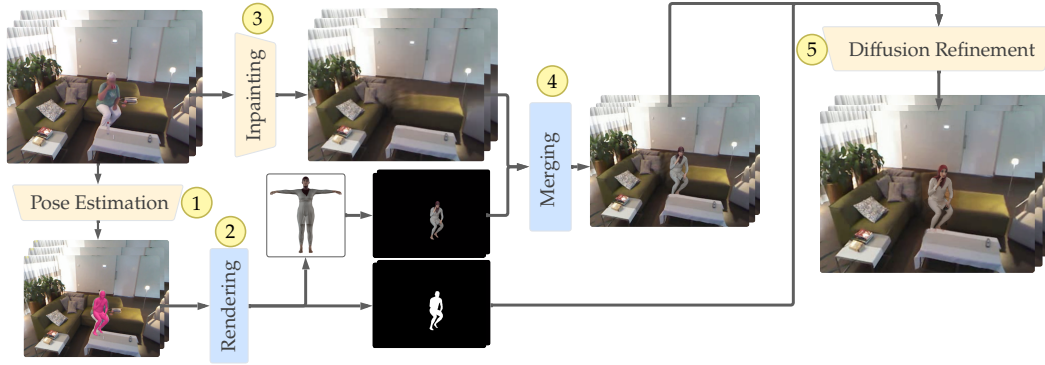


Figure 37: Overview of our video anonymization framework. The process includes: ① 3D pose estimation, ② Synthetic human rendering using SMPL-H, ③ Background inpainting to remove original persons, ④ Merging of synthetic avatar and inpainted background, and ⑤ Diffusion-based refinement. This pipeline effectively anonymizes individuals while preserving video quality and context.

to build up user trust under regulatory frameworks, such as GDPR [202], CCPA [118], and EU AI Act [184]. Without rigorous privacy measures, there are substantial risks of misuse, emphasizing the need for effective privacy-preserving techniques.

Recent efforts to integrate privacy-preserving measures into video-based activity recognition can be categorized into three approaches: obfuscation-based methods [55, 78], training on purely synthetic data [44, 56, 109, 266], and training with Differential Privacy [152]. However, each approach faces limitations: obfuscation methods typically lack formal guarantees, while standard DP incurs substantial utility penalties when treating entire videos as protected units.

Modern advances in 3D human pose and shape estimation [89, 150, 211] enable automatic generation of synthetic avatar replacements for real humans without manual intervention, providing scalable anonymization of human subjects in video data. However, such rendering-based methods cannot reliably synthesize environmental contexts, background interactions, or scene-specific details that remain critical for recognition performance. This technical asymmetry (reliable human synthesis but limited context generation) creates a challenge for privacy-preserving video analysis.

While much existing anonymization work focuses exclusively on protecting human identity through facial blurring, person removal, or avatar replacement, background information itself poses significant re-identification risks that such approaches overlook. Living environments, room layouts, personal objects, and contextual cues can reveal identifying information about individuals even when human subjects are fully anonymized. Home monitoring scenarios (such as those in the Toyota Smarthome dataset) particularly exemplify this privacy gap, where environmental context may be as identifying as the individuals themselves.

Our framework addresses these challenges through complete pixel-level coverage using two complementary privacy mechanisms. Every pixel in the video is protected by exactly one of two methods: (1) human subject regions are replaced with synthetic avatars generated from 3D pose estimation, removing identifying visual appearance; (2) all remaining

regions (background, objects, environmental context) receive formal differential privacy protection during model training (Figure 36). This decomposition ensures comprehensive privacy coverage while preserving the contextual information that isolated avatars would discard (Table 24 demonstrates that avatar-only approaches lose up to 84% performance on context-dependent datasets).

Alternative approaches prove insufficient for achieving both privacy and utility. Applying differential privacy to entire videos results in severe utility degradation, with prior work [152] demonstrating that even at lenient privacy budgets ($\epsilon = 5$), standard DP substantially reduces performance. Conversely, avatar-based anonymization without background preservation discards critical context. Neither approach alone satisfies the dual requirements of comprehensive privacy and task utility.

A critical property distinguishing our approach from detector-based anonymization methods is the exact pixel-level knowledge provided by rendering-based avatar generation. The rendering process provides perfect binary masks indicating precisely which pixels contain synthesized avatars: when pose estimation succeeds and rendering occurs, we know exactly which pixels were rendered; when pose estimation fails, no rendering occurs and those pixels remain as original background requiring differential privacy protection. This complete knowledge eliminates coverage gaps that could arise from detector failures in segmentation-based approaches; every pixel is definitively classified as either rendered (avatar region) or non-rendered (background region requiring DP). Unlike person segmentation networks that may produce uncertain or incorrect masks on difficult poses, unusual viewpoints, or occluded subjects, our masks reflect the ground truth of what was actually synthesized, ensuring that privacy mechanisms are applied across all frames and pixels without exception. Applications relying on imperfect detectors risk privacy leakage when segmentation errors leave sensitive regions unprotected.

Based on this complete coverage framework, we introduce Masked Differential Privacy (MaskDP), a novel relaxation of standard Differential Privacy that leverages the perfect pixel-level masks from rendering. MaskDP decomposes videos into public tokens (synthetic avatar regions assumed non-identifying through visual appearance) and private tokens (remaining regions requiring formal protection). By applying differential privacy exclusively to private tokens, MaskDP achieves higher utility than standard DP (which protects entire videos indiscriminately) while maintaining formal (ϵ, δ) -differential privacy guarantees for all non-anonymized regions.

The MaskDP framework explicitly does not provide differential privacy guarantees for full video samples. Rather, MaskDP provides formal (ϵ, δ) -differential privacy guarantees exclusively for explicitly defined privacy-sensitive spatio-temporal regions marked by mask function M . Non-protected regions rely on anonymization techniques (avatar replacement). This hybrid privacy model combines the utility benefits of selective protection with formal guarantees for designated sensitive regions.

5.2.1 Preliminaries

Differential privacy (DP) [67] serves as a gold standard for ensuring the privacy of individuals' sensitive data. In simple terms, DP ensures that the output of an algorithm remains consistent regardless of whether a specific individual's data is included in the dataset. This means that an external observer who has access to the entire dataset except

for one individual’s information cannot determine with high probability whether that specific individual’s data is part of the input. We provide a formal definition of DP below.

Definition 5.1 (Differential Privacy [67]). *Given $\epsilon > 0$ and $\delta > 0$, a randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ) -differential privacy, if for all possible pairs of adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ differing by at most one record, denoted as $\mathcal{D} \sim \mathcal{D}'$, and for any possible set of outputs $S \subseteq \text{Range}(\mathcal{A})$, we have:*

$$\Pr[\mathcal{A}(\mathcal{D}) \in S] \leq e^\epsilon \Pr[\mathcal{A}(\mathcal{D}') \in S] + \delta.$$

The parameter ϵ is referred to as the “privacy budget” or “privacy cost,” which balances the trade-off between privacy and utility. A smaller ϵ provides stronger privacy guarantees but might result in reduced utility. The parameter δ serves as a small failure probability, ensuring the robustness of the privacy mechanism. In the context of deep learning, DP techniques are employed during the model training process to protect the privacy of training data. The most popular algorithm, Differentially Private Stochastic Gradient Descent (DP-SGD) [1], involves adding carefully calibrated noise to the clipped gradients computed during training, obscuring the contribution of individual data points.

THREAT MODEL. Our framework addresses privacy risks in activity recognition where the training videos contain sensitive information in specific spatio-temporal regions. This framework protects the training dataset under the assumption that an adversary has access to the trained model. The privacy guarantees ensure that observing the model’s behavior does not reveal whether specific training samples were included in the dataset. This is *not* a method for safe data release; rather, it enables training models on sensitive data while protecting members of the training set from information extraction attacks by parties who deploy or interact with the trained model.

We assume an adversary with full access to the trained model (white-box access including model parameters and architecture) and knowledge of the training algorithm, but without direct access to the training dataset itself. The adversary may possess auxiliary information about individuals or contexts, and may attempt membership inference attacks (determining whether a specific video was in the training set) or attribute inference attacks (inferring sensitive attributes about individuals or environments from the model). We assume activity labels themselves are not privacy-sensitive and do not require protection; our privacy guarantees apply only to the video content (pixels) within designated private regions.

Our privacy guarantees protect against such inference attacks on regions marked as private by mask function M , providing formal (ϵ, δ) -differential privacy bounds on information leakage from those specific regions. This means an adversary observing the trained model cannot reliably determine whether any specific private token (e.g., background pixel) from a single video frame was present in the training data, with probability bounded by $e^\epsilon + \delta$.

However, our threat model explicitly excludes several attack scenarios: (1) behavioral biometric re-identification through gait or movement patterns preserved in synthetic avatars, (2) privacy breaches during the preprocessing/anonymization stage that operates on raw data, and (3) composition attacks that combine information from avatar movements with differentially-private background information. These excluded scenarios represent known limitations of our hybrid privacy approach rather than failures of the differential privacy mechanism itself.

5.2.2 Masked Differentially Private Training

While differential privacy protects information of dataset records, defining the composition of a record significantly influences both privacy and utility. Previous methods in the field of computer vision apply DP on the individual image or video level. For masked differential privacy we assume additional knowledge about the spatial or temporal region of sensitive information *within* a dataset sample and limit the application of differential privacy to that region. This allows us to decompose visual data into sets of tokens $\mathcal{T}_{\text{pr}}$ and $\mathcal{T}_{\text{pu}}$ where only $\mathcal{T}_{\text{pr}}$ are considered private while $\mathcal{T}_{\text{pu}}$ are non-sensitive public tokens.

Hence, we introduce Masked Differential Privacy (MaskDP) to formally analyze privacy guarantees of an algorithm that offers exclusive protection of specific components of individual data points, rather than the entire record. To this end, we first define the notion of masked adjacency, which extends Definition 5.1 and is the key concept in our definition of MaskDP:

Definition 5.2 (Masked Adjacency). *Let $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ be a dataset of size N , where each record $x^{(i)} = \{x_1^{(i)}, \dots, x_K^{(i)}\}$ is partitioned into K tokens. Let $M : [N] \times [K] \rightarrow \{0, 1\}$ be a mask function indicating which tokens are protected, where $M(i, k) = 1$ if and only if $x_k^{(i)}$ is protected. Consider another dataset $\mathcal{D}' = \{x'^{(1)}, \dots, x'^{(N)}\}$, and for all $i \in \{1, \dots, N\}$, let $S_i = \{k \in [K] \mid M(i, k) = 1\}$. Then $\mathcal{D}$ is masked-adjacent to $\mathcal{D}'$, denoted as $\mathcal{D} \sim_M \mathcal{D}'$, if and only if $\exists i \in [N]$ such that $\forall j \in [N], j \neq i \mid x^{(j)} = x'^{(j)}$ and $\forall k \in [K] \setminus S_i \mid x_k^{(i)} = x_k'^{(i)}$ and $\exists l \in S_i$ such that $x_l^{(i)} \neq x_l'^{(i)}$.*

Intuitively, when two datasets $\mathcal{D}$ and $\mathcal{D}'$ are masked adjacent, it implies that they are almost identical except in one record, where a change occurs only in the tokens that are specifically marked as sensitive by the mask function M .

Based on the above definition, we can now formally define MaskDP (Definition 5.3):

Definition 5.3 (Masked Differential Privacy (MaskDP)). *Given $\epsilon > 0$, $\delta > 0$, and mask function M , a randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ) -MaskDP, if for all possible pairs of masked adjacent datasets $\mathcal{D} \sim_M \mathcal{D}'$ under M , and for any possible set of outputs $S \subseteq \text{Range}(\mathcal{A})$, we have:*

$$\Pr[\mathcal{A}(\mathcal{D}) \in S] \leq e^\epsilon \Pr[\mathcal{A}(\mathcal{D}') \in S] + \delta.$$

This definition formalizes the idea that privacy protection targets specific components within individual records, as opposed to protecting each record in its entirety.

REMARK. Note that the only distinction between standard DP and MaskDP lies in the concept of dataset adjacency. As a result, fundamental properties of DP, including post-processing, composition, and privacy amplification theorems [1, 67, 279], also hold for MaskDP as they do not depend on any specific definition of adjacency. Moreover, the above definition readily extends to other variants of DP, such as Renyi DP [169], only by considering Definition 5.2 as the notion of adjacency between datasets.

LEARNING WITH MASKDP Algorithm 1 presents our proposed framework for implementing MaskDP-SGD, designed for training robust privacy-preserving neural networks on sensitive datasets.

Algorithmus 1 : MaskDP-SGD

Input : Dataset $\mathcal{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_N, \mathbf{y}_N)\}$; Task model $\mathcal{M}_\Theta$ with parameters Θ ; Video function f_{tok} ; Loss function $\mathcal{L}$; Number of epochs T ; Batch size B ; Gradient clipping threshold C ; noise standard deviation σ ; Learning rate η ;

Result : MaskDP Trained model $\mathcal{M}_\Theta$

```

1 Initialize  $\Theta$ 
2 for  $t \in [T \cdot \frac{N}{B}]$  do
3   Sample a batch  $\mathcal{B}_t$  by selecting each  $i \in [N]$  independently with probability  $B/N$ 
4   for  $i \in \mathcal{B}_t$  do
5     ▶ Tokenize video into private and public tokens.
6      $\mathcal{T}_{\text{pr}}^{(t,i)}, \mathcal{T}_{\text{pu}}^{(t,i)} \leftarrow f_{\text{tok}}(\tilde{\mathbf{x}}_i, \mathbf{r}_i)$ 
7     ▶ Compute gradients for both token sets
8      $\mathbf{g}_{\text{pu}}^{t,i} \leftarrow \nabla_{\Theta} \mathcal{L}(\mathcal{M}_\Theta(\mathcal{T}_{\text{pu}}^{(t,i)}), \mathbf{y}_i)$ 
9      $\mathbf{g}_{\text{pr}}^{t,i} \leftarrow \nabla_{\Theta} \mathcal{L}(\mathcal{M}_\Theta(\mathcal{T}_{\text{pr}}^{(t,i)}), \mathbf{y}_i)$ 
10    ▶ Clip gradients of private tokens
11     $\tilde{\mathbf{g}}_{\text{pr}}^{t,i} \leftarrow \mathbf{g}_{\text{pr}}^{t,i} \cdot \min(1, \frac{C}{\|\mathbf{g}_{\text{pr}}^{t,i}\|_2})$ 
12  end
13  ▶ Compute total gradient, add noise, and update parameters
14   $\mathbf{g}_{\text{pu}}^t \leftarrow \sum_{i=1}^{|\mathcal{B}_t|} \mathbf{g}_{\text{pu}}^{t,i}, \quad \tilde{\mathbf{g}}_{\text{pr}}^t \leftarrow \sum_{i=1}^{|\mathcal{B}_t|} \tilde{\mathbf{g}}_{\text{pr}}^{t,i} + \mathcal{N}(0, \sigma^2 \mathbf{I})$ 
15   $\mathbf{g}_{\text{total}}^t \leftarrow \frac{1}{|\mathcal{B}_t|} (\mathbf{g}_{\text{pu}}^t + \tilde{\mathbf{g}}_{\text{pr}}^t)$ 
16   $\Theta \leftarrow \Theta - \eta \mathbf{g}_{\text{total}}^t$ 
17 end
18 return  $\mathcal{M}_\Theta$ 

```

The algorithm computes the gradients of the model's loss function for both $\mathcal{T}_{\text{pu}}$ and $\mathcal{T}_{\text{pr}}$ token sets, separately. Only the gradients of background tokens $\mathcal{T}_{\text{pr}}$ are clipped to bound their contribution to the model update, followed by the addition of Gaussian noise.

The following theorem analyzes the privacy cost of our proposed algorithm. The proof is deferred to Section A.1 in the appendix.

Theorem 5.4. *Given a dataset $\mathcal{D}$ of size N , batch-size $B < N$, number of training epochs T , gradient clipping threshold $C > 0$, and Gaussian noise standard deviation $\sigma > 0$, Algorithm 1 satisfies (ϵ, δ) -MaskDP for any given $\delta \in (0, 1)$, where:*

$$\begin{aligned}
\epsilon \leq \min_{\alpha > 1} TN/B(\alpha-1) \log \left\{ (1 - B/N)^{\alpha-1} (\alpha B/N - B/N + 1) \right. \\
& + \binom{\alpha}{2} (B/N)^2 (1 - B/N)^{\alpha-2} e^{C^2/\sigma^2} \\
& \left. + \sum_{l=3}^{\alpha} \binom{\alpha}{l} (1 - B/N)^{\alpha-l} (B/N)^l e^{(l-1)(C^2/2\sigma^2)} \right\} \\
& + \log(\alpha^{-1}/\alpha) - (\log \delta + \log \alpha)/(\alpha-1),
\end{aligned}$$

providing that the labels are not protected, i.e., assuming the label to be the first token of every record, then $\forall i \in [N], M(i, 1) = 0$.



Figure 38: Anonymization examples from *NTU RGB+D 60* (a), *Toyota Smarthome* (b) and *UCF-101* (c). Our anonymization method can be applied to difficult poses (a) and maintains important context information such as food on a table or a kitchen environment (b).

5.2.3 Anonymization

This section presents our video anonymization approach that effectively preserves individual privacy while maintaining video quality and contextual integrity. Our methodology integrates advanced pose estimation, 3D avatar rendering, and diffusion models to achieve high-fidelity, privacy-preserving video anonymization, as illustrated in Figure 37 and delineated in Algorithm 2. Anonymization examples are shown in Figure 38.

The process initiates with the application of a pose estimation network f_{pos} [89] on a video x to extract temporally tracked 3D pose information, denoted as $\mathbf{p}$. Subsequently, our custom rendering function f_{ren} synthesizes diverse virtual humans using the SMPL-H model [150, 211]. The output of $f_{\text{ren}}(\mathbf{p})$ comprises rendered avatars $\mathbf{h}$, corresponding depth maps, and image masks $\mathbf{r}$ for the rendered subjects. Concurrently, we generate a person mask $\mathbf{q}$ utilizing the segmentation network f_{seg} [138]. To ensure temporal consistency, we apply operation B to $\mathbf{q}$; unflickering the masks across frames while simultaneously growing them slightly and applying a Gaussian kernel for alpha blending on edges.

An inpainting mechanism f_{inp} [143] is applied on x to remove any originally visible person, using the person mask $\mathbf{q}$, resulting in $\hat{x}$. The synthetic avatars are then merged with this inpainted background, yielding $\tilde{x}$. A temporally smooth tubelet video crop operation is then executed on $\tilde{x}$, $\mathbf{r}$, and $\mathbf{q}$, yielding one or multiple tubelets $\tilde{x}_j$, $\mathbf{r}_j$, and $\mathbf{q}_j$.

5.2.3.1 Multi-Conditional Diffusion Refinement

For each tubelet crop $\tilde{x}_j$, which contains both the synthetic avatar and its surrounding background context, we extract three distinct conditioning signals: depth maps C_d using a ZoeDepth detector $E(\tilde{x}_j)$, human pose keypoints C_p via pose detection $P(\tilde{x}_j)$, and facial landmark keypoints C_f through face detection $F(\tilde{x}_j)$. These conditioning signals provide structural and anatomical constraints that ensure the refined output maintains accurate human proportions and movements while enhancing visual fidelity.

Our diffusion function G implements a specialized pipeline integrating a pre-trained Stable Diffusion model with AnimateDiff [98] temporal layers and three parallel controlnet

Algorithmus 2 : Diffusion-refined synthetic humans

Input : Raw Video $\mathbf{x}$; Pose estimation f_{pos} ; Rendering f_{ren} ; Person segmentation f_{seg} ; Video inpainting f_{inp}

Result : Anonymized Video $\tilde{\mathbf{x}}$

```

1  $\mathbf{p} \leftarrow f_{\text{pos}}(\mathbf{x})$                                  $\triangleright$  Estimate 4D poses
2  $\mathbf{h}, \mathbf{r} \leftarrow f_{\text{ren}}(\mathbf{p})$                          $\triangleright$  Generate avatars and render masks
3  $\mathbf{q} \leftarrow f_{\text{seg}}(\mathbf{x})$                                  $\triangleright$  Generate person masks
4  $\mathbf{q} \leftarrow B(\mathbf{q})$                                  $\triangleright$  Unflicker and smooth masks
5  $\hat{\mathbf{x}} \leftarrow f_{\text{inp}}(\mathbf{x}, \mathbf{q})$                      $\triangleright$  Inpaint to remove persons
6  $\tilde{\mathbf{x}} \leftarrow \text{Merge}(\hat{\mathbf{x}}, \mathbf{h}, \mathbf{r})$                  $\triangleright$  Overlay avatars
7  $\{\tilde{\mathbf{x}}_j, \mathbf{r}_j, \mathbf{q}_j\}_{j=1}^M \leftarrow C(\hat{\mathbf{x}}, \mathbf{r}, \mathbf{q})$      $\triangleright$  Smooth video crops
8  $\hat{\mathbf{x}} \leftarrow []$  for  $j \leftarrow 1$  to  $M$  do
  // Depth, pose, face keypoint detection
9    $C_d, C_p, C_f \leftarrow E(\tilde{\mathbf{x}}_j), P(\tilde{\mathbf{x}}_j), F(\tilde{\mathbf{x}}_j)$ 
10   $\hat{\mathbf{x}}_j \leftarrow G(\tilde{\mathbf{x}}_j, C_d, C_p, C_f)$              $\triangleright$  Video diffusion
11   $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}} \cup \{\hat{\mathbf{x}}_j\}$ 
12 end
13  $\tilde{\mathbf{x}} \leftarrow K(\hat{\mathbf{x}}, \{\hat{\mathbf{x}}_j\}_{j=1}^M, \{\mathbf{q}_j\}_{j=1}^M)$      $\triangleright$  Reintegrate
14 return  $\tilde{\mathbf{x}}, \mathbf{r}$ 

```

modules. While standard image diffusion models operate on individual frames independently, AnimateDiff extends this capability to the temporal dimension by introducing motion modeling layers that enforce inter-frame consistency. However, AnimateDiff’s architecture is inherently constrained to processing a limited number of frames (typically 16) simultaneously due to its training configuration and memory constraints.

To overcome this limitation and enable processing of arbitrarily long videos, we adapt and extend a multi-resolution sampling approach for unbounded video generation that emerged from the open-source community. We implement this technique within our multi-conditional diffusion pipeline, applying it to both the input frames and the control signals that guide video generation.

This method generates overlapping frame sets at exponentially increasing stride values. For a video with N_f frames and a model capacity of L frames, the maximum stride level S is determined as:

$$S = \min(S_{\text{max}}, \lceil \log_2(N_f/L) \rceil + 1)$$

For each stride $s \in \{2^0, 2^1, \dots, 2^{S-1}\}$, frame sets F_j are generated starting at positions j with:

$$F_j = \{(j + i \cdot s) \bmod N_f \mid i \in [0, L)\}$$

During each denoising step, the scheduler iterates through multiple context windows, processing subsets of frames and accumulating predictions in a weighted manner. We maintain running totals of noise predictions and frame counts, aggregating contributions from all context windows that include each frame. The final noise prediction for each frame is computed as a weighted average, ensuring smooth transitions between context windows.

In our implementation we extend this sampling strategy to multiple conditioning signals. The same context windows are applied consistently to both the input frames and all three conditioning signals (depth maps c_d , pose keypoints c_p , and facial landmarks c_f),

ensuring proper alignment throughout the diffusion process. For each context window, the controlnet modules process the corresponding conditioning frames and modulate the UNet backbone to guide the generation.

For each action category in our dataset, we generate class-specific text prompts that semantically guide the diffusion model. These prompts describe realistic human appearances while preserving action-specific characteristics, significantly improving the quality and relevance of the generated outputs. The prompts are consistently applied across all frame sets belonging to the same video.

After processing all frame sets at all stride levels, we obtain multiple refined predictions for each original frame. These predictions are aggregated by averaging in latent space, with each prediction contributing equally to the final result. The accumulation process naturally gives more weight to frames that appear in multiple context windows, ensuring a smooth blending between different temporal scales.

The final aggregated latent representations are decoded back to pixel space, yielding refined versions of the tubelet crops that include both the enhanced avatar and potentially modified background regions. To preserve the original background integrity, we use the person masks $\{\mathbf{q}_j\}_{j=1}^M$ to extract only the refined avatar portions $\hat{\mathbf{x}}_j$ from these outputs, which are then carefully blended with the original inpainted background $\hat{\mathbf{x}}$. This approach ensures improved integration of the avatar with its environment while preventing unwanted modifications to the background.

This process is repeated for each video in the dataset, resulting in an anonymized video dataset $\mathcal{D}'' = \{(\tilde{\mathbf{x}}_1, \mathbf{r}_1, \mathbf{y}_1), \dots, (\tilde{\mathbf{x}}_N, \mathbf{r}_N, \mathbf{y}_N)\}$. Since $\mathbf{r}_i$ provides pixel-wise exact information of the presence of synthetic data, spatio-temporal regions which are masked as such can be considered completely anonymous and do not need additional privacy measures, while non-synthesized regions undergo masked differentially private training.

There are limitations to this concept of anonymization which are addressed in Section 5.2.6.

5.2.4 Experiments

In this section, we evaluate MaskDP in comparison to standard differentially private training (DP-SGD) on synthetically anonymized datasets which only require region specific privacy. We follow similar work [152] and focus our experiments on the utility privacy tradeoff given a certain privacy budget ϵ , rather than trying to approximate effective privacy with simulated attacks.

5.2.4.1 Datasets

Our experiments target multiple diverse and common human activity recognition datasets. NTU RGB+D [228] is a balanced dataset containing 56K video samples labeled with 60 action categories and recorded in a very controlled setup. Toyota Smarthome [52] comprises over 16K video clips containing 31 activities of daily living performed by multiple actors within an apartment which results in background and context information being descriptive of the performed activity. UCF-101[242] is a comparatively small activity recognition dataset containing videos of actions performed in complex environments. Utility on NTU RGB+D as well as UCF-101 is measured by classification accuracy, Toy-

ota Smarthome is an unbalanced dataset and evaluation is measured by mean per-class accuracy (mPCA).

We apply our anonymization pipeline to these datasets as described in Section 5.2.3. This process replaces real human subjects with synthetically-generated avatars while maintaining the original background and context. This allows us to create partially synthetic datasets where synthetically generated data is considered non-sensitive. To ablate the diffusion step in our pipeline we provide experiments for both, rendered synthetic avatars as well as the additional refinement, referencing to this data as either *Render* or *Render+Diff* in Tables 25 and 26 in contrast to the unmodified data (*Real*). Note, that in this setup it is the background information which is considered sensitive due to its insight about the living environment of a specific subject. This is in contrast to approaches which ignore sensitive information contained in background environments.

5.2.4.2 Implementation Details

We utilize MViT2-S[142] for action classification with Kinetics-400 [123] pre-trained weights, selectively fine-tuned using DP-SGD, similar to [152]. This approach maintains higher utility than full fine-tuning, as DP-SGD may erode prior knowledge. MViT2-S, a Transformer model, uses LayerNorm instead of BatchNorm and meets the requirements of differential privacy. Experiments are conducted with batch size 128, employing (Masked) DP-SGD with a clipping threshold of 1 and a 0.01 learning rate over 150 epochs for NTU RGB+D and Toyota Smarthome and 300 epochs for UCF-101. The training includes a 20-epoch warmup followed by cosine decay. Non-private training uses standard SGD with identical parameters.

Data augmentation is used in non-private training but was excluded in our privacy-focused training since the noise added during training already acts as a regularizer and especially with low- ϵ settings training might not converge. We set δ for (Masked) DP-SGD at 10^{-6} for NTU RGB+D and 10^{-5} for Toyota Smarthome and UCF-101. We use Opacus to implement (Masked) DP-SGD and report results for multiple ϵ in all DP-related experiments. For each ϵ value, we calibrate the noise multiplier to achieve the desired privacy level.

5.2.5 Results

5.2.5.1 The Importance of Context



Figure 39: Examples for the data used in each row of Table 24.

Table 24: Performances on NTU RGB+D, Toyota Smarthome (TS) and UCF-101 without differentially private training.

Description	Privacy	NTU Acc	TS mPCA	UCF Acc
Real World	None	83.8	52.7	97.6
Isolated Avatar	Full	54.8	8.8	46.8
Inpainted Background	Person	9.9	10.6	93.0
Render on Background	Person	68.3	26.7	94.7
RoB + Diffusion	Person	75.4	39.5	95.4

To quantify our anonymization pipeline’s impact and to assess contextual importance, we analyze performance across NTU RGB+D 60 (NTU), Toyota Smarthome (TS), and UCF-101 (UCF) datasets without the application of differential privacy (Table 24, examples in Figure 39). Real-world baselines establish upper bounds, with TS exhibiting the highest complexity. Isolated avatars preserve privacy but significantly compromise utility, particularly in TS (8.8% accuracy). NTU’s controlled environment with uniform office background mitigates this effect (54.8% accuracy), benefiting moderately from contextual information. Notably, inpainted backgrounds alone maintain high accuracy only in UCF (93.0%), indicating its strong context dependence and suggesting heavy reliance on contextual cues.

Rendering avatars on real backgrounds substantially improves performance: NTU (68.3% vs 54.8%), TS (26.7% vs 8.8%), UCF (94.7% vs 46.8%). Incorporating our diffusion technique further enhances results, with TS showing the most significant gain (39.5%, +12.8%). These improvements underscore the critical role of real-world context, motivating our approach to combine synthetic avatars with real-world backgrounds. Especially on NTU and UCF we reach full person anonymization while achieving a comparatively small synthetic-to-real utility domain gap. For setups which consider background information non-sensitive, our approach can directly be applied, for setups in which context itself is considered private we refer to our MaskDP experiments in Section 5.2.5.2.

5.2.5.2 MaskDP on Anonymized Activity Datasets

In order to evaluate Masked Differential Privacy (MaskDP) introduced in Section 5.2.2, we perform experiments with varying ϵ on combinations of fully real-world and person-anonymized data as well as normal differential privacy (DP) and MaskDP. We compare training on anonymized and non-anonymized datasets and report the non-private training results (indicated by $\epsilon = \infty$) as an upper bound. Table 25 displays our results on NTU RGB+D [228] as well as Toyota Smarthome [52].

Two competing effects are visible, a synthetic-to-real domain gap as analyzed in Section 5.2.5.1 reducing performance by switching from purely real world to anonymized data as well as a loss of performance with progressively tighter privacy budget ϵ . Standard DP provides full differential privacy on a complete video clip, but suffers from prohibitive utility loss for $\epsilon \leq 1$ which is considered strongly private. For input data

Table 25: Accuracy for NTU RGB+D and mean Per Class Accuracy for Toyota Smarthome contrasting the model’s performance with synthetic and anonymized data under Differential Privacy (DP) and Masked Differential Privacy (MaskDP) across various epsilon values.

	Data	Privacy	{Acc,mPCA}@ ϵ						
			0.1	0.25	0.5	0.75	1	5	
NTU RGB+D 60	Real	DP	1.9	2.4	15.9	49.9	55.2	70.8	83.8
	Render	DP	1.7	2.3	14.3	30.7	39.2	52.9	68.3
	Render	MaskDP	4.2	16.2	35.5	48.3	51.0	58.0	68.3
	Rend.+Diff	DP	2.1	2.8	34.5	51.2	56.3	65.0	75.4
	Rend.+Diff	MaskDP	8.1	19.9	48.6	54.0	60.0	64.9	75.4
	<i>Real*</i>	<i>MaskDP*</i>	4.6	26.6	60.8	61.0	72.1	77.3	83.8
TSH	Real	DP	2.6	3.8	8.0	14.0	20.2	29.9	52.7
	Render	DP	3.0	3.6	4.4	4.7	7.3	16.9	26.7
	Render	MaskDP	3.2	9.1	11.1	14.4	16.9	17.1	26.7
	Rend.+Diff	DP	4.0	7.0	14.9	23.8	28.0	32.5	39.5
	Rend.+Diff	MaskDP	12.4	15.4	22.3	24.9	25.6	27.8	39.5
	<i>Real*</i>	<i>MaskDP*</i>	3.9	15.3	18.9	24.3	27.3	32.3	52.7

* *Real human visible, non-private, listed to show domain gap.*

which is anonymized, MaskDP can be used, limiting DP application to sensitive image regions and significantly improving utility on both datasets. To further analyze the effect if the domain gap on MaskDP results, we also applied MaskDP on non-anonymized data, utilizing the person masks from our anonymization pipeline and find that performance significantly increases in all settings, e.g. from 48.6 % to 60.8 % on NTU RGB+D for $\epsilon = 0.5$, indicating that MaskDP profits from increasingly realistic data synthesis methods.

In Table 26 we list corresponding results for UCF-101. For $\epsilon = 5$, this also allows us to compare our implementation to the experiments described in [152]. In our work we leverage MviTv2[142] while [152] makes use of MViT [72], as a result we are able to achieve significantly better results on the same baseline settings. [152] provide additional results for their complementary method of multi-clip training, for comparability we refer to their single clip results. MaskDP on UCF-101 generally displays similar patterns to the experiments on NTU RGB+D and Toyota Smarthome, but limited to improvements within $\epsilon \leq 0.75$.

5.2.5.3 Differential Privacy on Imbalanced Datasets

Differentially private training limits the impact of individual data samples on the final model, which can significantly impair performance on minority classes with few training samples. This is particularly evident in the imbalanced Toyota Smarthome dataset, as illustrated in Figure 40 for $\epsilon = 5$.

Each bar contains a smaller dark gray bar representing the class’s share of samples in the dataset. The upper graph shows DP performance in comparison to standard SGD training,

Table 26: Accuracy on UCF-101 contrasting the model’s performance with synthetic and anonymized data under Differential Privacy (DP) and Masked Differential Privacy (MaskDP) across various epsilon values.

	Data	Privacy	Acc@ ϵ				
			0.25	0.5	0.75	1	5
UCF-101	Real [152]	DP	—				
	Real	DP	1.4	2.7	28.4	77.1	93.1
	Render	DP	1.6	2.4	6.8	57.4	89.0
	Render	MaskDP	2.0	19.0	38.5	58.4	85.3
	Rend.+Diff	DP	1.0	2.0	10.4	57.8	92.0
	Rend.+Diff	MaskDP	8.6	36.2	59.1	72.9	91.0

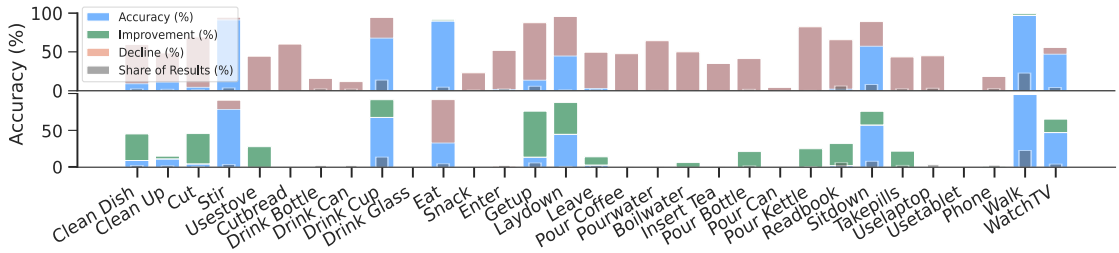


Figure 40: Class-wise accuracy loss on Toyota Smarthome compared to real world non-private training when applying DP (upper) and regain of accuracy upon applying MaskDP (lower) on anonymized data for $\epsilon = 5$. Class distribution marked in dark grey bars within.

while the lower graph shows MaskDP performance in comparison to DP training, thereby first displaying the detrimental effects of DP on minority classes and then presenting how MaskDP alleviates this effect.

DP training sacrifices performance on minority classes, maintaining good results on few majority classes like *Stir*, *Eat* and *Walk*. While MaskDP does not match the performance of SGD training, it demonstrates a notable ability to recover utility across multiple classes. This improvement is particularly visible for minority classes such as *Usetove*, *Pour Bottle*, and *Takepill*s which are not recognized at all with DP training.

5.2.6 Limitations

Our framework employs a hybrid privacy approach combining anonymization with formal differential privacy. This design choice introduces important limitations that users should carefully consider:

PRIVACY MODEL BOUNDARIES. The combination of avatar anonymization and selective differential privacy does not provide the same guarantees as full-video differential privacy. MaskDP provides formal (ϵ, δ) -differential privacy guarantees exclusively for regions marked as private by mask function M . Avatar regions rely on anonymization without formal privacy analysis. While synthetic avatars remove identifying visual features (faces,

skin tone, clothing), they preserve motion patterns and body shapes that may enable re-identification through gait signatures [32, 71, 144], body proportions, or behavioral dynamics. Body shape information leakage is easily mitigated by randomizing SMPL body shape parameters during avatar generation, whereas motion patterns present a more substantial privacy challenge. The assumption that avatars adequately anonymize human identity holds when visual appearance is the primary identifying factor, but may not hold when behavioral biometrics themselves are sensitive or when adversaries possess auxiliary information about movement patterns. Future research can address this by privatizing avatar motion through techniques such as differentially private motion generation or by adding calibrated noise to pose sequence parameters prior to rendering.

UTILITY-PRIVACY TRADEOFFS. Avatar anonymization induces a synthetic-to-real domain gap, limiting MaskDP’s advantage to strongly private regimes (low ϵ). Tables 25 and 26 demonstrate that as privacy requirements relax ($\epsilon \geq 5$), standard DP on real data often outperforms MaskDP on anonymized data due to this domain gap. Further improvements in synthetic human generation methods could shift this crossover point toward higher ϵ values. Additionally, differentially private training suffers on imbalanced datasets (Section 5.2.5.3), though MaskDP alleviates this effect by reducing noise impact.

5.3 CHAPTER CONCLUSION

This chapter addressed two questions about synthetic data for real world problems: (1) *Can we advance the field of human fall detection by generating synthetic data that dominates lab based real-world data on out-of-distribution generalization?* (2) *Can we advance privacy preserving action recognition by combining synthetic avatar anonymization with differential privacy to provide comprehensive privacy protection for both subjects and environmental background context?*

Section 5.1 addressed the first question, demonstrating that synthetic data can outperform real data as a training resource for specialized recognition tasks. OmniFall presents a unified fall detection benchmark which comprises OF-Staged, OF-In-the-Wild, and OF-Synthetic. The critical finding: synthetic video generated with systematic variation in age, ethnicity, body type, and environment achieves better out-of-distribution generalization over real staged datasets (Fall F1: 64.2% vs 61.2% on in-the-wild evaluation). We believe our work to be among the first to demonstrate that generatively produced synthetic video outperforms real staged data for a specialized activity recognition task without further constraints. OF-Syn is not larger than OF-Staged, in fact, considering multi-view recordings it is significantly smaller. Its benefits derive from combined realism and diversity. This contrasts with our findings in Chapter 3 and demonstrates the significant progress in synthetic video data generation, based on the advent of generative foundation models for video synthesis.

In Section 5.2, we address the second question: synthetic data enables privacy-preserving deployment through comprehensive protection mechanisms. MaskDP introduces a hybrid privacy framework and decomposes videos into two parts: synthetic avatar regions, visually anonymized through parametric body model rendering, and sensitive background regions, which receive formal differential privacy protection during training. Since only explicitly defined sensitive spatio-temporal areas are protected rather than entire videos, MaskDP achieves (ϵ, δ) -differential privacy guarantees for background con-

texts. At the same time it clearly improves recognition accuracy over standard differential privacy approaches which protect entire videos indiscriminately.

5.3.1 Contributions to Real World Problems

Contribution 1: The *OmniFall Benchmark* consolidates fall detection research by unifying eight public datasets under a standardized 16-class taxonomy with dense frame-level annotations, comprising 33 hours of densely annotated video in three complementary components: OF-Staged, which adds cross-subject and cross-view protocols; OF-In-the-Wild, which establishes the first standardized benchmark for staged-to-wild evaluation, curating 2.6 hours of genuine accident videos from uncontrolled environments; and OF-Synthetic, contributing 12,000 videos generated via the Wan 2.2 video diffusion model with demographic variation across age, ethnicity, body type, gender, and environments. It provides the first large-scale demonstration that synthetic video outperforms real staged data for in-the-wild fall detection (Fall F1: 64.2% vs 60.5-61.2%). This demonstrates that generatively produced synthetic video can be preferable to real-world data and signals the maturity of modern video diffusion models, pointing towards synthetic generation as a primary data-collection method for specialized recognition tasks.

Contribution 2: *Masked Differential Privacy (MaskDP)* introduces spatially-selective differential privacy by exploiting knowledge about sensitive versus non-sensitive spatio-temporal regions within video frames. It is a novel relaxation of standard Differential Privacy that leverages the perfect pixel-level masks from rendering: the framework decomposes videos into synthetic avatar regions, generated through 3D pose estimation and SMPL rendering, and sensitive background regions, to which (ϵ, δ) -differential privacy is applied exclusively. Hence, MaskDP offers exclusive protection of specific components of individual data points, rather than the entire record, guaranteeing anonymization of the subjects as well as formal privacy for the background contexts. Experiments demonstrate that MaskDP provides better utility-privacy trade-offs in strongly private low- ϵ regimes, significantly improving recognition accuracy compared to standard differential privacy while reducing the utility barriers that previously rendered differential privacy impractical for visual applications. It trades full protection for practical utility in scenarios where avatar anonymization addresses human identity protection while backgrounds require formal privacy mechanisms.

Across Chapters 3, 4, and 5, this thesis explored synthetic data generation through three complementary lenses: addressing cross-domain challenges through benchmarking, generalization, and adaptation (Chapter 3), enabling controllable generation through parametric body models (Chapter 4), and demonstrating practical deployment applications where synthetic data provides superior solutions to real data limitations (Chapter 5). The following chapter synthesizes insights from these contributions, examining how synthetic data generation reshapes human activity analysis and identifying directions for future research.

Part III

INSIGHTS

IMPACT ON THE FIELD

This thesis leverages synthetic data to address fundamental limitations in video-based human activity recognition where real data collection proves insufficient or infeasible. It demonstrates that synthetic data can not only substitute missing real-world recordings, but also enable new benchmarks, modalities, and privacy applications otherwise out of reach. This chapter synthesizes these contributions, tracing how the proposed datasets, methods, and evaluation protocols reshape core research directions, establish new standards for fall detection and biomechanical analysis, and inform future work on privacy-preserving activity recognition.

6.1 RESEARCH DIRECTIONS

SYNTHETIC-TO-REAL CROSS-DOMAIN BENCHMARKING FOR ACTIVITIES OF DAILY LIVING. This work presents systematic cross-domain evaluation infrastructure for synthetic→real transfer in ADL recognition. Sims4Action [207] introduces standardized evaluation protocols pairing life simulation game-based synthetic data with real-world ADL recordings. This infrastructure enables reproducible comparison of domain adaptation methods and systematic assessment of synthetic data quality for real-world deployment.

PARAMETRIC 3D BODY MODELS FOR HUMAN ACTIVITY RECOGNITION. Parametric body models offer scalable and controllable alternatives to web-scraped video datasets, whereas they achieve competitive performance for both, cross-view action recognition and depth-based activity understanding. We demonstrate this with AMARV [219], a large-scale synthetic video dataset with systematic variation in appearance, environment, and view-point. AMARV avoids the privacy concerns which are inherent in large-scale internet video collection. The approach allows for multi-view and multi-modal learning, without expensive multi-camera infrastructure and without real-world data.

DEEP LEARNING FOR POSE-TO-MUSCLE ACTIVATION ESTIMATION. This work opens biomechanical analysis to deep learning approaches through large-scale training data which couples visual motion with physiological signals. MinT [221] provides the first large-scale dataset with 402-muscle activation patterns. It opens research towards visual inference models which could replace expensive laboratory equipment and physics-based simulation. Apart from that, the dataset serves as a pre-training and augmentation resource for human motion understanding [148]. It enables future applications in rehabilitation monitoring, ergonomics assessment and injury risk prediction as well as sports science, which were so far constrained by the cost and invasiveness of electromyography measurement.

UNIFIED VISUAL FALL DETECTION RESEARCH. The landscape of visual fall detection had remained fragmented; OmniFall [220] unifies and standardizes it through systematic benchmarking infrastructure. It consolidates thirteen years of research in visual fall detection: eight public datasets are harmonized under a standardized 16-class taxonomy with dense frame-level annotations. This allows for the first systematic staged-to-wild evaluation on genuine accident videos. Our experiments demonstrate, that synthetic data generated with controlled demographic diversity outperforms real staged data for in-the-wild generalization. Note, that this reveals a shift in approach: synthetic video generation becomes preferable to staged data for out-of-distribution robustness. The work yields open and robust visual fall detection models which are practically deployable for healthcare and in-home use, and it lowers deployment barriers which so far constrained real-world adoption.

FULL-VIDEO PRIVACY PRESERVATION IN ACTIVITY RECOGNITION. This work extended privacy protection from subject-only anonymization to comprehensive full-video privacy by addressing the commonly overlooked background privacy gap. While related work in privacy for computer vision focuses on removing human features through blurring or avatar replacement, background information itself poses significant re-identification risks in home monitoring scenarios through environmental context, room layouts, and personal objects. MaskDP [222] addresses this problem by combining avatar-based visual anonymization with differential privacy guarantees for background regions.

6.2 NEW DATASETS

CROSS-DOMAIN SYNTHETIC-TO-REAL BENCHMARKS FOR ADL. **Sims4Action** [207] provides 10 hours of annotated gameplay from The Sims 4 across 10 ADL categories, paired with Toyota Smarthome recordings for systematic gaming→real evaluation.

LARGE-SCALE DATASETS FROM PARAMETRIC 3D BODY MODELS. **AMARV (Archive of Motion Capture as Rendered Videos)** [219] integrates AMASS motion capture [155] with scalable Unity-based parametric rendering, generating over 800,000 multi-view video clips across 260 action categories from 227 subjects. Synchronized modalities include RGB, depth, instance segmentation, surface normals, and 2D/3D pose across four camera viewpoints with random $\pm 30^\circ$ variation. Visual diversity stems from 1,240 body textures and 201 HDRI backgrounds.

LARGE-SCALE MUSCLE ACTIVATION ESTIMATION DATASETS. **MinT (Muscles in Time)** [221] couples 10 hours of motion sequences from 227 AMASS subjects (187 actions) with 402-muscle strand activation patterns generated via OpenSim musculoskeletal simulation [58, 227], representing an order of magnitude increase in subject count and substantially broader muscle coverage (4–20 muscles previously) compared to experimental electromyography datasets [38, 76, 103].

UNIFIED FALL DETECTION BENCHMARK. Under a standardized 16-class taxonomy with dense frame-level annotations, **OmniFall** [220] unifies eight public datasets [2, 11, 33, 94, 163, 256, 302] and consolidates fall detection research into three distinct components

totaling 33 hours single-view footage (82 hours multi-view): **OF-Staged** provides 14 hours of unique recordings (42 hours multi-view) from 101 subjects across 29 camera views with systematic cross-subject and cross-view evaluation splits; **OF-Synthetic** contributes 17 hours (12,000 videos) generated via Wan 2.2 [271] diffusion model with controlled demographic variation across age, ethnicity, body type, gender, and environments; **OF-In-the-Wild** introduces 2.6 hours of genuine accident videos from OOPS! [68] and serves as the first test-only benchmark for staged-to-wild evaluation of fall detection.

6.3 METHODS, FRAMEWORKS, AND EVALUATIONS

OPTIMAL TRANSPORT-BASED DOMAIN ADAPTATION. OT-DAKD, a domain adaptation framework, combines optimal transport-based prediction calibration with knowledge distillation for synthetic→real video transfer. It directly shapes output prediction distributions to match the respective target domain characteristics through Sinkhorn-Knopp optimization. The approach applies to any pre-trained feature extractor: it offers a plug-in calibration mechanism for cross-domain deployment which improves transfer performance on Sims4Action→Toyota Smarthome evaluation.

MULTIMODAL DOMAIN GENERATION FRAMEWORK. To generate novel training modalities from existing multi-modal representations (RGB, optical flow, pose), MMGen [162] presents a framework built on generative adversarial networks which maximize Sinkhorn distance and preserve semantic content through classification loss. Systematic evaluation of 15 modality combinations reveals that distributional diversity drives cross-domain robustness, whereas pose-based representations (limbs, heatmaps) achieve stronger transfer than appearance-based modalities. Synthetic domain augmentation yields substantial Synthetic→Real improvements on Toyota Smarthome. This shows that learned generative models narrow cross-domain performance gaps without explicit parameter specification or collection of new real-world training data.

ARCHITECTURES FOR POSE-TO-MUSCLE ACTIVATION PREDICTION. We present the first comprehensive evaluation of deep learning architectures for visual motion to 402-dimensional muscle activation time series prediction. Experiments on MinT show that transformer-based sequence models substantially outperform temporal convolutional networks and state space models as well as recurrent architectures for this task. These results give architecture insights for future work on visual-physiological coupling and set baselines for biomechanical research applications. They open paths toward models which could enable real-time biomechanical analysis at scale in the real world.

MASKED DIFFERENTIAL PRIVACY FRAMEWORK. MaskDP [222] introduces spatially-selective differential privacy and makes use of knowledge about sensitive versus non-sensitive spatio-temporal regions within video frames. The framework decomposes videos into synthetic avatar regions, which are visually anonymized through parametric body model rendering, and remaining background regions, which receive formal (ϵ, δ) -differential privacy protection during training. Since the noise is applied exclusively to non-anonymized regions rather than to entire videos, MaskDP achieves improved utility over standard differentially private stochastic gradient descent [1].

OPEN RESEARCH QUESTIONS AND FUTURE WORK

This thesis demonstrates that synthetic data can match or exceed real data as a training resource for human activity recognition under certain conditions. In the following, we explore five future research directions. We extend the central contributions through multi-modal coupling and ask when synthetic superiority generalizes beyond fall detection. We then validate synthetic data against real-world measurements, together with benchmarking infrastructure which compares generation approaches. Finally, we look at privacy through differentially private anonymization as well as at adaptive systems which generate data dynamically.

7.1 EXTENDING CORE CONTRIBUTIONS

The thesis contributions present several bases which open immediate research directions. Here we study natural extensions which couple multiple modalities, as well as the implications of doing so.

7.1.1 Multi-Modal Learning from Shared Motion Representations

Chapter 4 illustrates that AMARV and MinT make use of the same AMASS motion sequences: AMARV generates visual modalities (RGB, depth, pose), whereas MinT couples motion with 402-muscle activation patterns. This shared basis creates multi-modal learning opportunities which were so far constrained by the difficulty of simultaneous visual and physiological capture. *Can models learn cross-modal mappings directly from RGB to muscle activations, without a detour over pose estimation?* Such a capability might enable effort estimation and injury risk assessment as well as rehabilitation monitoring without invasive sensors.

Technical challenges include addressing synthetic-to-real domain gaps for biomechanical signals. Liu *et al.* [148] demonstrated that MinT improves real EMG prediction, yet further research is needed to characterize generalization gaps and develop domain adaptation methods for physiological signals, which face variability across subjects, devices, and protocols. Establishing network architectures for cross-modal relationships and optimizing modality alignment across SMPL rendering and OpenSim simulation remain open questions.

7.2 TOWARDS GENERAL SYNTHETIC DATA SUPERIORITY

Several questions emerge about the generalizability and scalability of synthetic data.

7.2.1 *Generalization Beyond Specialized Tasks*

This thesis describes a progression from strong synthetic-to-real domain gaps (Section 3.1) through task-specific advantages (Section 4.1) to out-of-distribution superiority on fall detection (Section 5.1). OmniFall showcased a clear case in which synthetic data generation should be chosen for its performance advantages. *Whether this synthetic superiority extends beyond fall detection* is unclear. The underlying mechanism, i. e. factorial variation which produces better deployment coverage than limited real data collection, may apply broadly. The *conditions under which synthetic provides better training distributions* need to be identified: advantages emerge when (1) real data exhibits limited variation, (2) systematic data variation better captures the deployment characteristics, and (3) tasks require rare events or diverse demographics. Conversely, synthetic data may not excel when real datasets are abundant and representative.

7.2.2 *Long-Term Coherent Video Generation*

Can generative models produce long-term video whose temporal context spans minutes rather than seconds? Current foundation models do not have this capability, which is why OmniFall-Synthetic consists of short video clips. Autoregressive approaches like Loong [280] achieve minute-level data generation, yet their output lacks biomechanically plausible motion and activity diversity. Complex household activities, industrial processes with fine-grained action sequences, rehabilitation programs, and multi-person interactions all demand multi-minute coherent data generation: output which keeps actors and environments as well as causal structure and physically plausible movement consistent throughout.

7.3 UNDERSTANDING AND VALIDATING SYNTHETIC DATA

Understanding when and why synthetic excels requires rigorous validation against real-world measurements and systematic characterization of generative model limitations.

7.3.1 *Crossing the Synthetic-to-Real Domain Gap for Biomechanical Signals*

In Section 4.2, pose-to-muscle prediction models achieved a Pearson correlation of > 0.8 against OpenSim simulations. However, additional research on the synthetic-to-real gap for muscle activation and EMG data is required. Liu *et al.* [148] demonstrated that MinT data clearly improved real prediction on Muscles in Action [38], which suggests that synthetic biomechanical data aids transfer effectively, whereas further research is still needed to characterize the generalization gaps and to set up training pipelines. This will affect fields like rehabilitation, sports science, and ergonomics, i. e. wherever physiological inference from vision becomes practical.

7.3.2 Generative Model Biases: Opportunity and Challenge

Regarding bias, synthetic data generators offer both, advantages and disadvantages. On the one hand, they enable controlled cross-demographic generation. OmniFall demonstrates how explicit demographic variation parameters allow the creation of datasets with known distributional properties for rigorous bias evaluation. On the other hand, the same generative models inherit and increase the biases in their training data. We do not yet understand *how demographic representations affect the generated diversity, nor which motion patterns or body types end up systematically under- or over-represented*. A systematic characterization would help to identify strategies such as targeted data collection and bias-aware loss functions as well as post-hoc filtering.

7.3.3 Quality Assessment and Failure Modes for Generative Models

Apart from the concern about biases, a broader question needs to be asked: *under what conditions does synthetic data fail to improve, or even actively harm, model performance?* An answer would let us set quality thresholds and identify failure conditions for practical deployment decisions. Generation artifacts can introduce spurious correlations, physical inconsistencies appear and pipeline errors accumulate. Beyond that, generated data distributions can drift away from the deployment distribution. Whether a *universal quality assessment framework can span different generation approaches and target tasks* is itself an open problem. Such a framework must check that the generated data is diverse, looks realistic and, more importantly, that it preserves task-relevant features at the same time.

7.4 DIFFERENTIALLY PRIVATE ANONYMIZATION

Beyond data quality, comprehensive privacy preservation remains essential for deployment in sensitive contexts, particularly addressing limitations in current approaches.

Avatar-based anonymization using SMPL-based fitting [20, 89, 120, 127], as demonstrated in Section 5.2, preserves motion patterns that enable gait-based re-identification despite visual anonymization [32, 71, 144]. Human motion acts as a biometric, i. e. it creates privacy risks even when appearance cues are removed. Recent work approaches this problem [99], yet many works in this field focus on how to train pose estimation models in a privacy preserving manner [239] much rather than on how to anonymize the motion data itself. MaskDP addressed background privacy gaps through selective differential privacy, but did not eliminate the behavioral risks or provide formal guarantees for the avatar pipeline. It remains open *whether differentially private pose and shape estimation, integrated into avatar generation, can achieve full differential privacy for video anonymization*.

MaskDP's current implementation assumes that generated avatars require no further privacy protection, whereas this does not maintain differentially private proofs for the anonymization component. Integrating differential privacy into avatar generation itself, i. e. applying formal guarantees to all user information used in creating the SMPL avatar, could potentially achieve full differential privacy. This approach is promising, as 3D pose and shape models require far fewer parameters than deep learning models, potentially mitigating the severe utility degradation. Since SMPL operates in a low-dimensional parame-

ter space, applying DP at this stage could enable fully differentially private anonymization with better utility preservation.

The technical challenges here are maintaining utility when the estimated poses must be privacy preserving, ensuring privacy composition across pipeline stages, and developing temporal perturbations that obscure gait while preserving action semantics. *Which motion features carry identity versus activity information* is not yet understood, and *whether learned transformations can map individual gaits to generic patterns without losing task-relevant semantics* remains to be shown.

7.5 ACTIVE AND ADAPTIVE SYNTHETIC DATA GENERATION

The synthetic data generation mechanisms established in this thesis enable systems that generate synthetic data dynamically in response to deployment needs. One direction is *a system that dynamically generates and learns from synthetic data during deployment for continual improvement*. Rather than static pretraining datasets, such methods could generate data adaptively and continue learning based on deployment feedback. *Which uncertainty quantification methods effectively identify the distribution gaps that targeted synthetic generation should address* is, so far, an open question.

At the same time, recent work on model collapse shows that repeatedly training on recursively generated synthetic data can induce degenerative feedback loops in which performance and diversity deteriorate over generations [235], whereas related analyses of fairness feedback loops demonstrate that similar mechanisms can amplify bias and reduce minority-group representation when models repeatedly train on their own synthetic or model-labeled data [286]. Both results underscore the need for safeguards against synthetic-data feedback drift. Models which encounter high-uncertainty scenarios could estimate underrepresented demographics, viewpoints, or environments and request targeted generation, whereas the feedback is constrained so that already model-preferred modes are not over-amplified. *How to set up mechanisms which recognize and prevent model collapse and fairness degradation under synthetic data training* remains the harder question.

A further direction is *curriculum learning which makes use of the systematic parameter control in synthetic data generation*. Model competence could guide progressive difficulty increases: simple synthetic data, e. g. frontal views, standard demographics, and isolated motions, would transition to complex scenarios such as extreme viewpoints and rare demographics as well as multi-person interactions.

7.6 CONCLUDING SYNTHESIS

This thesis traced the evolution of synthetic data from a training supplement to a practical alternative in human action recognition. The progression began with severe domain gaps which we quantified (Chapter 3), advanced over controllable generation at scale (Chapter 4), and ultimately demonstrated the out-of-distribution dominance of synthetic data over real-world datasets (Chapter 5). Along this trajectory we showed how to successfully make use of synthetic data for video-based human activity recognition.

Part IV

APPENDIX

APPENDIX

A.1 DEFERRED THEORETICAL PROOFS

This section provides the complete formal proofs for the privacy guarantees of the MaskDP framework introduced in Section 5.2. The main result, Theorem 5.4, establishes that the MaskDP training algorithm (Algorithm 1) satisfies masked Rényi differential privacy (MRDP) with respect to the mask function that designates protected and unprotected tokens. The proof proceeds by first establishing a composition lemma showing that applying differential privacy selectively to protected tokens while leaving unprotected tokens unchanged preserves the desired privacy guarantees.

A.1.1 Proof of Theorem 5.4

To prove Theorem 5.4, we first establish the following lemma, which shows that selective application of differential privacy to masked regions preserves privacy guarantees:

Lemma A.1. *Given a dataset $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$, where each record $x^{(i)} = \{x_1^{(i)}, \dots, x_K^{(i)}\}$ is partitioned into K tokens, and a mask function $M : [N] \times [K] \rightarrow \{0, 1\}$, let*

$$\mathcal{D}_p = \left\{ \{x_k^{(i)} \mid \forall k \in [K] \text{ s.t. } M(i, k) = 1\} \mid i \in [N] \right\}$$

be a dataset consisting of only protected tokens. Correspondingly, let

$$\mathcal{D}_u = \left\{ \{x_k^{(i)} \mid \forall k \in [K] \text{ s.t. } M(i, k) = 0\} \mid i \in [N] \right\}$$

be the set of unprotected tokens. Assume that $f_1(\cdot)$ is an (α, ϵ) -RDP algorithm and $f_2(\cdot)$ is an arbitrary function. Then the mechanism $g(\cdot)$ defined as

$$g(\mathcal{D}) = \{f_1(\mathcal{D}_p), f_2(\mathcal{D}_u)\} \tag{8}$$

satisfies (α, ϵ) -MRDP under the same mask function M .

Proof. Let $\mathcal{D}$ and $\mathcal{D}'$ be two masked-adjacent datasets the mask function M . The goal is to show

$$D_\alpha(g(\mathcal{D}) \| g(\mathcal{D}')) \leq \epsilon.$$

Let $P_{\mathcal{D}}$ and $P_{\mathcal{D}'}$ denote the output distribution of $g(\mathcal{D})$ and $g(\mathcal{D}')$, respectively. Then, for each $x \in \text{Range}(g)$, we have:

$$P_{\mathcal{D}}(x) = P_{\mathcal{D}_p}(x_1)P_{\mathcal{D}_u}(x_2) \tag{9}$$

$$P_{\mathcal{D}'}(x) = P_{\mathcal{D}'_p}(x_1)P_{\mathcal{D}'_u}(x_2), \tag{10}$$

where $P_{\mathcal{D}_p}$, $P_{\mathcal{D}_u}$, $P_{\mathcal{D}_p'}$, and $P_{\mathcal{D}_u'}$ are the output distributions of $f_1(\mathcal{D}_p)$, $f_2(\mathcal{D}_u)$, $f_1(\mathcal{D}_p')$, and $f_2(\mathcal{D}_u')$, respectively, and $x = (x_1, x_2)$, $x_1 \in \text{Range}(f_1)$, and $x_2 \in \text{Range}(f_2)$. Therefore, we have:

$$\begin{aligned} D_\alpha(g(\mathcal{D}) \| g(\mathcal{D}')) &= \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim P_{\mathcal{D}'}} \left[\frac{P_{\mathcal{D}}(x)}{P_{\mathcal{D}'}(x)} \right]^\alpha \\ &= \frac{1}{\alpha - 1} \log \mathbb{E}_{(x_1, x_2) \sim P_{\mathcal{D}'}} \left[\frac{P_{\mathcal{D}_p}(x_1) P_{\mathcal{D}_u}(x_2)}{P_{\mathcal{D}_p'}(x_1) P_{\mathcal{D}_u'}(x_2)} \right]^\alpha \\ &= \frac{1}{\alpha - 1} \log \mathbb{E}_{x_1 \sim P_{\mathcal{D}_p'}} \left[\frac{P_{\mathcal{D}_p}(x_1)}{P_{\mathcal{D}_p'}(x_1)} \right]^\alpha \leq \epsilon. \end{aligned}$$

The second equality follows from (9) and (10). The third equality follows from the fact that $\mathcal{D}_u = \mathcal{D}_u'$, which itself follows from the definition of masked adjacency (Definition 5.2). Finally, the last inequality follows from the fact that $f_1(\cdot)$ is an (α, ϵ) -RDP algorithm. $\square$

Now, we are ready to prove Theorem 5.4:

Proof of Theorem 5.4. Each iteration of Algorithm 1 applies the Gaussian mechanism on the set of non-anonymized tokens, denoted as $\mathcal{D}_p^{(t)} = \{\mathcal{T}_{\text{pr}}^{(t,i)} \mid \forall i \in \mathcal{B}_t\}$, which according to Corollary 3 of [169], is $(\alpha, \alpha^{C^2/2\sigma^2})$ -RDP. It also uses the anonymized tokens for the parameter update. According to Lemma A.1, these two steps together satisfy $(\alpha, \alpha^{C^2/2\sigma^2})$ -MRDP w.r.t. the batch $\mathcal{B}$. As each batch is randomly subsampled, according to Theorem 5 of [307], each iteration of Algorithm 1 satisfies (α, ϵ_t) -MRDP w.r.t. the entire dataset, where ϵ_t is given by:

$$\begin{aligned} \epsilon_t \leq \frac{1}{\alpha - 1} \log \left\{ (1 - B/N)^{\alpha-1} (\alpha B/N - B/N + 1) \right. \\ \left. + \binom{\alpha}{2} (B/N)^2 (1 - B/N)^{\alpha-2} e^{\frac{C^2}{\sigma^2}} \right. \\ \left. + \sum_{l=3}^{\alpha} \binom{\alpha}{l} (1 - B/N)^{\alpha-l} (B/N)^l e^{(l-1)(\frac{C^2}{2\sigma^2})} \right\}. \quad (11) \end{aligned}$$

As the algorithm runs for TN/B iterations, based on the composition theorem of RDP (Proposition 1 of [169]), Algorithm 1 satisfies $(\alpha, TN\epsilon_t/B)$ -MRDP. The proof is completed by replacing ϵ_t with (11) and applying the RDP to DP conversion theorem from [169]. $\square$

A.2 OMNIFALL SYNTHETIC DATA GENERATION METHODOLOGY

Existing fall detection datasets exhibit systematic demographic biases, i. e. they predominantly feature young to middle-aged participants with limited ethnic diversity. To address these limitations and to avoid the ethical complexities of collecting fall data from vulnerable elderly populations, we generated OF-Synthetic with explicit control over demographic attributes through structured prompt design for the Wan 2.2 video diffusion model, ensuring systematic coverage across age groups and ethnicities as well as body types and environmental contexts.

A.2.1 *Demographic Attribute Design*

AGE GROUPS. Six age categories follow common reporting practices, i. e. toddlers (1–4 years), children (5–12 years), teenagers (13–17 years) and young adults (18–34 years), as well as middle-aged (35–64 years) and elderly (65+ years). While elderly individuals constitute our primary target demographic, all age groups allow for both, age-specific evaluation and training diversity. Each age group receives equal representation with 200 generated videos per action class.

RACE AND ETHNICITY. For systematic ethnic diversity, we adopt the seven categories defined by the U.S. Office of Management and Budget (OMB) standard for race and ethnicity reporting [262]: Black, Asian, White, Hispanic/Latino, Middle Eastern/North African (MENA), American Indian/Alaska Native (AIAN), and Native Hawaiian/Pacific Islander (NHPI). This standardized taxonomy allows for consistent demographic reporting and comparison with population statistics.

SKIN TONE. Beyond categorical ethnicity labels, we incorporate the Monk Skin Tone Scale [172], a perceptually uniform 10-point scale, i. e. MST1–MST10 ranging from very light to completely dark skin. This continuous representation captures intra-ethnic skin tone variation and offers finer-grained control over visual appearance diversity than ethnicity categories alone.

BODY MORPHOLOGY. Four BMI bands (underweight, normal, overweight, obese) and three height categories (short, average, tall) give systematic body type coverage. BMI bands translate to appearance descriptors in prompts, e. g. thin, average build, chubby, or stocky. Gender presentation is specified as male or female.

Beyond person-level attributes, environmental context and scenario characteristics contribute additional sources of controlled variation.

A.2.2 *Environmental and Contextual Diversity*

AGE-APPROPRIATE ENVIRONMENTS. Each age group is associated with both, 100 indoor and 100 outdoor environment descriptors, reflecting activities for that demographic.

Younger age groups feature educational and play-oriented settings: toddler environments include daycare play areas, indoor playgrounds, and sandbox areas; children’s environments span elementary classrooms, playgrounds, and sports fields; teenage environments encompass high school hallways, athletic facilities, and teen activity centers. Adult age groups reflect professional and residential contexts: young adult environments include offices, gyms, universities, and apartments; middle-aged environments feature corporate settings, professional spaces, and home environments; elderly environments comprise senior centers and assisted living facilities as well as medical clinics and residential areas. This age-specific environment mapping keeps the generated scenarios in realistic activity contexts for each demographic.

CLOTHING DIVERSITY. Each age group has 200+ age-appropriate clothing options, i. e. everyday casual wear (60%), seasonal attire (25%) and sleepwear (10%) as well as special

occasion clothing (5%). Examples range from toddlers’ colorful casual wear to elderly soft comfortable clothing, ensuring contextually appropriate attire with visual diversity.

A.2.3 *Action-Specific Attributes*

Apart from demographic and environmental attributes, which remain constant across action classes, each action class incorporates action-specific descriptors to control both, temporal dynamics and scenario characteristics.

FALL-SPECIFIC ATTRIBUTES. Fall videos include 30 fall mechanism types (e. g. slip on wet surface, trip over obstacle, lose balance from dizziness), 20 fall direction specifications (falling forward, backward, sideways, diagonally, while spinning) and 20 recovery attempt descriptions (e. g. reaching for support, windmilling arms, bracing for impact), as well as age-appropriate prior activities and 20 facial expression descriptors.

Note, that the inclusion of facial expressions addresses a specific challenge with the Wan 2.2 model: the model exhibits a tendency to generate individuals with smiling or happy expressions. Presumably, this is due to training data distribution or alignment objectives. This creates inappropriate and unrealistic scenarios in which individuals smile while falling. To counteract this bias, we explicitly specify contextually appropriate facial expressions (e. g. startled, panicked, focused, surprised) in fall scenario prompts.

OTHER ACTION CLASSES. Similar attribute sets were defined for other action classes (lie down, stand up, sit down, walk), incorporating action-specific motion descriptors and environmental constraints as well as contextual details appropriate to each activity. The “other” category receives special attention for minimal overlap with the 15 primary action classes through pre-screening of candidate activities for distinctiveness.

A.2.4 *Camera Specification*

To complete the prompt specification, camera parameters control viewpoint variation. Each video specifies camera vantage point through four attributes: elevation (e. g. eye-level, low, high/CCTV perspective, top-down/birds eye) and azimuth (front, rear, left, right), as well as distance (medium/distant shot, far/long-distance shot) and shot type (static wide, static medium wide). Frame rates are specified as either 24fps, 25fps or 30fps real-time. This systematic camera parameter variation produces videos from diverse viewpoints representative of both, real-world surveillance and monitoring camera placements.

A.2.5 *Prompt Construction and Generation*

PROMPT STRUCTURE. Each prompt combines demographic and environmental attributes with action-specific and camera attributes into a structured natural language description. A typical fall prompt follows this template:

“A [height] [body type] [age descriptor] [gender] with [skin tone] and [ethnicity] features wearing [clothing] is in the process of falling due to [fall mechanism], [fall direction], in [environment], while [recovery attempt], after [prior activity], with

[facial expression], [framerate], [shot type], camera positioned [elevation] from the [azimuth], it is a [distance]. Realistic. Fully dressed upper body."

Note, that the modifiers "Realistic" and "Fully dressed upper body" were found beneficial for generating appropriate video content.

DETERMINISTIC ATTRIBUTE ASSIGNMENT. For reproducibility and systematic coverage of the attribute space, each video receives a unique identifier (e. g. `fall_el_001` for elderly fall video 1). This identifier deterministically seeds attribute selection via hash-based random number generation. This ensures, that identical attributes are selected for the same identifier across generation runs, whereas the attribute combinations stay diverse across the full dataset.

DATASET SCALE. We generated 200 videos per age group for 10 base action classes (fall, fallen, lie down, lying, sit down, sitting, stand up, standing, walk, other), yielding 200×6 age groups $\times$ 10 actions = 12,000 videos totaling 17 hours of synthetic footage at 1280 $\times$ 720 resolution and 16fps. While the initial generation targeted these 10 action classes, subsequent dense annotation revealed the utility of additional fine-grained distinctions. This led to the full 16-class taxonomy used for timeline segmentation evaluation.

A.2.6 Safety Filtering and Iterative Refinement

The Wan 2.2 model occasionally produced inappropriate content requiring iterative filtering: generated videos were manually reviewed, inappropriate content removed, and prompts refined to reduce recurrence. This process repeated until the dataset contained only safe-for-work content suitable for public release.

A.3 OMNIFALL DENSE ANNOTATION PROCESS

The unification of eight staged datasets, the curation of genuine accident videos and the annotation of 12,000 synthetic videos into a consistent 16-class taxonomy required substantial annotation effort. In the following, we detail the annotation workflow and software tooling as well as the quality assurance measures we used for consistent, frame-level labels across the 33 hours of single-view footage (82 hours with multi-view recordings).

A.3.1 Annotation Software Selection and Customization

After systematic comparison, we selected VGG Image Annotator (VIA) [66] as the most efficient platform among available open-source video annotation tools. Custom modifications introduced keyboard-only annotation capabilities, i. e. all operations run without mouse interaction. Thereby, annotation throughput accelerated through rapid temporal navigation and label assignment as well as boundary adjustment via keyboard shortcuts, particularly beneficial for precise frame-level timeline segmentation.

A.3.2 Annotation Team and Training

Approximately 10 annotators received detailed instruction on the 16-class taxonomy, with emphasis on both, visually similar action pairs (e. g. *fall* vs. *lie down*) and temporal boundaries between transient actions and static states. Regular team meetings built consensus on challenging cases, such as transition frame determination (*fall* to *fallen*) and controlled vs. uncontrolled descents, as well as the handling of occluded or out-of-frame actions.

A.3.3 Multi-View Annotation Strategy

For datasets with synchronized multi-view recordings (CMDFall, UP Fall, EDF, OCCU, MCFD), we made use of an efficient annotation transfer approach rather than independently annotating all camera views. For this, we systematically selected for each recording the primary camera view most likely to maintain the subject in frame throughout the activity sequence. Hence, the selection criterion minimized instances in which the subject would be visible in other views, yet occluded or out-of-frame in the annotated view.

Annotations were performed on this primary view and subsequently transferred to all synchronized camera views. During transfer, we accounted for temporal offsets between synchronized cameras, so that labels align at frame level across views. This approach substantially reduced annotation effort (14 hours of unique recordings vs. 42 hours of multi-view footage), while providing complete timeline annotations for all camera perspectives and thereby enabling cross-view evaluation protocols.

A.3.4 Quality Assurance Protocol

Nearly all videos underwent two-pass review, i. e. initial frame-level annotation followed by quality control verifying both, temporal boundary accuracy and label correctness, as well as taxonomy consistency. Discrepancies were corrected, with ambiguous cases referred to team discussion. This workflow rests on specialized tooling and detailed training as well as consensus meetings and efficient multi-view transfer, yielding dense and consistent annotations for OmniFall’s evaluation protocol.

A.4 MINT DATASET SIMULATION QUALITY ANALYSIS

The Muscles in Time (MinT) dataset was generated by applying trajectory optimization and static optimization procedures to motion capture sequences from AMASS. Processed sequences are analyzed using metrics derived from the OpenSim evaluation guidelines [179, 180]. This appendix presents the quantitative quality assessment results on the Lai Uhlrich lower body model, documenting tracking accuracy and knee loading metrics as well as adherence to biomechanically-informed thresholds.

A.4.1 Quality Metrics and Thresholds

Five quality criteria were evaluated for each sequence, following the OpenSim simulation evaluation checklist [179]: (1) Rotational tracking accuracy measures RMSE between simu-

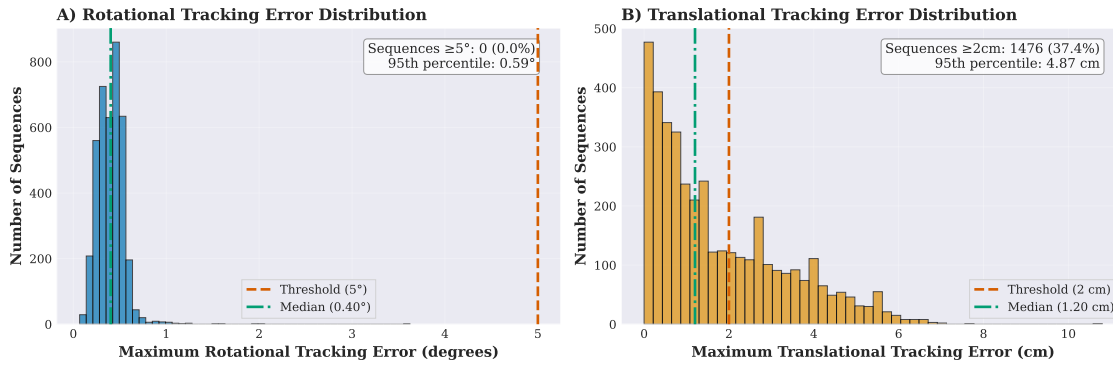


Figure 41: Distribution of maximum tracking errors across 3,947 simulated sequences. **(A)** Rotational tracking errors show median 0.40° with 95th percentile at 0.59° . All sequences remain below the 5° threshold. **(B)** Translational tracking errors show median 1.20 cm with 95th percentile 4.87 cm. 37.4% of sequences ($1,476$ sequences) exceeded the 2 cm threshold.

lated and target joint angles (threshold: 5° , based on typical motion capture accuracy [179]); (2) Translational tracking accuracy measures RMSE for marker positions (threshold: 2 cm, reflecting typical marker tracking accuracy [179]); (3) Peak ground reaction force (threshold: 2.0 BW, where normal walking produces 1.0 – 1.2 BW [124] and running/jumping reach 1.5 – 2.0 BW [166]); (4) Peak muscle activation across all strands (threshold: 0.9 , identifying potentially unrealistic recruitment patterns [179]); (5) Peak knee adduction moment (threshold: 0.08 BW·ht, where typical walking produces 0.02 – 0.04 BW·ht [133]). Ground reaction force data could not be evaluated due to a metadata storage issue.

A.4.2 Tracking Error Results

Figure 41 shows the distribution of maximum tracking errors across all 3,947 sequences.

ROTATIONAL TRACKING. All 3,947 sequences achieved rotational tracking errors below the 5° threshold (0% failure rate). The median maximum rotational error was 0.40° , with mean 0.40° and standard deviation 0.15° as well as a 95th percentile of 0.59° , indicating a tight distribution. The maximum observed rotational error across all sequences was 3.63° .

TRANSLATIONAL TRACKING. Of the 3,947 sequences, $1,476$ sequences (37.4%) showed maximum translational errors over the 2 cm threshold. The median maximum translational error was 1.20 cm, with mean 1.79 cm and standard deviation 1.55 cm as well as a 95th percentile of 4.87 cm, while the maximum observed translational error was 10.80 cm.

A.4.3 Coordinate-Specific Errors

Figure 42 identifies the 20 coordinates with highest median tracking error.

In the OpenSim model structure, the pelvis serves as the root segment with six degrees of freedom (three translations, three rotations), whereas all other joints are modeled as rotational only. The metatarsophalangeal (MTP) joint angles exhibited the highest rotational

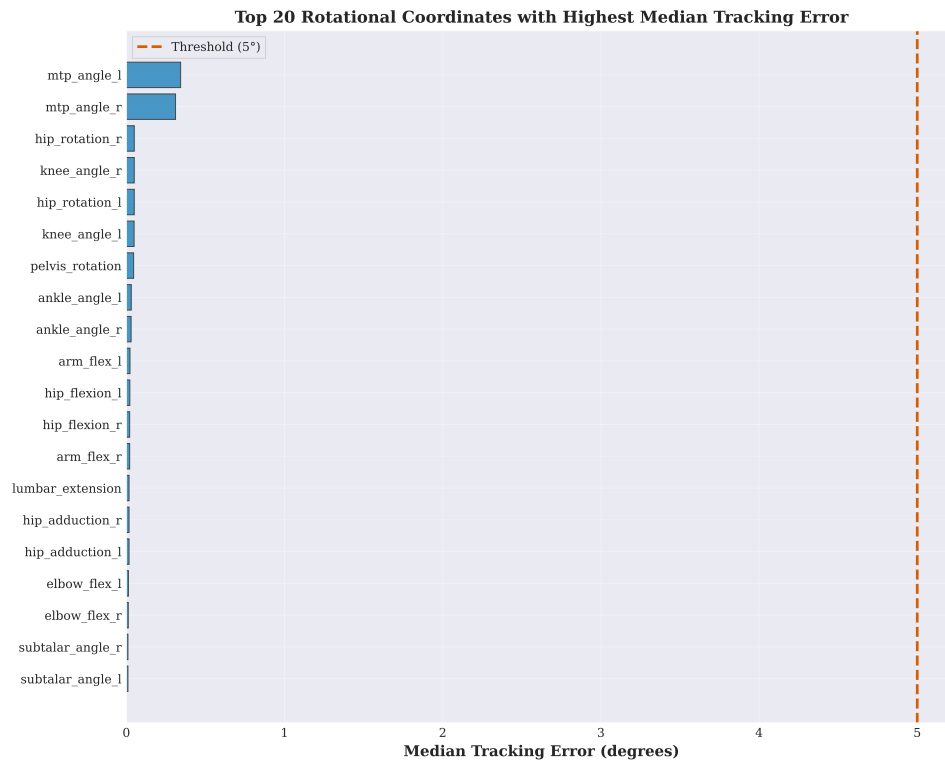


Figure 42: The 20 rotational coordinates (joint angles) with highest median tracking error. The metatarsophalangeal (MTP) joint angles exhibit the highest errors. All values remain well below the 5° threshold (dashed line).

errors, whereas pelvis_tx (anteroposterior translation) showed the highest translational error among the three pelvis translation coordinates.

A.4.4 Knee Loading Metrics

Figure 43 presents the distribution of peak knee adduction moments (KAM) and peak medial contact forces (MCF).

KNEE ADDUCTION MOMENT. Peak KAM values exhibited median 0.023 BW·ht, mean 0.025 BW·ht and standard deviation 0.010 BW·ht as well as a 95th percentile of 0.043 BW·ht. Eight sequences (0.2%) exceeded the 0.08 BW·ht threshold. The maximum observed KAM was 0.133 BW·ht. One sequence with KAM exceeding 1.0 BW·ht was identified as an outlier and removed from analysis.

MEDIAL CONTACT FORCE. Peak MCF values exhibited median 2.69 BW, mean 2.89 BW and standard deviation 1.77 BW as well as a 95th percentile of 6.22 BW. A total of 812 sequences (20.6%) exceeded the 4.0 BW threshold. The maximum observed MCF was 13.90 BW. One sequence with MCF exceeding 20 BW was identified as an outlier and removed from analysis.

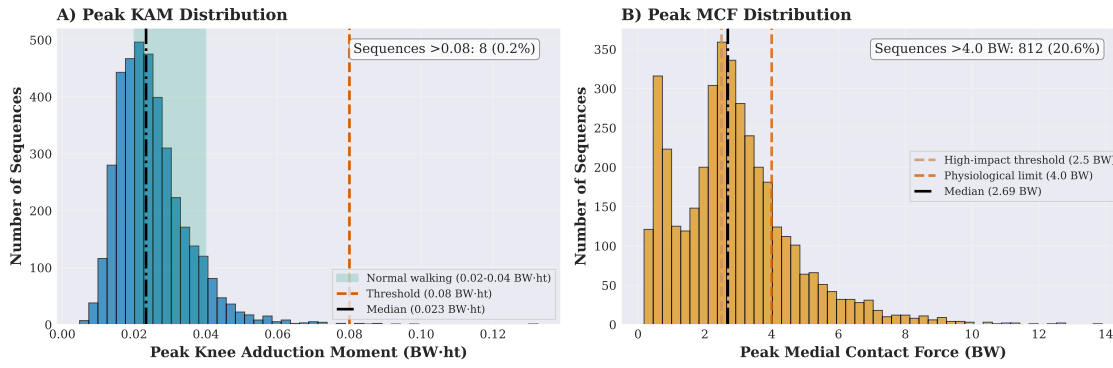


Figure 43: Knee loading metric distributions across all sequences. **(A)** Peak knee adduction moment (KAM) distribution shows median 0.023 BW·ht with 95th percentile 0.043 BW·ht. The shaded region (0.02–0.04 BW·ht) indicates the range reported for normal walking [133]. 0.2% of sequences exceed the 0.08 BW·ht threshold. **(B)** Peak medial contact force (MCF) distribution shows median 2.69 BW with 95th percentile 6.22 BW. 20.6% of sequences exceed the 4.0 BW threshold.

Table 27: Summary statistics for MinT dataset quality metrics across 3,947 sequences.

Metric	Median	95th Percentile	Threshold	% Exceeding
Max Rotational Error (°)	0.40	0.59	5.0	0.0%
Max Translational Error (cm)	1.20	4.87	2.0	37.4%
Peak KAM (BW·ht)	0.023	0.043	0.08	0.2%
Peak MCF (BW)	2.69	6.22	4.0	20.6%

A.4.5 Summary Dashboard

Figure 44 shows an integrated visualization of the primary quality metrics.

The dashboard presents the distribution characteristics for both, rotational tracking error and knee adduction moment, as well as for medial contact force. The rotational tracking error distribution is concentrated below 1° and the KAM distribution below 0.05 BW·ht, while the MCF distribution shows broader spread extending to higher values.

A.4.6 Summary Statistics

Table 27 summarizes the quality metrics across all analyzed sequences.

These results document the quality characteristics of the MinT dataset simulations according to accepted biomechanical simulation evaluation criteria. Both, the 0% failure rate for rotational tracking and the 0.2% failure rate for knee adduction moment, indicate adherence to these thresholds across nearly all sequences. The 37.4% exceedance rate for translational tracking and 20.6% for medial contact force, however, identify specific quality characteristics which may inform filtering strategies for particular research applications.

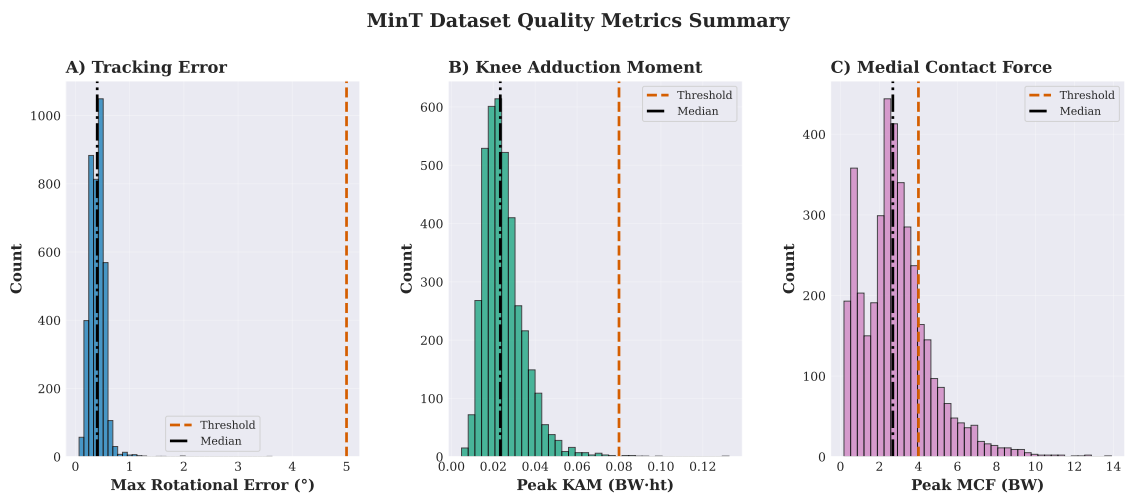
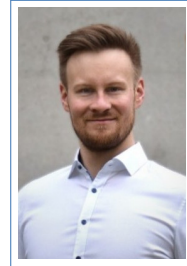


Figure 44: Summary of key quality metrics across 3,947 sequences. **(A)** Maximum rotational tracking error distribution (median 0.40° , threshold 5° shown as dashed red line). **(B)** Peak knee adduction moment distribution (median $0.023 \text{ BW}\cdot\text{ht}$, threshold $0.08 \text{ BW}\cdot\text{ht}$). **(C)** Peak medial contact force distribution (median 2.69 BW , threshold 4.0 BW). Black dashed lines indicate medians.

SHORT CV



David Schneider

Research Interests

Visual Human Activity Recognition, Synthetic Data Generation

Education and Career

- Apr 2021 – Jan 2026 **PhD Student**, *Computer Vision for Human-Computer Interaction Lab (cv:hci)*,
Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology
Research Topic:
“**Leveraging Synthetic Data for Video-based Human Activity Recognition**”
- Oct 2023 – Feb 2024 **PhD Internship**, *Sony Europe B.V.*, Hoofddorp, Schlieren/Switzerland
Focus on Privacy Preserving Machine Learning
- 2017 – 2021 **M.Sc. Computer Science**, *Karlsruhe Institute of Technology*
Focus: Algorithm Engineering and Cognitive Systems
- 2019 **Study Abroad**, *Instituto Tecnológico de Buenos Aires*, Argentina
One semester exchange program
Scholarship: **Baden-Württemberg-STIPENDIUM**
- 2012 – 2017 **B.Sc. Computer Science**, *Karlsruhe Institute of Technology*
- 2015 – 2016 **Erasmus Exchange**, *Universidad Politécnica de Madrid*, Spain
Two semester exchange program
- 2009 – 2012 **Abitur**, *Kaufmännische Schulen Hausach*
Honored for outstanding academic performance, excellence in English and Informatics

Teaching and Supervision

Supervision Master's and Bachelor's theses, Seminars and Practical Courses

Academic Service

- Reviewing Regularly serving as reviewer for AAAI, NeurIPS, CVPR, ICCV, IROS, ICRA, ACMMM, WACV, BMVC
- Award CVPR 2025 Outstanding Reviewer Award

Conference Participation

IROS 2021 (virtual), CVPR 2022, ICRA 2024, ECCV 2024, NeurIPS 2024

AUTHORED PUBLICATIONS

The following thesis-related publications and submissions currently under review resulted from this doctoral research (* Marks equal contribution.):

- [1] David Schneider, Zdravko Marinov, Moritz Mistol, Zeyun Zhong, Alexander Jaus, Rodi Düger, Rafael Baur, M. Saquib Sarfraz, and Rainer Stiefelbogen. **"OmniFall: From Staged Through Synthetic to Wild, A Unified Multi-Domain Dataset for Robust Fall Detection."** *arXiv preprint arXiv:2505.19889*, 2026. Accepted at: *European Conference on Computer Vision*, 2026. <https://arxiv.org/abs/2505.19889>.
- [2] David Schneider, Simon Reiß, Marco Kugler, Alexander Jaus, Kunyu Peng, Susanne Sutschet, M. Saquib Sarfraz, Sven Matthiesen, and Rainer Stiefelbogen. **"Muscles in Time: Learning to Understand Human Motion In-Depth by Simulating Muscle Activations."** In: *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 67251–67281. https://papers.nips.cc/paper_files/paper/2024/hash/7c3465ba08732cc2db38f070bfae601a-Abstract-Datasets_and_Benchmarks_Track.html.
- [3] David Schneider, Marco Keller, Zeyun Zhong, Kunyu Peng, Alina Roitberg, Jürgen Beyerer, and Rainer Stiefelbogen. **"SynthAct: Towards Generalizable Human Action Recognition based on Synthetic Data."** In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13038–13045. doi: [10.1109/icra57147.2024.10611486](https://doi.org/10.1109/icra57147.2024.10611486).
- [4] Alina Roitberg*, David Schneider*, Aulia Djamal, Constantin Seibold, Simon Reiß, and Rainer Stiefelbogen. **"Let's Play for Action: Recognizing Activities of Daily Living by Learning from Life Simulation Video Games."** In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021. doi: [10.1109/IROS51168.2021.9636381](https://doi.org/10.1109/IROS51168.2021.9636381).
- [5] Zdravko Marinov*, David Schneider*, Alina Roitberg*, and Rainer Stiefelbogen. **"Multimodal Generation of Novel Action Appearances for Synthetic-to-Real Recognition of Activities of Daily Living."** In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11320–11327. doi: [10.1109/iros47612.2022.9981946](https://doi.org/10.1109/iros47612.2022.9981946).
- [6] David Schneider, Sina Sajadmanesh, Vikash Sehwal, Saquib Sarfraz, Rainer Stiefelbogen, Lingjuan Lyu, and Vivek Sharma. **"Activity Recognition on Avatar-Anonymized Datasets with Masked Differential Privacy."** Presented at the *2nd International Workshop on Privacy-Preserving Computer Vision*, 2024. <https://arxiv.org/abs/2410.17098v2>.
- [7] David Schneider, Saquib Sarfraz, Alina Roitberg, and Rainer Stiefelbogen. **"Pose-Based Contrastive Learning for Domain Agnostic Activity Representations."** In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2022, pp. 3433–3443. doi: [10.1109/cvprw56347.2022.00387](https://doi.org/10.1109/cvprw56347.2022.00387).

The following scientific publications were co-authored by David Schneider during the doctoral research period but are not directly incorporated into the thesis content:

- [1] Yifei Chen, Kunyu Peng, Alina Roitberg, David Schneider, Jiaming Zhang, Junwei Zheng, Ruiping Liu, Yufan Chen, Kailun Yang, and Rainer Stiefelbogen. "Unveiling the Hidden Realm: Self-supervised Skeleton-based Action Recognition in Occluded Environments." In: *arXiv preprint arXiv:2309.12029* (2023).
- [2] Christian R. G. Dreher, Manuel Zaremski, Fabian Leven, David Schneider, Alina Roitberg, Rainer Stiefelbogen, Michael Heizmann, Barbara Deml, and Tamim Asfour. "Erfassung und Interpretation menschlicher Handlungen für die Programmierung von Robotern in der Produktion." In: *at - Automatisierungstechnik* 70.6 (2022), pp. 517–533. doi: [10.1515/auto-2022-0006](https://doi.org/10.1515/auto-2022-0006).
- [3] Zdravko Marinov, Alina Roitberg, David Schneider, and Rainer Stiefelbogen. "Modselect: Automatic modality selection for synthetic-to-real domain generalization." In: *European Conference on Computer Vision*. Springer. 2022, pp. 326–346. doi: [10.1007/978-3-031-25085-9_19](https://doi.org/10.1007/978-3-031-25085-9_19).
- [4] Kunyu Peng, Alina Roitberg, David Schneider, Marios Koulakis, Kailun Yang, and Rainer Stiefelbogen. "Affect-DML: Context-Aware One-Shot Recognition of Human Affect using Deep Metric Learning." In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE. 2021, pp. 1–8. doi: [10.1109/fg52635.2021.9666940](https://doi.org/10.1109/fg52635.2021.9666940).
- [5] Kunyu Peng, David Schneider, Alina Roitberg, Kailun Yang, Jiaming Zhang, Chen Deng, Kaiyu Zhang, M Saquib Sarfraz, and Rainer Stiefelbogen. "Towards Video-based Activated Muscle Group Estimation in the Wild." In: *ACM Multimedia 2024*. 2024. doi: [10.1145/3664647.3680630](https://doi.org/10.1145/3664647.3680630).
- [6] Kunyu Peng, Di Wen, M Saquib Sarfraz, Yufan Chen, Junwei Zheng, David Schneider, Kailun Yang, Jiamin Wu, Alina Roitberg, and Rainer Stiefelbogen. "Mitigating label noise using prompt-based hyperbolic meta-learning in open-set domain generalization." In: *International Journal of Computer Vision* 134.3 (2026), p. 99.
- [7] Kunyu Peng, Di Wen, David Schneider, Jiaming Zhang, Kailun Yang, M Saquib Sarfraz, Rainer Stiefelbogen, and Alina Roitberg. "Exploring Few-Shot Adaptation for Activity Recognition on Diverse Domains." In: *arXiv preprint arXiv:2305.08420* (2023).
- [8] Kunyu Peng, Cheng Yin, Junwei Zheng, Ruiping Liu, David Schneider, Jiaming Zhang, Kailun Yang, M Saquib Sarfraz, Rainer Stiefelbogen, and Alina Roitberg. "Navigating open set scenarios for skeleton-based action recognition." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 5. 2024, pp. 4487–4496. doi: [10.1609/aaai.v38i5.28247](https://doi.org/10.1609/aaai.v38i5.28247).
- [9] Alina Roitberg, Kunyu Peng, Zdravko Marinov, Constantin Seibold, David Schneider, and Rainer Stiefelbogen. "A Comparative Analysis of Decision-Level Fusion for Multimodal Driver Behaviour Understanding." In: *Intelligent Vehicles Symposium 2022, IEEE*. 2022. doi: [10.48550/ARXIV.2204.04734](https://doi.org/10.48550/ARXIV.2204.04734). URL: <https://arxiv.org/abs/2204.04734>.
- [10] Alina Roitberg, Kunyu Peng, David Schneider, Kailun Yang, Marios Koulakis, Manuel Martinez, and Rainer Stiefelbogen. "Is My Driver Observation Model Overconfident? Input-Guided Calibration Networks for Reliable and Interpretable Confidence Estimates." In: *IEEE Transactions on Intelligent Transportation Systems* 23.12 (2022), pp. 25271–25286. doi: [10.1109/TITS.2022.3196410](https://doi.org/10.1109/TITS.2022.3196410).
- [11] Manuel Zaremski et al. "Learning human actions from complex manipulation tasks and their transfer to robots in the circular factory." In: *at - Automatisierungstechnik* 72.9 (2024), pp. 844–859. doi: [10.1515/auto-2024-0008](https://doi.org/10.1515/auto-2024-0008).
- [12] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelbogen, and Jürgen Beyerer. "Anticipative feature fusion transformer for multi-modal action anticipation." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 6068–6077. doi: [10.1109/wacv56688.2023.00601](https://doi.org/10.1109/wacv56688.2023.00601).

DISCLOSURE OF GENERATIVE AI AND AI-ASSISTED TOOLS IN THE WRITING PROCESS

Following the "Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG"¹ from September 2023, the author discloses the usage of Grammarly, Claude (Sonet 4.5 and Opus 4.1) and ChatGPT (version 5) to enhance the language quality of this work and to support typesetting tasks (e.g., formatting tables).

¹ <https://web.archive.org/web/20260624154117/https://www.dfg.de/resource/blob/289674/ff57cf46c5ca109cb18533b21fba49bd/230921-stellungnahme-praesidium-ki-ai-data.pdf>

BIBLIOGRAPHY

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. “Deep learning with differential privacy.” In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318. doi: [10.1145/2976749.2978318](https://doi.org/10.1145/2976749.2978318).
- [2] Ekram Alam, Abu Sufian, Paramartha Dutta, Marco Leo, and Ibrahim A. Hameed. “GMDCSA-24: A dataset for human fall detection in videos.” In: *Data in Brief* 57 (Dec. 1, 2024), p. 110892. issn: 2352-3409. doi: [10.1016/j.dib.2024.110892](https://doi.org/10.1016/j.dib.2024.110892). url: <https://www.sciencedirect.com/science/article/pii/S2352340924008552> (visited on 01/27/2025).
- [3] Mohammad Arif Ul Alam. “AI-Fairness Towards Activity Recognition of Older Adults.” In: *MobiQuitous 2020: 17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 2020, pp. 108–117. doi: [10.1145/3448891.3448943](https://doi.org/10.1145/3448891.3448943).
- [4] Mohammad Mehdi Alemi, Jacob J. Banks, Andrew C. Lynch, Brett T. Allaire, Mary L. Bousein, and Dennis E. Anderson. “EMG Validation of a Subject-Specific Thoracolumbar Spine Musculoskeletal Model During Dynamic Activities in Older Adults.” In: *Annals of Biomedical Engineering* 51.10 (2023), pp. 2313–2322. doi: [10.1007/s10439-023-03273-3](https://doi.org/10.1007/s10439-023-03273-3).
- [5] Mona Saleh Alzahrani, Salma Kammoun Jarraya, Manar Ali Salamah, and Hanène Ben-Abdallah. “FallFree: Multiple fall scenario dataset of cane users for monitoring applications using kinect.” In: *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2017, pp. 327–333. doi: [10.1109/sitis.2017.61](https://doi.org/10.1109/sitis.2017.61).
- [6] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. “Differentially private learning with adaptive clipping.” In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17455–17466. url: https://openreview.net/forum?id=RUQ1zwZR8_.
- [7] Sandy Ardianto and H. Hang. “NCTU-GTAV360: A 360° Action Recognition Video Dataset.” In: *IEEE International Workshop on Multimedia Signal Processing (IEEE MMSP)*. 2019, pp. 1–5. doi: [10.1109/mmisp.2019.8901740](https://doi.org/10.1109/mmisp.2019.8901740).
- [8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. “Wasserstein Generative Adversarial Networks.” In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 214–223. url: <https://proceedings.mlr.press/v70/arjovsky17a.html>.

- [9] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. “Vivit: A video vision transformer.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6836–6846. doi: [10.1109/iccv48922.2021.00676](https://doi.org/10.1109/iccv48922.2021.00676).
- [10] Reyhane Askari-Hemmat, Mohammad Pezeshki, Elvis Dohmatob, Florian Bordes, Pietro Astolfi, Melissa Hall, Jakob Verbeek, Michal Drozdal, and Adriana Romero-Soriano. “Improving the Scaling Laws of Synthetic Data with Deliberate Practice.” In: *International Conference on Machine Learning (ICML)*. PMLR, 2025, pp. 1833–1863. URL: <https://openreview.net/forum?id=0LZRtvK871>.
- [11] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. *Multiple cameras fall data set*. Tech. rep. 1350. DIRO, Université de Montréal, 2010. URL: <https://www.iro.umontreal.ca/~labimage/Dataset/>.
- [12] Greet Baldewijns, Glen Debar, Gert Mertes, Bart Vanrumste, and Tom Croonenborghs. “Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms.” In: *Healthcare technology letters* 3.1 (2016), pp. 6–11. doi: [10.1049/htl.2015.0047](https://doi.org/10.1049/htl.2015.0047).
- [13] Jacob J. Banks, Brian R. Umberger, and Graham E. Caldwell. “EMG optimization in OpenSim: A model for estimating lower back kinetics in gait.” In: *Medical Engineering & Physics* 103 (2022), p. 103790. doi: [10.1016/j.medengphy.2022.103790](https://doi.org/10.1016/j.medengphy.2022.103790).
- [14] Omer Bar-Tal et al. “Lumiere: A Space-Time Diffusion Model for Video Generation.” In: *ACM SIGGRAPH Asia 2024 Conference Papers*. 2024, pp. 1–11. doi: [10.1145/3680528.3687614](https://doi.org/10.1145/3680528.3687614).
- [15] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. “Is Space-Time Attention All You Need for Video Understanding?” In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 813–824. URL: <https://proceedings.mlr.press/v139/bertasius21a.html>.
- [16] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. “Multi-garment net: Learning to dress 3d people from images.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 5420–5430. doi: [10.1109/iccv.2019.00552](https://doi.org/10.1109/iccv.2019.00552).
- [17] Marian Bittner, Wei-Tse Yang, Xucong Zhang, Ajay Seth, Jan van Gemert, and Frans C. T. van der Helm. “Towards Single Camera Human 3D-Kinematics.” In: *Sensors* 23.1 (Jan. 2023), p. 341. ISSN: 1424-8220. doi: [10.3390/s23010341](https://doi.org/10.3390/s23010341). (Visited on 10/14/2023).
- [18] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. “BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 8726–8737. doi: [10.1109/cvpr52729.2023.00843](https://doi.org/10.1109/cvpr52729.2023.00843).

- [19] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. "Align Your Latents: High-Resolution Video Synthesis with Latent Diffusion Models." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 22563–22575. doi: [10.1109/cvpr52729.2023.02161](https://doi.org/10.1109/cvpr52729.2023.02161).
- [20] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. "Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image." In: *Proceedings of the European Conference on Computer Vision*. 2016, pp. 561–578. doi: [10.1007/978-3-319-46454-1_34](https://doi.org/10.1007/978-3-319-46454-1_34).
- [21] Andrew Brock, Jeff Donahue, and Karen Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis." In: *International Conference on Learning Representations*. 2019. URL: <https://openreview.net/forum?id=B1xsqj09Fm>.
- [22] XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. "Mm-net: A model-based multimodal network for human action recognition in rgb-d videos." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (2022), pp. 3522–3538. doi: [10.1109/TPAMI.2022.3177813](https://doi.org/10.1109/TPAMI.2022.3177813).
- [23] Alexander G. Bruno, Mary L. Buxsein, and Dennis E. Anderson. "Development and Validation of a Musculoskeletal Model of the Fully Articulated Thoracolumbar Spine and Rib Cage." In: *Journal of Biomechanical Engineering* 137.8 (Aug. 2015), p. 081003. issn: 1528-8951. doi: [10.1115/1.4030408](https://doi.org/10.1115/1.4030408).
- [24] Angel Bu, Mhairi K. MacLean, and Daniel P. Ferris. "EMG-informed neuromuscular model assesses the effects of varied bodyweight support on muscles during overground walking." In: *Journal of Biomechanics* 151 (2023), p. 111532. doi: [10.1016/j.jbiomech.2023.111532](https://doi.org/10.1016/j.jbiomech.2023.111532).
- [25] Daniel J Butler, Justin Huang, Franziska Roesner, and Maya Cakmak. "The privacy-utility tradeoff for remotely teleoperated robots." In: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 2015, pp. 27–34. doi: [10.1145/2696454.2696484](https://doi.org/10.1145/2696454.2696484).
- [26] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A Dos Santos, and William Robson Schwartz. "Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition." In: *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE. 2019, pp. 1–8. doi: [10.1109/avss.2019.8909840](https://doi.org/10.1109/avss.2019.8909840).
- [27] Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. "MyoSuite: A Contact-rich Simulation Suite for Musculoskeletal Motor Control." In: *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*. Vol. 168. Proceedings of Machine Learning Research. PMLR, 2022, pp. 492–507. URL: <https://proceedings.mlr.press/v168/caggiano22a.html>.
- [28] Jonathan Camargo, Aditya Ramanathan, Will Flanagan, and Aaron Young. "A comprehensive, open-source dataset of lower limb biomechanics in multiple conditions of stairs, ramps, and level-ground ambulation and transitions." In: *Journal of Biomechanics* 119 (2021), p. 110320. doi: [10.1016/j.jbiomech.2021.110320](https://doi.org/10.1016/j.jbiomech.2021.110320).

- [29] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. “Extracting training data from diffusion models.” In: *32nd USENIX security symposium (USENIX Security 23)*. 2023, pp. 5253–5270. arXiv: [2301.13188](#). URL: <https://www.usenix.org/conference/usenixsecurity23/presentation/carlini>.
- [30] João Carreira and Andrew Zisserman. “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset.” In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4724–4733. DOI: [10.1109/CVPR.2017.502](#).
- [31] Angel X. Chang et al. “ShapeNet: An Information-Rich 3D Model Repository.” In: *arXiv preprint* (2015). arXiv: [1512.03012](#).
- [32] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. “GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 1. 2019, pp. 8126–8133. DOI: [10.1609/aaai.v33i01.33018126](#).
- [33] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. “Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification.” In: *Journal of Electronic Imaging* 22.4 (July 2013). Publisher: SPIE, p. 041106. ISSN: 1017-9909, 1560-229X. DOI: [10.1117/1.JEI.22.4.041106](#). (Visited on 12/06/2024).
- [34] Chun-Fu Chen, Shaohan Hu, Zhonghao Shi, Prateek Gulati, Bill Moriarty, Marco Pistoia, Vincenzo Piuri, and Pierangela Samarati. “MaSS: Multi-attribute Selective Suppression.” In: *arXiv preprint* (2022). arXiv: [2210.09904](#).
- [35] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. “Temporal attentive alignment for large-scale video domain adaptation.” In: *IEEE International Conference on Computer Vision (ICCV)*. 2019, pp. 6320–6329. DOI: [10.1109/iccv.2019.00642](#).
- [36] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A Simple Framework for Contrastive Learning of Visual Representations.” In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1597–1607. URL: <https://proceedings.mlr.press/v119/chen20j.html>.
- [37] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. “Human daily action analysis with multi-view and color-depth data.” In: *Computer Vision–ECCV 2012. Workshops and Demonstrations: Florence, Italy, October 7–13, 2012, Proceedings, Part II* 12. Springer. 2012, pp. 52–61. DOI: [10.1007/978-3-642-33868-7_6](#).
- [38] Mia Chiquier and Carl Vondrick. “Muscles in action.” In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 22034–22044. DOI: [10.1109/ICCV51070.2023.02019](#).
- [39] Edward Chou, Matthew Tan, Cherry Zou, Michelle Guo, Albert Haque, Arnold Milstein, and Li Fei-Fei. “Privacy-preserving action recognition for smart hospitals using low-resolution depth images.” In: *arXiv preprint* (2018). arXiv: [1811.09950](#).

- [40] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. "Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package)." In: *Neurocomputing* 307 (2018), pp. 72–77. DOI: [10.1016/j.neucom.2018.03.067](https://doi.org/10.1016/j.neucom.2018.03.067).
- [41] Eftychios G Christoforou, Andreas S Panayides, Sotiris Avgousti, Panicos Moursas, and Constantinos S Pattichis. "An overview of assistive robotics and technologies for elderly care." In: *Mediterranean Conference on Medical and Biological Engineering and Computing*. Springer. 2019, pp. 971–976. DOI: [10.1007/978-3-030-31635-8_118](https://doi.org/10.1007/978-3-030-31635-8_118).
- [42] Jia-Luen Chua, Yoong Choon Chang, and Wee Keong Lim. "A simple vision-based fall detection technique for indoor video surveillance." In: *Signal, Image and Video Processing* 9 (2015), pp. 623–633. DOI: [10.1007/s11760-013-0493-7](https://doi.org/10.1007/s11760-013-0493-7).
- [43] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation. Stichting Blender Foundation, Amsterdam, 2018. URL: <http://www.blender.org>.
- [44] Victor G. Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. "Dual-Head Contrastive Domain Adaptation for Video Action Recognition." In: *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2022, pp. 2234–2243. DOI: [10.1109/WACV51458.2022.00229](https://doi.org/10.1109/WACV51458.2022.00229).
- [45] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. "Randaugment: Practical automated data augmentation with a reduced search space." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 702–703. DOI: [10.1109/cvprw50498.2020.00359](https://doi.org/10.1109/cvprw50498.2020.00359).
- [46] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." In: *Advances in Neural Information Processing Systems*. Vol. 26. 2013. URL: https://papers.nips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.
- [47] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. "Toyota Smarthome Untrimmed: Real-World Untrimmed Videos for Activity Detection." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.2 (2023), pp. 2533–2550. DOI: [10.1109/TPAMI.2022.3169976](https://doi.org/10.1109/TPAMI.2022.3169976).
- [48] Dima Damen et al. "Scaling Egocentric Vision: The EPIC-KITCHENS Dataset." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 720–736. URL: https://openaccess.thecvf.com/content_ECCV_2018/html/Dima_Damen_Scaling_Egocentric_Vision_ECCV_2018_paper.html.
- [49] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. "DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation." In: *European Conference on Computer Vision (ECCV)*. 2018, pp. 467–483. DOI: [10.1007/978-3-030-01225-0_28](https://doi.org/10.1007/978-3-030-01225-0_28).

- [50] Tri Dao and Albert Gu. “Transformers are SSMs: generalized models and efficient algorithms through structured state space duality.” In: *Proceedings of the 41st International Conference on Machine Learning*. ICML’24. Vienna, Austria: JMLR.org, 2024. URL: <https://openreview.net/forum?id=ztn8FCR1td>.
- [51] Srijan Das, Arpit Chaudhary, Francois Bremond, and Monique Thonnat. “Where to focus on for human action recognition?” In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 71–80. DOI: [10.1109/wacv.2019.00015](https://doi.org/10.1109/wacv.2019.00015).
- [52] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. “Toyota Smarthome: Real-World Activities of Daily Living.” In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 833–842. DOI: [10.1109/ICCV.2019.00092](https://doi.org/10.1109/ICCV.2019.00092).
- [53] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. “Vpn++: Rethinking video-pose embeddings for understanding activities of daily living.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (2021), pp. 9703–9717. DOI: [10.1109/tpami.2021.3127885](https://doi.org/10.1109/tpami.2021.3127885).
- [54] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. “Vpn: Learning video-pose embedding for activities of daily living.” In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX* 16. Springer. 2020, pp. 72–90. DOI: [10.1007/978-3-030-58545-7_5](https://doi.org/10.1007/978-3-030-58545-7_5).
- [55] Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. “Spact: Self-supervised privacy preservation for action recognition.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20164–20173. DOI: [10.1109/cvpr52688.2022.01953](https://doi.org/10.1109/cvpr52688.2022.01953).
- [56] César Roberto De Souza, Adrien Gaidon, Yohann Cabon, and Antonio Manuel López. “Procedural Generation of Videos to Train Deep Action Recognition Networks.” In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2594–2604. DOI: [10.1109/CVPR.2017.278](https://doi.org/10.1109/CVPR.2017.278).
- [57] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. “Unlocking high-accuracy differentially private image classification through scale.” In: *arXiv preprint* (2022). arXiv: [2204.13650](https://arxiv.org/abs/2204.13650).
- [58] Scott L. Delp, Frank C. Anderson, Allison S. Arnold, Peter Loan, Ayman Habib, Chand T. John, Eran Guendelman, and Darryl G. Thelen. “OpenSim: Open-Source Software to Create and Analyze Dynamic Simulations of Movement.” In: *IEEE Transactions on Biomedical Engineering* 54.11 (Nov. 2007), pp. 1940–1950. ISSN: 1558-2531. DOI: [10.1109/TBME.2007.901024](https://doi.org/10.1109/TBME.2007.901024). (Visited on 10/25/2023).
- [59] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database.” In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255. DOI: [10.1109/cvpr.2009.5206848](https://doi.org/10.1109/cvpr.2009.5206848).
- [60] Stefan Denkovski, Shehroz S. Khan, Brandon Malamis, Sae Young Moon, Bing Ye, and Alex Mihailidis. “Multi Visual Modality Fall Detection Dataset.” In: *IEEE Access* 10 (2022). Conference Name: IEEE Access, pp. 106422–106435. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2022.3211939](https://doi.org/10.1109/ACCESS.2022.3211939). (Visited on 12/06/2024).

- [61] Chhavi Dhiman and Dinesh Kumar Vishwakarma. “View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics.” In: *IEEE Transactions on Image Processing* 29 (2020), pp. 3835–3844. doi: [10.1109/tip.2020.2965299](https://doi.org/10.1109/tip.2020.2965299).
- [62] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. “Large scale holistic video understanding.” In: *European Conference on Computer Vision*. Springer. 2020, pp. 593–610. doi: [10.1007/978-3-030-58558-7_35](https://doi.org/10.1007/978-3-030-58558-7_35).
- [63] My H Dinh and Ferdinando Fioretto. “Context-Aware Differential Privacy for Language Modeling.” In: *arXiv preprint* (2023). arXiv: [2301.12288](https://arxiv.org/abs/2301.12288).
- [64] Friedrich Dörmann, Osvald Frisk, Lars Nørvang Andersen, and Christian Fischer Pedersen. “Not all noise is accounted equally: How differentially private learning benefits from large sampling rates.” In: *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2021, pp. 1–6. doi: [10.1109/MLSP52302.2021.9596307](https://doi.org/10.1109/MLSP52302.2021.9596307).
- [65] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=YicbFdNTTy>.
- [66] Abhishek Dutta and Andrew Zisserman. “The VIA Annotation Software for Images, Audio and Video.” In: *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 2019, pp. 2276–2279. doi: [10.1145/3343031.3350535](https://doi.org/10.1145/3343031.3350535).
- [67] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. “Calibrating noise to sensitivity in private data analysis.” In: *Theory of cryptography conference*. Springer. 2006, pp. 265–284. doi: [10.1007/11681878_14](https://doi.org/10.1007/11681878_14).
- [68] Dave Epstein, Boyuan Chen, and Carl Vondrick. “Oops! Predicting Unintentional Action in Video.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 919–929. doi: [10.1109/cvpr42600.2020.00100](https://doi.org/10.1109/cvpr42600.2020.00100).
- [69] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. “A survey on bias in visual datasets.” In: *Computer Vision and Image Understanding* 223 (2022), p. 103552. doi: [10.1016/j.cviu.2022.103552](https://doi.org/10.1016/j.cviu.2022.103552).
- [70] Antoine Falisse, Gil Serrancoli, Christopher L. Dembia, Joris Gillis, and Friedl De Groote. “Algorithmic differentiation improves the computational efficiency of OpenSim-based trajectory optimization of human movement.” In: *PLOS ONE* 14.10 (Oct. 17, 2019), e0217730. ISSN: 1932-6203. doi: [10.1371/journal.pone.0217730](https://doi.org/10.1371/journal.pone.0217730). (Visited on 03/04/2024).
- [71] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. “GaitPart: Temporal Part-based Model for Gait Recognition.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 14225–14233. doi: [10.1109/cvpr42600.2020.01423](https://doi.org/10.1109/cvpr42600.2020.01423).

- [72] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. "Multiscale vision transformers." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6824–6835. DOI: [10.1109/iccv48922.2021.00675](https://doi.org/10.1109/iccv48922.2021.00675).
- [73] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. "RMPE: Regional Multi-person Pose Estimation." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2353–2362. DOI: [10.1109/iccv.2017.256](https://doi.org/10.1109/iccv.2017.256).
- [74] Gunnar Farneback. "Two-frame motion estimation based on polynomial expansion." In: *Scandinavian conference on Image analysis*. Springer. 2003, pp. 363–370. DOI: [10.1007/3-540-45103-x_50](https://doi.org/10.1007/3-540-45103-x_50).
- [75] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. "SlowFast Networks for Video Recognition." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 6201–6210. DOI: [10.1109/ICCV.2019.00630](https://doi.org/10.1109/ICCV.2019.00630).
- [76] Benedikt Feldotto, Cristian Soare, Alois Knoll, Piyanee Sriya, Sarah Astill, Marc de Kamps, and Samit Chakrabarty. "Evaluating Muscle Synergies With EMG Data and Physics Simulation in the Neurorobotics Platform." In: *Frontiers in Neurorobotics Volume 16 - 2022* (2022). ISSN: 1662-5218. DOI: [10.3389/fnbot.2022.856797](https://doi.org/10.3389/fnbot.2022.856797).
- [77] Yusen Feng, Xiyan Xu, and Libin Liu. "MuscleVAE: Model-Based Controllers of Muscle-Actuated Characters." In: *SIGGRAPH Asia 2023 Conference Papers*. 2023, pp. 1–11. DOI: [10.1145/3610548.3618137](https://doi.org/10.1145/3610548.3618137).
- [78] Joseph Fiorese, Ishan Rajendrakumar Dave, and Mubarak Shah. "Ted-spade: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 13598–13609. DOI: [10.1109/iccv51070.2023.01251](https://doi.org/10.1109/iccv51070.2023.01251).
- [79] Jane Fleming and Carol Brayne. "Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90." In: *BMJ* 337 (2008), a2227. DOI: [10.1136/bmj.a2227](https://doi.org/10.1136/bmj.a2227).
- [80] Jie Fu, Zhili Chen, and Xiao Han. "Adap DP-FL: Differentially Private Federated Learning with Adaptive Noise." In: *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE. 2022, pp. 656–663. DOI: [10.1109/trustcom56396.2022.00094](https://doi.org/10.1109/trustcom56396.2022.00094).
- [81] Mariusz P Furmanek, Madhur Mangalam, Mathew Yarossi, Kyle Lockwood, and Eugene Tunik. "A kinematic and EMG dataset of online adjustment of reach-to-grasp movements to visual perturbations." In: *Scientific data* 9.1 (2022), p. 23. DOI: [10.1038/s41597-021-01107-2](https://doi.org/10.1038/s41597-021-01107-2).
- [82] Ziliang Gan, Lei Jin, Lei Nie, Zheng Wang, Li Zhou, Liang Li, Zhecan Wang, Jianshu Li, Junliang Xing, and Jian Zhao. "ASQuery: A Query-based Model for Action Segmentation." In: *2024 IEEE International Conference on Multimedia and Expo (ICME)* (2024), pp. i–vi. DOI: [10.1109/icme57554.2024.10687535](https://doi.org/10.1109/icme57554.2024.10687535).

- [83] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. "Domain-Adversarial Training of Neural Networks." In: *Journal of Machine Learning Research* 17.59 (2016), pp. 1–35. URL: <http://jmlr.org/papers/v17/15-239.html>.
- [84] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. "Convolutional Sequence to Sequence Learning." In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1243–1252. URL: <https://proceedings.mlr.press/v70/gehring17a.html>.
- [85] Mona Fathollahi Ghezelghieh, Rangachar Kasturi, and Sudeep Sarkar. "Learning camera viewpoint using CNN to improve 3D body pose estimation." In: *2016 fourth international conference on 3D vision (3DV)*. IEEE. 2016, pp. 685–693. doi: [10.1109/3dv.2016.75](https://doi.org/10.1109/3dv.2016.75).
- [86] Saeed Ghorbani, Kimia Mahdavian, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. "MoVi: A large multi-purpose human motion and video dataset." In: *PLoS ONE* 16.6 (2021), e0253157. doi: [10.1371/journal.pone.0253157](https://doi.org/10.1371/journal.pone.0253157).
- [87] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. "Omnivore: A single model for many visual modalities." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16102–16112. doi: [10.1109/cvpr52688.2022.01563](https://doi.org/10.1109/cvpr52688.2022.01563).
- [88] Athanasios Gkrekidis, Georgios Giarmatzis, Dimitrios Menychtas, Evangelos Karakasis, Vassilios Gourgoulis, Maria Michalopoulou, Ilias Smilios, Helen T. Douda, Georgios Ch. Sirakoulis, and Nikolaos Aggelousis. "Clinical Validation of Estimated Muscle Activations during Phases of Elderly Gait." In: *Biomechanics* 3.4 (2023), pp. 552–560. doi: [10.3390/biomechanics3040044](https://doi.org/10.3390/biomechanics3040044).
- [89] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. "Humans in 4D: Reconstructing and Tracking Humans with Transformers." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 14737–14748. doi: [10.1109/ICCV51070.2023.01358](https://doi.org/10.1109/ICCV51070.2023.01358).
- [90] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014. URL: https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.
- [91] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. "Actions as space-time shapes." In: *IEEE transactions on pattern analysis and machine intelligence* 29.12 (2007), pp. 2247–2253. doi: [10.1109/TPAMI.2007.70711](https://doi.org/10.1109/TPAMI.2007.70711).
- [92] Raghav Goyal et al. "The 'something something' video database for learning and evaluating visual common sense." In: *IEEE international conference on computer vision* (2017), pp. 5842–5850. doi: [10.1109/iccv.2017.622](https://doi.org/10.1109/iccv.2017.622).
- [93] Chunhui Gu et al. "AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6047–6056. doi: [10.1109/cvpr.2018.00633](https://doi.org/10.1109/cvpr.2018.00633).

- [94] José Camilo Eraso Guerrero, Elena Muñoz España, Mariela Muñoz Añasco, and Jesús Emilio Pinto Lopera. “Dataset for human fall recognition in an uncontrolled environment.” In: *Data in Brief* 45 (Dec. 1, 2022), p. 108610. issn: 2352-3409. doi: [10.1016/j.dib.2022.108610](https://doi.org/10.1016/j.dib.2022.108610). (Visited on 12/06/2024).
- [95] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. “Improved Training of Wasserstein GANs.” In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/892c3b1c6dccd52936e27cbd0ff683d6-Abstract.html.
- [96] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. “Generating diverse and natural 3d human motions from text.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5142–5151. doi: [10.1109/cvpr52688.2022.00509](https://doi.org/10.1109/cvpr52688.2022.00509).
- [97] Xi Guo, Wei Wu, Dongliang Wang, Jing Su, Haisheng Su, Weihao Gan, Jian Huang, and Qin Yang. “Learning Video Representations of Human Motion from Synthetic Data.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 20165–20175. doi: [10.1109/CVPR52688.2022.01956](https://doi.org/10.1109/CVPR52688.2022.01956).
- [98] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. “AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning.” In: *International Conference on Learning Representations* (2024). URL: <https://openreview.net/forum?id=k0UgAl3EVq>.
- [99] Simon Hanisch, Julian Todt, and Thorsten Strufe. “Pantomime: Motion Data Anonymization Using Foundation Motion Models.” In: *Proceedings on Privacy Enhancing Technologies* 2026.2 (2026), pp. 384–396. doi: [10.56553/popets-2026-0053](https://doi.org/10.56553/popets-2026-0053).
- [100] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. “Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA*. 2018, pp. 18–22. doi: [10.1109/cvpr.2018.00685](https://doi.org/10.1109/cvpr.2018.00685).
- [101] Kaibo He, Chenhui Zuo, Chengtian Ma, and Yanan Sui. “DynSyn: Dynamical Synergistic Representation for Efficient Learning and Control in Overactuated Embodied Systems.” In: *International Conference on Machine Learning (ICML)*. PMLR, 2024, pp. 18115–18132. URL: <https://openreview.net/forum?id=q0MQ0UGLYL>.
- [102] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778. doi: [10.1109/cvpr.2016.90](https://doi.org/10.1109/cvpr.2016.90).
- [103] Óscar G. Hernández, Jose M. Lopez-Castellanos, Carlos A. Jara, Gabriel J. Garcia, Andres Ubeda, Vicente Morell-Gimenez, and Francisco Gomez-Donoso. “A kinematic, imaging and electromyography dataset for human muscular manipulability index prediction.” In: *Scientific Data* 10.1 (Mar. 2023), p. 132. issn: 2052-4463. doi: [10.1038/s41597-023-02031-3](https://doi.org/10.1038/s41597-023-02031-3).
- [104] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. “Distilling the Knowledge in a Neural Network.” In: *NIPS Deep Learning and Representation Learning Workshop*. 2015. arXiv: [1503.02531](https://arxiv.org/abs/1503.02531).

- [105] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising Diffusion Probabilistic Models.” In: *Advances in Neural Information Processing Systems*. 2020, pp. 6840–6851. URL: https://proceedings.neurips.cc/paper_files/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.
- [106] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David Fleet. “Video Diffusion Models.” In: 35 (2022), pp. 8633–8646. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/39235c56aef13fb05a6adc95eb9d8d66-Abstract-Conference.html.
- [107] Jonathan Ho et al. “Imagen Video: High Definition Video Generation with Diffusion Models.” In: *arXiv preprint* (2022). arXiv: [2210.02303](https://arxiv.org/abs/2210.02303).
- [108] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory.” In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [109] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. “ElderSim: A Synthetic Data Generation Platform for Human Action Recognition in Eldercare Applications.” In: *IEEE Access* 9 (2021), pp. 24237–24248. DOI: [10.1109/ACCESS.2021.3051842](https://doi.org/10.1109/ACCESS.2021.3051842).
- [110] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.” In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 448–456. URL: <https://proceedings.mlr.press/v37/ioffe15.html>.
- [111] Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. “Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 12627–12637. DOI: [10.1109/cvpr.2019.01291](https://doi.org/10.1109/cvpr.2019.01291).
- [112] Jinhyeok Jang, Dohyung Kim, Cheonshu Park, Minsu Jang, Jaeyeon Lee, and Jaehong Kim. “ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly.” In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020, pp. 10990–10997. DOI: [10.1109/IROS45743.2020.9341160](https://doi.org/10.1109/IROS45743.2020.9341160).
- [113] Néstor J Jarque-Bou, Margarita Vergara, Joaquín L Sancho-Bru, Verónica Gracia-Ibáñez, and Alba Roda-Sales. “A calibrated database of kinematics and EMG of the forearm and hand during activities of daily living.” In: *Scientific data* 6.1 (2019), p. 270. DOI: [10.1038/s41597-019-0285-1](https://doi.org/10.1038/s41597-019-0285-1).
- [114] Bargav Jayaraman and David Evans. “Evaluating Differentially Private Machine Learning in Practice.” In: *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 1895–1912. ISBN: 978-1-939133-06-9. arXiv: [1902.08874](https://arxiv.org/abs/1902.08874). URL: <https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman>.
- [115] Yifeng Jiang, Tom Van Wouwe, Friedl De Groote, and C Karen Liu. “Synthesis of biologically realistic human motion using joint torque actuation.” In: *ACM Transactions On Graphics (TOG)* 38.4 (2019), pp. 1–12. DOI: [10.1145/3306346.3322966](https://doi.org/10.1145/3306346.3322966).

- [116] Zach Jorgensen, Ting Yu, and Graham Cormode. "Conservative or liberal? Personalized differential privacy." In: *2015 IEEE 31st international conference on data engineering*. IEEE. 2015, pp. 1023–1034. DOI: [10.1109/icde.2015.7113353](https://doi.org/10.1109/icde.2015.7113353).
- [117] Indu Joshi, Marcel Grimmer, Christian Rathgeb, Christoph Busch, Francois Bremond, and Antitza Dantcheva. "Synthetic Data in Human Analysis: A Survey." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.7 (2024), pp. 4957–4976. DOI: [10.1109/TPAMI.2024.3362821](https://doi.org/10.1109/TPAMI.2024.3362821).
- [118] State of California Department of Justice (.gov). *California Consumer Privacy Act (CCPA)*. 2023. URL: <https://oag.ca.gov/privacy/ccpa>.
- [119] Rie Kamikubo, Lining Wang, Crystal Marte, Amnah Mahmood, and Hernisa Kacorri. "Data Representativeness in Accessibility Datasets: A Meta-Analysis." In: *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*. 2022, pp. 1–15. DOI: [10.1145/3517428.3544826](https://doi.org/10.1145/3517428.3544826).
- [120] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. "End-to-End Recovery of Human Shape and Pose." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7122–7131. DOI: [10.1109/cvpr.2018.00744](https://doi.org/10.1109/cvpr.2018.00744).
- [121] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. "Large-Scale Video Classification with Convolutional Neural Networks." In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1725–1732. DOI: [10.1109/CVPR.2014.223](https://doi.org/10.1109/CVPR.2014.223).
- [122] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4401–4410. DOI: [10.1109/cvpr.2019.00453](https://doi.org/10.1109/cvpr.2019.00453).
- [123] Will Kay et al. "The Kinetics Human Action Video Dataset." In: *arXiv preprint* (2017). arXiv: [1705.06950](https://arxiv.org/abs/1705.06950).
- [124] TS Keller, AM Weisberger, JL Ray, SS Hasan, RG Shiavi, and DM Spengler. "Relationship between vertical ground reaction force and speed during walking, slow jogging, and running." In: *Clinical Biomechanics* 11.5 (1996), pp. 253–259. DOI: [10.1016/0268-0033\(95\)00068-2](https://doi.org/10.1016/0268-0033(95)00068-2).
- [125] Doyoung Kim, Inwoong Lee, Dohyung Kim, and Sanghoon Lee. "Action recognition using close-up of maximum activation and etri-activity3d livinglab dataset." In: *Sensors* 21.20 (2021), p. 6774. DOI: [10.3390/s21206774](https://doi.org/10.3390/s21206774).
- [126] Václav Knapp and Maty Bohacek. "Synthetic Human Action Video Data Generation with Pose Transfer." In: *Synthetic Data for Computer Vision Workshop @ CVPR 2025*. 2025. URL: <https://openreview.net/forum?id=KTXL0idiky>.
- [127] Muhammed Kocabas, Chun-Hao Paul Huang, Otmar Hilliges, and Michael J. Black. "PARE: Part Attention Regressor for 3D Human Body Estimation." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11127–11137. DOI: [10.1109/iccv48922.2021.01094](https://doi.org/10.1109/iccv48922.2021.01094).

- [128] Amir Kolaman, Dan Malowany, Rami R Hagege, and Hugo Guterman. "Light invariant video imaging for improved performance of convolution neural networks." In: *IEEE Transactions on Circuits and Systems for Video Technology* 29.6 (2018), pp. 1584–1594. doi: [10.1109/tcsvt.2018.2846695](https://doi.org/10.1109/tcsvt.2018.2846695).
- [129] Ios Kotsogiannis, Stelios Doudalis, Sam Haney, Ashwin Machanavajjhala, and Sharad Mehrotra. "One-sided differential privacy." In: *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE. 2020, pp. 493–504. doi: [10.1109/icde48307.2020.00049](https://doi.org/10.1109/icde48307.2020.00049).
- [130] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: 25 (2012). URL: https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.
- [131] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. "HMDB: a large video database for human motion recognition." In: *2011 International conference on computer vision*. IEEE. 2011, pp. 2556–2563. doi: [10.1109/iccv.2011.6126543](https://doi.org/10.1109/iccv.2011.6126543).
- [132] Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. "Toward training at imagenet scale with differential privacy." In: *arXiv preprint* (2022). arXiv: [2201.12328](https://arxiv.org/abs/2201.12328).
- [133] Ines Kutzner, Adam Trepczynski, Markus O. Heller, and Georg Bergmann. "Knee adduction moment and medial contact force—facts about their correlation during gait." In: *PloS one* 8.12 (2013), e81036. doi: [10.1371/journal.pone.0081036](https://doi.org/10.1371/journal.pone.0081036).
- [134] Bogdan Kwolek and Michal Kepski. "Human fall detection on embedded platform using depth maps and wireless accelerometer." In: *Computer Methods and Programs in Biomedicine* 117.3 (Dec. 1, 2014), pp. 489–501. ISSN: 0169-2607. doi: [10.1016/j.cmpb.2014.09.005](https://doi.org/10.1016/j.cmpb.2014.09.005). (Visited on 12/06/2024).
- [135] Adrian KM Lai, Allison S Arnold, and James M Wakeling. "Why are antagonist muscles co-activated in my simulation? A musculoskeletal model for analysing human locomotor tasks." In: *Annals of Biomedical Engineering* 45 (2017), pp. 2762–2774. doi: [10.1007/s10439-017-1920-7](https://doi.org/10.1007/s10439-017-1920-7).
- [136] Tiziana Lencioni, Ilaria Carpinella, Marco Rabuffetti, Alberto Marzegan, and Maurizio Ferrarin. "Human kinematic, kinetic and EMG data during different walking and stair ascending and descending tasks." In: *Scientific data* 6.1 (2019), p. 309. doi: [10.1038/s41597-019-0323-z](https://doi.org/10.1038/s41597-019-0323-z).
- [137] Chengshu Li et al. "iGibson 2.0: Object-Centric Simulation for Robot Learning of Everyday Household Tasks." In: *Proceedings of the 5th Conference on Robot Learning*. Vol. 164. Proceedings of Machine Learning Research. PMLR, 2022, pp. 455–465. URL: <https://proceedings.mlr.press/v164/li22b.html>.
- [138] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. "Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 3041–3050. doi: [10.1109/CVPR52729.2023.00297](https://doi.org/10.1109/CVPR52729.2023.00297).

- [139] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. "CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10855–10864. DOI: [10.1109/CVPR.2019.01112](https://doi.org/10.1109/CVPR.2019.01112).
- [140] Ming Li, Xiangyu Xu, Hehe Fan, Pan Zhou, Jun Liu, Jia-Wei Liu, Jiahe Li, Jussi Keppo, Mike Zheng Shou, and Shuicheng Yan. "STPrivacy: Spatio-Temporal Privacy-Preserving Action Recognition." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 5083–5092. DOI: [10.1109/ICCV51070.2023.00471](https://doi.org/10.1109/ICCV51070.2023.00471).
- [141] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. "Large Language Models Can Be Strong Differentially Private Learners." In: *International Conference on Learning Representations (ICLR)*. 2022. URL: <https://openreview.net/forum?id=bVuP3ltATMz>.
- [142] Y Li, CY Wu, H Fan, K Mangalam, B Xiong, J Malik, and C Feichtenhofer. "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 4794–4804. DOI: [10.1109/CVPR52688.2022.00476](https://doi.org/10.1109/CVPR52688.2022.00476).
- [143] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. "Towards An End-to-End Framework for Flow-Guided Video Inpainting." In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 17541–17550. DOI: [10.1109/cvpr52688.2022.01704](https://doi.org/10.1109/cvpr52688.2022.01704).
- [144] Beibei Lin, Shunli Zhang, Ming Wang, Lincheng Li, and Xin Yu. "Gaitgl: Learning discriminative global-local feature representations for gait recognition." In: *arXiv preprint (2022)*. arXiv: [2208.01380](https://arxiv.org/abs/2208.01380).
- [145] Ji Lin, Chuang Gan, and Song Han. "TSM: Temporal Shift Module for Efficient Video Understanding." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 7083–7093. DOI: [10.1109/ICCV.2019.00718](https://doi.org/10.1109/ICCV.2019.00718).
- [146] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context." In: *European conference on computer vision*. Springer. 2014, pp. 740–755. DOI: [10.1007/978-3-319-10602-1_48](https://doi.org/10.1007/978-3-319-10602-1_48).
- [147] Jian Liu, Hossein Rahmani, Naveed Akhtar, and Ajmal Mian. "Learning human pose models from synthesized data for robust RGB-D action recognition." In: *International Journal of Computer Vision* 127 (2019), pp. 1545–1564. DOI: [10.1007/s11263-019-01192-2](https://doi.org/10.1007/s11263-019-01192-2).
- [148] Xinpeng Liu, Junxuan Liang, Chenshuo Zhang, Zixuan Cai, Cewu Lu, and Yong-Lu Li. "Homogeneous Dynamics Space for Heterogeneous Humans." In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 27782–27793. DOI: [10.1109/cvpr52734.2025.02587](https://doi.org/10.1109/cvpr52734.2025.02587).
- [149] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022. DOI: [10.1109/iccv48922.2021.00986](https://doi.org/10.1109/iccv48922.2021.00986).

- [150] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. "SMPL: A Skinned Multi-Person Linear Model." In: *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16. doi: [10.1145/2816795.2818013](https://doi.org/10.1145/2816795.2818013).
- [151] Dennis Ludl, Thomas Gulde, and Cristóbal Curio. "Enhancing data-driven algorithms for human pose estimation and action recognition through simulation." In: *IEEE transactions on intelligent transportation systems* 21.9 (2020), pp. 3990–3999. doi: [10.1109/tits.2020.2988504](https://doi.org/10.1109/tits.2020.2988504).
- [152] Zelun Luo, Yuliang Zou, Yijin Yang, Zane Durante, De-An Huang, Zhiding Yu, Chaowei Xiao, Li Fei-Fei, and Animashree Anandkumar. "Differentially Private Video Activity Recognition." In: *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024, pp. 6643–6653. doi: [10.1109/WACV57701.2024.00652](https://doi.org/10.1109/WACV57701.2024.00652).
- [153] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE." In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: <http://jmlr.org/papers/v9/vandemaaten08a.html>.
- [154] Behrooz Mahasseni and Sinisa Todorovic. "Regularizing long short term memory with 3D human-skeleton sequences for action recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 3054–3062. doi: [10.1109/cvpr.2016.333](https://doi.org/10.1109/cvpr.2016.333).
- [155] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. "AMASS: Archive of Motion Capture as Surface Shapes." In: *International Conference on Computer Vision*. Oct. 2019, pp. 5442–5451. doi: [10.1109/iccv.2019.00554](https://doi.org/10.1109/iccv.2019.00554).
- [156] Saturnino Maldonado-Bascon, Cristian Iglesias-Iglesias, Pilar Martín-Martín, and Sergio Lafuente-Arroyo. "Fallen people detection capabilities using assistive robot." In: *Electronics* 8.9 (2019), p. 915. doi: [10.3390/electronics8090915](https://doi.org/10.3390/electronics8090915).
- [157] Nebojša Malešević, Alexander Olsson, Paulina Sager, Elin Andersson, Christian Cipriani, Marco Controzzi, Anders Björkman, and Christian Antfolk. "A database of high-density surface electromyogram signals comprising 65 isometric hand gestures." In: *Scientific Data* 8.1 (2021), p. 63. doi: [10.1038/s41597-021-00843-9](https://doi.org/10.1038/s41597-021-00843-9).
- [158] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. "The KIT whole-body human motion database." In: *2015 International Conference on Advanced Robotics (ICAR)*. IEEE. 2015, pp. 329–336. doi: [10.1109/icar.2015.7251476](https://doi.org/10.1109/icar.2015.7251476).
- [159] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. "Unifying Representations and Large-Scale Whole-Body Motion Databases for Studying Human Motion." In: *IEEE Transactions on Robotics* (2016), pp. 796–809. doi: [10.1109/tro.2016.2572685](https://doi.org/10.1109/tro.2016.2572685).
- [160] Leo Marco and Giovanni Maria Farinella. *Computer Vision for Assistive Healthcare*. 1st. GBR: Academic Press Ltd., 2018. ISBN: 0128134453. URL: <https://dl.acm.org/doi/10.5555/3235962>.

- [161] Zdravko Marinov, Alina Roitberg, David Schneider, and Rainer Stiefelhagen. "ModSelect: Automatic Modality Selection for Synthetic-to-Real Domain Generalization." In: *Computer Vision—ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*. Springer. 2023, pp. 326–346. doi: [10.1007/978-3-031-25085-9_19](https://doi.org/10.1007/978-3-031-25085-9_19).
- [162] Zdravko Marinov*, David Schneider*, Alina Roitberg*, and Rainer Stiefelhagen. "Multimodal Generation of Novel Action Appearances for Synthetic-to-Real Recognition of Activities of Daily Living." In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 11320–11327. doi: [10.1109/iros47612.2022.9981946](https://doi.org/10.1109/iros47612.2022.9981946).
- [163] Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafort-Asturiano. "UP-Fall Detection Dataset: A Multimodal Approach." In: *Sensors* 19.9 (Jan. 2019). Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 1988. issn: 1424-8220. doi: [10.3390/s19091988](https://doi.org/10.3390/s19091988). (Visited on 12/06/2024).
- [164] Ana Matran-Fernandez, Itzel Jared Rodríguez Martínez, Riccardo Poli, Christian Cipriani, and Luca Citi. "SEEDS, simultaneous recordings of high-density EMG and finger joint angles during multiple hand movements." In: *Scientific data* 6.1 (2019), p. 186. doi: [10.1038/s41597-019-0200-9](https://doi.org/10.1038/s41597-019-0200-9).
- [165] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Damos, Greg Damos, Lynn He, Alicia Parrish, Hannah Rose Kirk, et al. "Dataperf: Benchmarks for data-centric ai development." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 5320–5347. url: <https://openreview.net/forum?id=LaFKTgrZMG>.
- [166] Jill L McNitt-Gray. "Kinetics of the lower extremities during drop landings from three heights." In: *Journal of Biomechanics* 26.9 (1993), pp. 1037–1046. doi: [10.1016/s0021-9290\(05\)80003-x](https://doi.org/10.1016/s0021-9290(05)80003-x).
- [167] Celso M. de Melo, Brandon Rothrock, Prudhvi Gurram, Oytun Ulutan, and B.S. Manjunath. "Vision-Based Gesture Recognition in Human-Robot Teams Using Synthetic Data." In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 10278–10284. doi: [10.1109/iros45743.2020.9340728](https://doi.org/10.1109/iros45743.2020.9340728).
- [168] Ciro Mennella, Massimo Esposito, Giuseppe De Pietro, and Umberto Maniscalco. "Promoting Fairness in Activity Recognition Algorithms for Patient's Monitoring and Evaluation Systems in Healthcare." In: *Computers in Biology and Medicine* (2024), p. 108826. doi: [10.1016/j.combiomed.2024.108826](https://doi.org/10.1016/j.combiomed.2024.108826).
- [169] Ilya Mironov. "Rényi differential privacy." In: *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE. 2017, pp. 263–275. doi: [10.1109/CSF.2017.11](https://doi.org/10.1109/CSF.2017.11).
- [170] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The ethics of algorithms: Mapping the debate." In: *Big Data & Society* 3.2 (2016), p. 2053951716679679. doi: [10.1177/2053951716679679](https://doi.org/10.1177/2053951716679679).
- [171] Mathew Monfort et al. "Moments in Time Dataset: One Million Videos for Event Understanding." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2019), pp. 502–508. doi: [10.1109/tpami.2019.2901464](https://doi.org/10.1109/tpami.2019.2901464).

- [172] Ellis Monk. *Monk Skin Tone Scale*. 2019. URL: <https://skintone.google>.
- [173] Luís Moreira, Joana Figueiredo, Pedro Fonseca, João P Vilas-Boas, and Cristina P Santos. "Lower limb kinematic, kinetic, and EMG data from young healthy humans during walking at controlled speeds." In: *Scientific data* 8.1 (2021), p. 103. DOI: [10.1038/s41597-021-00881-3](https://doi.org/10.1038/s41597-021-00881-3).
- [174] Felix Morsbach, Jan Reubold, and Thorsten Strufe. "R+R: Understanding Hyperparameter Effects in DP-SGD." In: *Proceedings of the 2024 Annual Computer Security Applications Conference (ACSAC)*. IEEE, 2024, pp. 1217–1230. DOI: [10.1109/acsac63791.2024.00097](https://doi.org/10.1109/acsac63791.2024.00097).
- [175] David Mulero-Pérez, Manuel Benavent-Lledo, Jose Garcia-Rodriguez, and Markus Vincze. "Unrealgensyn: a framework for generating synthetic videos of Unfrequent human events." In: *Virtual Reality* 29.2 (Apr. 2025), p. 76. ISSN: 1434-9957. DOI: [10.1007/s10055-025-01146-9](https://doi.org/10.1007/s10055-025-01146-9).
- [176] Alhassan Mumuni, Fuseini Mumuni, and Nana Kobina Gerrar. "A Survey of Synthetic Data Augmentation Methods in Machine Vision." In: *Machine Intelligence Research* 21.5 (2024), pp. 831–869. DOI: [10.1007/s11633-022-1411-7](https://doi.org/10.1007/s11633-022-1411-7).
- [177] Alexander Quinn Nichol and Prafulla Dhariwal. "Improved Denoising Diffusion Probabilistic Models." In: *Proceedings of the 38th International Conference on Machine Learning*. 2021, pp. 8162–8171. URL: <https://openreview.net/forum?id=-NEXDKk8gZ>.
- [178] Sergey I. Nikolenko. *Synthetic Data for Deep Learning*. Springer Optimization and Its Applications. Springer, 2021. DOI: [10.1007/978-3-030-75178-4](https://doi.org/10.1007/978-3-030-75178-4).
- [179] OpenSim Documentation. *Checklist - Evaluating your Simulation*. <https://opensimconfluence.atlassian.net/wiki/spaces/OpenSim/pages/53114566/>. Accessed: 2024-11-20. 2012.
- [180] OpenSim Documentation. *Simulation with OpenSim - Best Practices*. <https://opensimconfluence.atlassian.net/wiki/spaces/OpenSim/pages/53087686/>. Accessed: 2024-11-20. 2012.
- [181] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." In: *BMJ* 372 (2021), n71. DOI: [10.1136/bmj.n71](https://doi.org/10.1136/bmj.n71).
- [182] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. "Scalable private learning with pate." In: *International Conference on Learning Representations (ICLR)*. 2018. URL: <https://openreview.net/forum?id=rkZB1XbRZ>.
- [183] Dennis Park and Deva Ramanan. "Articulated pose estimation with tiny synthetic videos." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2015, pp. 58–66. DOI: [10.1109/cvprw.2015.7301337](https://doi.org/10.1109/cvprw.2015.7301337).
- [184] European Parliament. *EU AI Act: first regulation on artificial intelligence*. 2023. URL: <https://www.europarl.europa.eu/news/en/headlines/society/20230601ST093804/eu-ai-act-first-regulation-on-artificial-intelligence>.

- [185] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. "Data and its (dis) contents: A survey of dataset development and use in machine learning research." In: *Patterns* 2.11 (2021), p. 100336. doi: [10.1016/j.patter.2021.100336](https://doi.org/10.1016/j.patter.2021.100336).
- [186] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. "Expressive Body Capture: 3D Hands, Face, and Body from a Single Image." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10975–10985. doi: [10.1109/cvpr.2019.01123](https://doi.org/10.1109/cvpr.2019.01123).
- [187] Kunyu Peng, David Schneider, Alina Roitberg, Kailun Yang, Jiaming Zhang, Chen Deng, Kaiyu Zhang, M Saquib Sarfraz, and Rainer Stiefelhagen. "Towards Video-based Activated Muscle Group Estimation in the Wild." In: *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 2024, pp. 4495–4504. doi: [10.1145/3664647.3680630](https://doi.org/10.1145/3664647.3680630).
- [188] Xingchao Peng, Ben Usman, Neela Kaushik, Dequan Wang, Judy Hoffman, and Kate Saenko. "Visda: A synthetic-to-real benchmark for visual domain adaptation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 2021–2026. doi: [10.1109/cvprw.2018.00271](https://doi.org/10.1109/cvprw.2018.00271).
- [189] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. "Articulated people detection and pose estimation: Reshaping the future." In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 3178–3185. doi: [10.1109/cvpr.2012.6248052](https://doi.org/10.1109/cvpr.2012.6248052).
- [190] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. "Learning privacy preserving encodings through adversarial training." In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 791–799. doi: [10.1109/wacv.2019.00089](https://doi.org/10.1109/wacv.2019.00089).
- [191] *Poly Haven HDRI*. <http://web.archive.org/web/20230223210457/https://polyhaven.com/hdri>. Accessed: 2023-03-01.
- [192] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Thakurta. "How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy." In: *Journal of Artificial Intelligence Research* 77 (2023), pp. 1113–1201. doi: [10.1613/jair.1.14649](https://doi.org/10.1613/jair.1.14649).
- [193] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. "Virtualhome: Simulating household activities via programs." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8494–8502. doi: [10.1109/cvpr.2018.00886](https://doi.org/10.1109/cvpr.2018.00886).
- [194] Prasanna Reddy Pulakurthi, Celso M. de Melo, Raghuveer Rao, and Majid Rabbani. "Enhancing human action recognition with GAN-based data augmentation." In: *Defense + Commercial Sensing*. Vol. 13035. SPIE. 2024, 130350O. doi: [10.1117/12.3021572](https://doi.org/10.1117/12.3021572).

- [195] Abhinanda R Punnakal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. "BABEL: bodies, action and behavior with English labels." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 722–731. doi: [10.1109/cvpr46437.2021.00078](https://doi.org/10.1109/cvpr46437.2021.00078).
- [196] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks." In: *International Conference on Learning Representations (ICLR)*. 2016. arXiv: [1511.06434](https://arxiv.org/abs/1511.06434).
- [197] Hossein Rahmani and Ajmal S. Mian. "3D Action Recognition from Novel Viewpoints." In: *Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1506–1515. doi: [10.1109/CVPR.2016.167](https://doi.org/10.1109/CVPR.2016.167).
- [198] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. "Learning a Deep Model for Human Action Recognition from Novel Viewpoints." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.3 (2018), pp. 667–681. doi: [10.1109/TPAMI.2017.2691768](https://doi.org/10.1109/TPAMI.2017.2691768).
- [199] Megani Rajendran, Chek Tien Tan, I. Atmosukarto, Aik Beng Ng, and Simon See. "Review on synergizing the Metaverse and AI-driven synthetic data: enhancing virtual realms and activity recognition in computer vision." In: *Visual Intelligence* 2.1 (2024), p. 32. doi: [10.1007/s44267-024-00059-6](https://doi.org/10.1007/s44267-024-00059-6).
- [200] Gundala Jhansi Rani, Mohammad Farukh Hashmi, and Aditya Gupta. "Surface electromyography and artificial intelligence for human activity recognition—A systematic review on methods, emerging trends applications, challenges, and future implementation." In: *IEEE access* 11 (2023), pp. 105140–105169. doi: [10.1109/access.2023.3316509](https://doi.org/10.1109/access.2023.3316509).
- [201] Arun V. Reddy, Ketul Shah, William Paul, Rohita Mocharla, Judy Hoffman, Kapil D. Katyal, Dinesh Manocha, Celso M. de Melo, and Rama Chellappa. "Synthetic-to-Real Domain Adaptation for Action Recognition: A Dataset and Baseline Performances." In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2023, pp. 4245–4251. doi: [10.1109/icra48891.2023.10160416](https://doi.org/10.1109/icra48891.2023.10160416).
- [202] General Data Protection Regulation. *General data protection regulation (GDPR)-official legal text* (2023). 2023. URL: <https://gdpr-info.eu/>.
- [203] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. "Learning to anonymize faces for privacy preserving action detection." In: *Proceedings of the european conference on computer vision (ECCV)*. 2018, pp. 620–636. doi: [10.1007/978-3-030-01246-5_38](https://doi.org/10.1007/978-3-030-01246-5_38).
- [204] Ziliang Ren, Qieshi Zhang, Xiangyang Gao, Pengyi Hao, and Jun Cheng. "Multi-modality learning for human action recognition." In: *Multimedia Tools and Applications* 80 (2021), pp. 16185–16203. doi: [10.1007/s11042-019-08576-z](https://doi.org/10.1007/s11042-019-08576-z).
- [205] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. "Playing for benchmarks." In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2213–2222. doi: [10.1109/iccv.2017.243](https://doi.org/10.1109/iccv.2017.243).
- [206] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. "Playing for data: Ground truth from computer games." In: *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II* 14. Springer. 2016, pp. 102–118. doi: [10.1007/978-3-319-46475-6_7](https://doi.org/10.1007/978-3-319-46475-6_7).

- [207] Alina Roitberg*, David Schneider*, Aulia Djamal, Constantin Seibold, Simon Reiß, and Rainer Stiefelhagen. “Let’s Play for Action: Recognizing Activities of Daily Living by Learning from Life Simulation Video Games.” In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021. doi: [10.1109/IROS51168.2021.9636381](https://doi.org/10.1109/IROS51168.2021.9636381).
- [208] Mónica Rojas-Martínez, Leidy Yanet Serna, Mislav Jordanic, Hamid Reza Marateb, Roberto Merletti, and Miguel Ángel Mañanas. “High-density surface electromyography signals during isometric contractions of elbow muscles of healthy humans.” In: *Scientific data* 7.1 (2020), p. 397. doi: [10.1038/s41597-020-00717-6](https://doi.org/10.1038/s41597-020-00717-6).
- [209] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695. doi: [10.1109/cvpr52688.2022.01042](https://doi.org/10.1109/cvpr52688.2022.01042).
- [210] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. “FitNets: Hints for Thin Deep Nets.” In: *International Conference on Learning Representations (ICLR)*. 2015. arXiv: [1412.6550](https://arxiv.org/abs/1412.6550).
- [211] Javier Romero, Dimitrios Tzionas, and Michael J. Black. “Embodied hands: modeling and capturing hands and bodies together.” In: *ACM Transactions on Graphics (TOG)* 36.6 (Nov. 2017). issn: 0730-0301. doi: [10.1145/3130800.3130883](https://doi.org/10.1145/3130800.3130883).
- [212] Michael S Ryoo, AJ Piergiovanni, Juhana Kangaspunta, and Anelia Angelova. “Assemblenet++: Assembling modality representations via attention connections.” In: *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*. Springer. 2020, pp. 654–671. doi: [10.1007/978-3-030-58565-5_39](https://doi.org/10.1007/978-3-030-58565-5_39).
- [213] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. “Temporal Generative Adversarial Nets with Singular Value Clipping.” In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2849–2858. doi: [10.1109/iccv.2017.308](https://doi.org/10.1109/iccv.2017.308).
- [214] Sina Sajadmanesh and Daniel Gatica-Perez. “Locally Private Graph Neural Networks.” In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. CCS ’21*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 2130–2145. isbn: 9781450384544. doi: [10.1145/3460120.3484565](https://doi.org/10.1145/3460120.3484565).
- [215] Sina Sajadmanesh and Daniel Gatica-Perez. “ProGAP: Progressive Graph Neural Networks with Differential Privacy Guarantees.” In: *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*. 2024, pp. 596–605. doi: [10.1145/3616855.3635761](https://doi.org/10.1145/3616855.3635761).
- [216] Sina Sajadmanesh, Ali Shahin Shamsabadi, Aurélien Bellet, and Daniel Gatica-Perez. “GAP: Differentially Private Graph Neural Networks with Aggregation Perturbation.” In: *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 3223–3240. isbn: 978-1-939133-37-3. arXiv: [2203.00949](https://arxiv.org/abs/2203.00949). url: <https://www.usenix.org/conference/usenixsecurity23/presentation/sajadmanesh>.

- [217] M. Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelbogen. "Efficient Parameter-free Clustering Using First Neighbor Relations." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 8934–8943. doi: [10.1109/cvpr.2019.00914](https://doi.org/10.1109/cvpr.2019.00914).
- [218] Saquib Sarfraz, Marios Koulakis, Constantin Seibold, and Rainer Stiefelbogen. "Hierarchical nearest neighbor graph embedding for efficient dimensionality reduction." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 336–345. doi: [10.1109/cvpr52688.2022.00043](https://doi.org/10.1109/cvpr52688.2022.00043).
- [219] David Schneider, Marco Keller, Zeyun Zhong, Kunyu Peng, Alina Roitberg, Jürgen Beyerer, and Rainer Stiefelbogen. "SynthAct: Towards Generalizable Human Action Recognition based on Synthetic Data." In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 13038–13045. doi: [10.1109/icra57147.2024.10611486](https://doi.org/10.1109/icra57147.2024.10611486).
- [220] David Schneider, Zdravko Marinov, Moritz Mistol, Zeyun Zhong, Alexander Jaus, Rodi Düger, Rafael Baur, M. Saquib Sarfraz, and Rainer Stiefelbogen. "OmniFall: From Staged Through Synthetic to Wild, A Unified Multi-Domain Dataset for Robust Fall Detection." In: (2026). Accepted at the European Conference on Computer Vision (ECCV), 2026. arXiv: [2505.19889](https://arxiv.org/abs/2505.19889).
- [221] David Schneider, Simon Reiß, Marco Kugler, Alexander Jaus, Kunyu Peng, Susanne Sutschet, M. Saquib Sarfraz, Sven Matthiesen, and Rainer Stiefelbogen. "Muscles in Time: Learning to Understand Human Motion In-Depth by Simulating Muscle Activations." In: *Advances in Neural Information Processing Systems*. Vol. 37. 2024, pp. 67251–67281. URL: https://papers.nips.cc/paper_files/paper/2024/hash/7c3465ba08732cc2db38f070bfae601a-Abstract-Datasets_and_Benchmarks_Track.html.
- [222] David Schneider, Sina Sajadmanesh, Vikash Sehwal, Saquib Sarfraz, Rainer Stiefelbogen, Lingjuan Lyu, and Vivek Sharma. "Activity Recognition on Avatar-Anonymized Datasets with Masked Differential Privacy." In: Presented at the 2nd International Workshop on Privacy-Preserving Computer Vision, ECCV. 2024. arXiv: [2410.17098](https://arxiv.org/abs/2410.17098).
- [223] David Schneider, Saquib Sarfraz, Alina Roitberg, and Rainer Stiefelbogen. "Pose-Based Contrastive Learning for Domain Agnostic Activity Representations." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2022, pp. 3433–3443. doi: [10.1109/cvprw56347.2022.00387](https://doi.org/10.1109/cvprw56347.2022.00387).
- [224] Céline Schreiber and Florent Moissenet. "A multimodal dataset of human gait at different walking speeds established on injury-free adult participants." In: *Scientific data* 6.1 (2019), p. 111. doi: [10.1038/s41597-019-0124-4](https://doi.org/10.1038/s41597-019-0124-4).
- [225] Christian Schudt, Ivan Laptev, and Barbara Caputo. "Recognizing human actions: a local SVM approach." In: *IEEE international conference on pattern recognition* (2004), pp. 32–36. doi: [10.1109/icpr.2004.1334462](https://doi.org/10.1109/icpr.2004.1334462).

- [226] Raphael Schumann and Ines Rehbein. “Active learning via membership query synthesis for semi-supervised sentence classification.” In: *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*. 2019, pp. 472–481. DOI: [10.18653/v1/k19-1044](https://doi.org/10.18653/v1/k19-1044).
- [227] Ajay Seth et al. “OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement.” In: *PLoS computational biology* 14.7 (2018), e1006223. DOI: [10.1371/journal.pcbi.1006223](https://doi.org/10.1371/journal.pcbi.1006223).
- [228] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. “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)*. 2016, pp. 1010–1019. DOI: [10.1109/cvpr.2016.115](https://doi.org/10.1109/cvpr.2016.115).
- [229] Ali Shahin Shamsabadi and Nicolas Papernot. “Losing less: A loss for differentially private deep learning.” In: *Proceedings on Privacy Enhancing Technologies* 2023.3 (2023), pp. 307–320. DOI: [10.56553/popets-2023-0083](https://doi.org/10.56553/popets-2023-0083).
- [230] Ali Shahin Shamsabadi, Brij Mohan Lal Srivastava, Aurélien Bellet, Nathalie Vauquier, Emmanuel Vincent, Mohamed Maouche, Marc Tommasi, and Nicolas Papernot. “Differentially private speaker anonymization.” In: *Proceedings on Privacy Enhancing Technologies* 2023.1 (2023), pp. 98–114. DOI: [10.56553/popets-2023-0007](https://doi.org/10.56553/popets-2023-0007).
- [231] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. “Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition.” In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 12026–12035. DOI: [10.1109/CVPR.2019.01230](https://doi.org/10.1109/CVPR.2019.01230).
- [232] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. “Skeleton-based action recognition with multi-stream adaptive graph convolutional networks.” In: *IEEE Transactions on Image Processing* 29 (2020), pp. 9532–9545. DOI: [10.1109/TIP.2020.3028207](https://doi.org/10.1109/TIP.2020.3028207).
- [233] Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. “Selective differential privacy for language modeling.” In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 2022, pp. 2848–2859. DOI: [10.18653/v1/2022.naacl-main.205](https://doi.org/10.18653/v1/2022.naacl-main.205).
- [234] Jungpil Shin, Najmul Hassan, Abu Saleh Musa Miah, and Satoshi Nishimura. “A Comprehensive Methodological Survey of Human Activity Recognition Across Diverse Data Modalities.” In: *Sensors* 25.13 (2025), p. 4028. DOI: [10.3390/s25134028](https://doi.org/10.3390/s25134028).
- [235] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. “AI Models Collapse When Trained on Recursively Generated Data.” In: *Nature* 631.8022 (2024), pp. 755–759. DOI: [10.1038/s41586-024-07566-y](https://doi.org/10.1038/s41586-024-07566-y).
- [236] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. “Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding.” In: *European Conference on Computer Vision*. Springer. 2016, pp. 510–526. DOI: [10.1007/978-3-319-46448-0_31](https://doi.org/10.1007/978-3-319-46448-0_31).

- [237] Karen Simonyan and Andrew Zisserman. “Two-Stream Convolutional Networks for Action Recognition in Videos.” In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014. URL: https://papers.nips.cc/paper_files/paper/2014/hash/ca007296a63f7d1721a2399d56363022-Abstract.html.
- [238] Uriel Singer et al. “Make-A-Video: Text-to-Video Generation without Text-Video Data.” In: *International Conference on Learning Representations (ICLR)*. 2023. URL: <https://openreview.net/forum?id=nJfylDvgzlq>.
- [239] Kaushik Bhargav Sivangi, Paul Henderson, and Fani Deligianni. “Differentially Private 2D Human Pose Estimation.” In: (2026), pp. 21143–21153. URL: https://openaccess.thecvf.com/content/CVPR2026/html/Sivangi_Differentially_Private_2D_Human_Pose_Estimation_CVPR_2026_paper.html.
- [240] Xingyu Song, Zhan Li, Shi Chen, and Kazuyuki Demachi. “GTAutoAct: An Automatic Datasets Generation Framework Based on Game Engine Redevelopment for Action Recognition.” In: *arXiv preprint* (2024). arXiv: [2401.13414](https://arxiv.org/abs/2401.13414).
- [241] Tetsuri Sonoda and Anders Grunnet-Jepsen. “Depth Image Compression by Colorization for Intel® RealSense™ Depth Cameras.” In: Rev 1.0 (2020). URL: <https://dev.realsenseai.com/download/9979/>.
- [242] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. “UCF101: A dataset of 101 human actions classes from videos in the wild.” In: *arXiv preprint* (2012). arXiv: [1212.0402](https://arxiv.org/abs/1212.0402).
- [243] Vinkle Srivastav, Afshin Gangi, and Nicolas Padoy. “Human pose estimation on privacy-preserving low-resolution depth images.” In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2019, pp. 583–591. DOI: [10.1007/978-3-030-32254-0_65](https://doi.org/10.1007/978-3-030-32254-0_65).
- [244] Baochen Sun and Kate Saenko. “From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains.” In: *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. DOI: [10.5244/C.28.82](https://doi.org/10.5244/C.28.82).
- [245] Baochen Sun and Kate Saenko. “Deep CORAL: Correlation Alignment for Deep Domain Adaptation.” In: *European Conference on Computer Vision Workshops (ECCVW)*. 2016, pp. 443–450. DOI: [10.1007/978-3-319-49409-8_35](https://doi.org/10.1007/978-3-319-49409-8_35). arXiv: [1607.01719](https://arxiv.org/abs/1607.01719).
- [246] Min Sun, Pushmeet Kohli, and Jamie Shotton. “Conditional regression forests for human pose estimation.” In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 3394–3401. DOI: [10.1109/cvpr.2012.6248079](https://doi.org/10.1109/cvpr.2012.6248079).
- [247] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. “Resolution-Robust Large Mask Inpainting with Fourier Convolutions.” In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 2149–2159. DOI: [10.1109/wacv51458.2022.00323](https://doi.org/10.1109/wacv51458.2022.00323).

- [248] Tatsuya Teramae, Takamitsu Matsubara, Tomoyuki Noda, and Jun Morimoto. "Neural-enhanced motion-to-EMG: refining simulated muscle activity from musculoskeletal models using a Seq2Seq approach." In: *Frontiers in Bioengineering and Biotechnology* Volume 13 - 2025 (2025). ISSN: 2296-4185. DOI: [10.3389/fbioe.2025.1611414](https://doi.org/10.3389/fbioe.2025.1611414). URL: <https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2025.1611414>.
- [249] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. "Motionclip: Exposing human motion generation to clip space." In: *European Conference on Computer Vision*. Springer. 2022, pp. 358–374. DOI: [10.1007/978-3-031-20047-2_21](https://doi.org/10.1007/978-3-031-20047-2_21).
- [250] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N. Metaxas, and Sergey Tulyakov. "A Good Image Generator Is What You Need for High-Resolution Video Synthesis." In: *International Conference on Learning Representations*. 2021. URL: <https://openreview.net/forum?id=6puCSjH3hwa>.
- [251] Mary E Tinetti, Wen-Liang Liu, and Elizabeth B Claus. "Predictors and prognosis of inability to get up after falls among elderly persons." In: *JAMA* 269.1 (1993), pp. 65–70. DOI: [10.1001/jama.1993.03500010075035](https://doi.org/10.1001/jama.1993.03500010075035).
- [252] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. "Domain randomization for transferring deep neural networks from simulation to the real world." In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE. 2017, pp. 23–30. DOI: [10.1109/iro.2017.8202133](https://doi.org/10.1109/iro.2017.8202133).
- [253] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training." In: 35 (2022), pp. 10078–10093. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/416f9cb3276121c42eebb86352a4354a-Abstract-Conference.html.
- [254] Florian Tramer and Dan Boneh. "Differentially private learning needs better features (or much more data)." In: *International Conference on Learning Representations (ICLR)*. 2021. URL: <https://openreview.net/forum?id=YTWGvpF0QD->.
- [255] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning Spatiotemporal Features with 3D Convolutional Networks." In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 4489–4497. DOI: [10.1109/ICCV.2015.510](https://doi.org/10.1109/ICCV.2015.510).
- [256] Thanh-Hai Tran, Thi-Lan Le, Dinh-Tan Pham, Van-Nam Hoang, Van-Minh Khong, Quoc-Toan Tran, Thai-Son Nguyen, and Cuong Pham. "A multi-modal multi-view dataset for human fall analysis and preliminary investigation on modality." In: *2018 24th International Conference on Pattern Recognition (ICPR)*. 2018 24th International Conference on Pattern Recognition (ICPR). ISSN: 1051-4651. Aug. 2018, pp. 1947–1952. DOI: [10.1109/ICPR.2018.8546308](https://doi.org/10.1109/ICPR.2018.8546308). (Visited on 12/06/2024).
- [257] Nikolaus F Troje. *Decomposing biological motion: A framework for analysis and synthesis of human gait patterns*. 2002. DOI: [10.1167/2.5.2](https://doi.org/10.1167/2.5.2).

- [258] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features." In: *arXiv preprint* (2025). arXiv: [2502.14786](#).
- [259] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. "MoCoGAN: Decomposing Motion and Content for Video Generation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 1526–1535. DOI: [10.1109/CVPR.2018.00165](#).
- [260] Scott D Uhlich, Antoine Falisse, Łukasz Kidziński, Julie Muccini, Michael Ko, Akshay S Chaudhari, Jennifer L Hicks, and Scott L Delp. "OpenCap: Human movement dynamics from smartphone videos." In: *PLoS computational biology* 19.10 (2023), e1011462. DOI: [10.1371/journal.pcbi.1011462](#).
- [261] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. "Instance normalization: The missing ingredient for fast stylization." In: *arXiv preprint* (2016). arXiv: [1607.08022](#).
- [262] United States Census Bureau. *Updates to OMB's Race/Ethnicity Standards*. <https://web.archive.org/web/20251006232303/https://www.census.gov/newsroom/blogs/random-samplings/2024/04/updates-race-ethnicity-standards.html>. Accessed: 2025-11-19. Apr. 2024.
- [263] Unity Technologies. *Unity Perception Package*. <https://github.com/Unity-Technologies/com.unity.perception>. 2020.
- [264] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. "FVD: A new Metric for Video Generation." In: *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop (DeepGenStruct)* (2019). URL: <https://openreview.net/forum?id=rylgEULtdN>.
- [265] Gul Varol, Ivan Laptev, and Cordelia Schmid. "Long-term temporal convolutions for action recognition." In: *IEEE transactions on pattern analysis and machine intelligence* (2017), pp. 1510–1517. DOI: [10.1109/tpami.2017.2712608](#).
- [266] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. "Synthetic humans for action recognition from unseen viewpoints." In: *International Journal of Computer Vision* 129.7 (2021), pp. 2264–2287. DOI: [10.1007/s11263-021-01467-7](#).
- [267] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. "Learning from synthetic humans." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 109–117. DOI: [10.1109/cvpr.2017.492](#).
- [268] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All you Need." In: 30 (2017). URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

- [269] Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbahn, and Anson Ho. "Position: Will we run out of data? Limits of LLM scaling based on human-generated data." In: *International Conference on Machine Learning (ICML)*. 2024, pp. 49523–49544. URL: <https://icml.cc/virtual/2024/poster/33903>.
- [270] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. "Generating Videos with Scene Dynamics." In: *Advances in Neural Information Processing Systems*. 2016. arXiv: [1609.02612](https://arxiv.org/abs/1609.02612).
- [271] Team Wan et al. "Wan: Open and Advanced Large-Scale Video Generative Models." In: *arXiv preprint* (2025). arXiv: [2503.20314](https://arxiv.org/abs/2503.20314).
- [272] Zhijing Wan, Zhixiang Wang, Cheukting Chung, and Zheng Wang. "A Survey of Dataset Refinement for Problems in Computer Vision Datasets." In: *ACM Computing Surveys* 56.7 (Apr. 2024). ISSN: 0360-0300. DOI: [10.1145/3627157](https://doi.org/10.1145/3627157).
- [273] Boxin Wang, Fan Wu, Yunhui Long, Luka Rimanic, Ce Zhang, and Bo Li. "Datalens: Scalable privacy preserving training via gradient compression and aggregation." In: *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 2021, pp. 2146–2168. DOI: [10.1145/3460120.3484579](https://doi.org/10.1145/3460120.3484579).
- [274] Heng Wang and Cordelia Schmid. "Action recognition with improved trajectories." In: *International Conference on Computer Vision (ICCV)*. 2013, pp. 3551–3558. DOI: [10.1109/iccv.2013.441](https://doi.org/10.1109/iccv.2013.441).
- [275] Huogen Wang, Zhanjie Song, Wanqing Li, and Pichao Wang. "A hybrid network for large-scale action recognition from rgb and depth modalities." In: *Sensors* 20.11 (2020), p. 3305. DOI: [10.3390/s20113305](https://doi.org/10.3390/s20113305).
- [276] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. "Generalizing to Unseen Domains: A Survey on Domain Generalization." In: *IEEE Transactions on Knowledge and Data Engineering* 35.8 (2023), pp. 8052–8072. DOI: [10.1109/TKDE.2022.3178128](https://doi.org/10.1109/TKDE.2022.3178128).
- [277] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. "Temporal Segment Networks: Towards Good Practices for Deep Action Recognition." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2016, pp. 20–36. DOI: [10.1007/978-3-319-46484-8_2](https://doi.org/10.1007/978-3-319-46484-8_2).
- [278] Yi Wang et al. "InternVideo: General Video Foundation Models via Generative and Discriminative Learning." In: *arXiv preprint* (2022). arXiv: [2212.03191](https://arxiv.org/abs/2212.03191).
- [279] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. "Subsampled Renyi Differential Privacy and Analytical Moments Accountant." In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1226–1235. URL: <https://proceedings.mlr.press/v89/wang19b.html>.
- [280] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. "Loong: Generating Minute-level Long Videos with Autoregressive Language Models." In: *arXiv preprint* (2024). arXiv: [2410.02757](https://arxiv.org/abs/2410.02757).

- [281] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. "Masked feature prediction for self-supervised visual pre-training." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 14668–14678. DOI: [10.1109/cvpr52688.2022.01426](https://doi.org/10.1109/cvpr52688.2022.01426).
- [282] World Health Organization. *Falls*. <https://web.archive.org/web/20250423060144/https://www.who.int/news-room/fact-sheets/detail/falls/>. 2021.
- [283] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. "Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.4 (2020), pp. 2126–2139. DOI: [10.1109/tpami.2020.3026709](https://doi.org/10.1109/tpami.2020.3026709).
- [284] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. "Towards privacy-preserving visual recognition via adversarial training: A pilot study." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 606–624. DOI: [10.1007/978-3-030-01270-0_37](https://doi.org/10.1007/978-3-030-01270-0_37).
- [285] Markus Wulfmeier, Alex Bewley, and Ingmar Posner. "Addressing appearance change in outdoor robotics with adversarial domain adaptation." In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 1551–1558. DOI: [10.1109/iros.2017.8205961](https://doi.org/10.1109/iros.2017.8205961).
- [286] Sierra Calanda Wyllie, Ilia Shumailov, and Nicolas Papernot. "Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias." In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 2024, pp. 2113–2147. DOI: [10.1145/3630106.3659029](https://doi.org/10.1145/3630106.3659029).
- [287] Taihong Xiao, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker, and Ming-Hsuan Yang. "Adversarial learning of privacy-preserving and task-oriented representations." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 12434–12441. DOI: [10.1609/aaai.v34i07.6930](https://doi.org/10.1609/aaai.v34i07.6930).
- [288] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. "Re-thinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 305–321. DOI: [10.1007/978-3-030-01267-0_19](https://doi.org/10.1007/978-3-030-01267-0_19).
- [289] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. "Pose Flow: Efficient Online Pose Tracking." In: *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 53. arXiv: [1802.00977](https://arxiv.org/abs/1802.00977). URL: <http://bmvc2018.org/contents/papers/0096.pdf>.
- [290] Sijie Yan, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018. ISBN: 978-1-57735-800-8. DOI: [10.1609/aaai.v32i1.12328](https://doi.org/10.1609/aaai.v32i1.12328).

- [291] Di Yang, Rui Dai, Yaohui Wang, Rupayan Mallick, Luca Minciullo, Gianpiero Francesca, and Francois Bremond. "Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos." In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 2363–2372. doi: [10.1109/wacv48630.2021.00241](https://doi.org/10.1109/wacv48630.2021.00241).
- [292] Zhuoyi Yang et al. "CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer." In: *International Conference on Learning Representations*. Vol. 2025. 2025, pp. 83048–83077. URL: https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce31378e9f41d8907e97dab172b6c559-Abstract-Conference.html.
- [293] Zhiyu Yao, Yunbo Wang, Jianmin Wang, Philip S. Yu, and Mingsheng Long. "VideoDG: Generalizing Temporal Relations in Videos to Novel Domains." In: vol. 44. 11. 2022, pp. 7989–8004. doi: [10.1109/TPAMI.2021.3116945](https://doi.org/10.1109/TPAMI.2021.3116945).
- [294] Ashkan Yousefpour et al. "Opacus: User-Friendly Differential Privacy Library in PyTorch." In: *NeurIPS 2021 Workshop Privacy in Machine Learning*. 2021. arXiv: [2109.12298](https://arxiv.org/abs/2109.12298). URL: <https://openreview.net/forum?id=EopKEYBoI->.
- [295] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. "Differentially private fine-tuning of language models." In: *International Conference on Learning Representations (ICLR)*. 2022. URL: <https://openreview.net/forum?id=Q42f0dfjECO>.
- [296] Sergey Zagoruyko and Nikos Komodakis. "Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer." In: *International Conference on Learning Representations*. 2017. URL: <https://openreview.net/forum?id=Sks9-ajex>.
- [297] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization." In: *International Conference on Learning Representations (ICLR)*. 2018. URL: <https://openreview.net/forum?id=r1Ddp1-Rb>.
- [298] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. "T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 14730–14740. doi: [10.1109/cvpr52729.2023.01415](https://doi.org/10.1109/cvpr52729.2023.01415).
- [299] Jie Zhang, Yihui Zhao, Fergus Shone, Zhenhong Li, Alejandro F. Frangi, Shengquan Xie, and Zhiqiang Zhang. "Physics-Informed Deep Learning for Musculoskeletal Modelling: Predicting Muscle Forces and Joint Kinematics from Surface EMG." In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (2022), pp. 484–493. doi: [10.1109/tnsre.2022.3226860](https://doi.org/10.1109/tnsre.2022.3226860).
- [300] Qiang Zhang. "Experimental data of sEMG, US imaging, GRF, and markers for walking on treadmill across multiple speeds." In: (2021). URL: <https://iee-dataport.org/documents/experimental-data-semg-us-imaging-grf-and-markers-walking-treadmill-across-multiple>.
- [301] Zhixiang Zhang, Thomas Cilloni, Charles Walter, and Charles Fleming. "Multi-scale, class-generic, privacy-preserving video." In: *Electronics* 10.10 (2021), p. 1172. doi: [10.3390/electronics10101172](https://doi.org/10.3390/electronics10101172).

- [302] Zhong Zhang, Christopher Conly, and Vassilis Athitsos. "Evaluating Depth-Based Computer Vision Methods for Fall Detection under Occlusions." In: *Advances in Visual Computing*. Vol. 8888. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 196–207. ISBN: 978-3-319-14363-7 978-3-319-14364-4. DOI: [10.1007/978-3-319-14364-4_19](https://doi.org/10.1007/978-3-319-14364-4_19). (Visited on 12/06/2024).
- [303] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. "Random erasing data augmentation." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07. 2020, pp. 13001–13008. DOI: [10.1609/aaai.v34i07.7000](https://doi.org/10.1609/aaai.v34i07.7000).
- [304] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. "Domain generalization: A survey." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2023), pp. 4396–4415. DOI: [10.1109/TPAMI.2022.3195549](https://doi.org/10.1109/TPAMI.2022.3195549).
- [305] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. "Learning to generate novel domains for domain generalization." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020, pp. 561–578. DOI: [10.1007/978-3-030-58517-4_33](https://doi.org/10.1007/978-3-030-58517-4_33).
- [306] Shangchen Zhou, Chongyi Li, Kelvin C. K. Chan, and Chen Change Loy. "ProPainter: Improving Propagation and Transformer for Video Inpainting." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 10443–10452. DOI: [10.1109/ICCV51070.2023.00961](https://doi.org/10.1109/ICCV51070.2023.00961).
- [307] Yuqing Zhu and Yu-Xiang Wang. "Poission Subsampled Rényi Differential Privacy." In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 7634–7642. URL: <https://proceedings.mlr.press/v97/zhu19c.html>.
- [308] Yuqing Zhu, Xiang Yu, Manmohan Chandraker, and Yu-Xiang Wang. "Private-knn: Practical differential privacy for computer vision." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 11854–11862. DOI: [10.1109/cvpr42600.2020.01187](https://doi.org/10.1109/cvpr42600.2020.01187).
- [309] Chenhui Zuo, Kaibo He, Jing Shao, and Yanan Sui. "Self model for embodied intelligence: Modeling full-body human musculoskeletal system and locomotion control with hierarchical low-dimensional representation." In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 13062–13069. DOI: [10.1109/icra57147.2024.10610081](https://doi.org/10.1109/icra57147.2024.10610081).