Karlsruher Institut für Technologie

Towards Generalizable Deep Learning-Based Human Action 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

Kunyu Peng

aus ChiFeng, CHINA

October, 2024

Tag der mündlichen Prüfung: 23.10.2024

Hauptreferent: Prof. Dr. -Ing. Rainer Stiefelhagen
Korreferent: Prof. Dr. -Ing. Jürgen Gall

Dedication

This thesis is dedicated to my loving parents, Li Peng and Xuemei Liang, who have always believed in me and provided unwavering support throughout my academic journey. Their constant encouragement and sacrifices have been my driving force.

To my husband, Jialin Zheng, whose patience, love, and understanding have been a source of strength and motivation during the most challenging times of my PhD journey. His belief in my abilities and his emotional support have made this achievement possible.

Acknowledgements

I am deeply grateful to the many individuals who supported me throughout my doctoral program at the Karlsruhe Institute of Technology, culminating in this dissertation. I would like to express my profound gratitude to my dissertation committee for their invaluable guidance, support, and expertise throughout my research. I am especially thankful to my advisors, Prof. Dr.-Ing. Rainer Stiefelhagen and Jun. Prof. Dr.-Ing. Alina Roitberg, for being inspiring, patient, and encouraging mentors. I also extend my thanks to my former colleague, Prof. Dr. Kailun Yang, and my current colleague, Dr.-Ing. Saquib Sarfraz, for sharing their rich experience in scientific research and publications. I am appreciative of my current colleagues, Corinna Haas-Hecker, David Schneider, Jiaming Zhang, Di Wen, Ruiping Liu, Junwei Zheng, Yufan Chen, Alexander Jaus, Zdravko Marinov, Omar Moured, and Simon Reiss, for their collaborations on papers and work support. The journey to achieve a PhD has been both exciting and challenging, leaving me with unforgettable memories. As Friedrich Nietzsche said, "That which does not kill us makes us stronger." Throughout this short three-and-a-half-year PhD period, I experienced numerous ups and downs. The hardships have ultimately given me a calm mind towards almost anything, which I believe will benefit me throughout my life. I must thank my parents, Li Peng and Xuemei Liang, and my husband, Jialin Zheng, for their encouragement during my darkest times. Additionally, ChatGPT is acknowledged for the grammer checking within this thesis.

Lastly, I sincerely acknowledge all the challenges that appeared during this PhD journey, which have all finally turned into unique shining starlight in my life.

Abstract

This thesis addresses the critical challenge of generalizability in deep learning based human action recognition models, focusing on both skeleton-based and video-based approaches. Generalizable human action recognition is essential for numerous applications, including surveillance, healthcare, sports analytics, and human-computer interaction. Ensuring these systems can accurately recognize actions across diverse and varying environments enhances their reliability and effectiveness, making them valuable tools in real-world scenarios. Through the proposal of a series of new methods and benchmarks, this thesis explored the generalizable challenges in human action recognition field in various perspectives. For skeleton-based action recognition, novel technique, Trans4SOAR, leveraging transformer architectures for multi-modal feature fusion and incorporating prototypical learning, and OPSTL, a two-stage imputation method, is designed separately for one-shot recognition and self-supervised learning under occlusions. The CrossMax approach is proposed to tackle open-set recognition, improving the model's ability to identify unseen actions for skeleton-based human action recognition using cross-modal logits calibration. For video-based action recognition, the RelaMiX method demonstrates data effective few-shot domain adaptation by incorporating temporal relational attentional dropout and cross-domain information alignment. Additionally, a cross-modal fall detection method is developed to achieve effective RGB to depth unsupervised domain adaptation, enhancing safety applications in real-world scenarios. Comprehensive benchmarks are constructed to evaluate these methods, highlighting their superior performance and potential for practical applications compared with numerous existing techniques in the related field. This thesis lays a strong foundation for future advancements in exploring generalizable human action recognition.

Zusammenfassung

Diese Dissertation befasst sich mit der kritischen Herausforderung der Generalisierbarkeit in Deep-Learning-basierten Modellen zur Erkennung menschlicher Aktionen, wobei sowohl skelettbasierte als auch videobasierte Ansätze im Fokus stehen. Generalisierbare Erkennung menschlicher Aktionen ist für zahlreiche Anwendungen, einschließlich Überwachung, Gesundheitswesen, Sportanalyse und Mensch-Computer-Interaktion, von entscheidender Bedeutung. Die Gewährleistung, dass diese Systeme Aktionen in unterschiedlichen und variierenden Umgebungen genau erkennen können, erhöht ihre Zuverlässigkeit und Effektivität und macht sie zu wertvollen Werkzeugen in realen Szenarien. Durch die Vorschläge einer Reihe neuer Methoden und Benchmarks untersucht diese Dissertation die Herausforderungen der Generalisierbarkeit im Bereich der Erkennung menschlicher Aktionen aus verschiedenen Perspektiven. Für die skelettbasierte Aktionserkennung wurde die neuartige Technik Trans4SOAR entwickelt, die Transformer-Architekturen zur multimodalen Merkmalsfusion nutzt und prototypisches Lernen integriert. Darüber hinaus wurde OPSTL, eine zweistufige Imputationsmethode, speziell für One-Shot-Erkennung und selbstüberwachtes Lernen unter Okklusionen entworfen. Der CrossMax-Ansatz wird vorgeschlagen, um das Open-Set-Erkennen zu bewältigen und die Fähigkeit des Modells zu verbessern, unbekannte Aktionen für die skelettbasierte menschliche Aktionserkennung durch cross-modale Logits-Kalibrierung zu identifizieren. Für die videobasierte Aktionserkennung demonstriert die RelaMiX-Methode eine datenwirksame Few-Shot-Domänenanpassung durch die Einbeziehung von temporalen relationalen Aufmerksamkeits-Dropouts und domänenübergreifender Informationsausrichtung. Zusätzlich wurde eine crossmodale Sturzerkennungsmethode entwickelt, um eine effektive, unbeaufsichtigte Domänenanpassung von RGB zu Tiefenbildern zu erreichen, was Sicherheitsanwendungen in realen Szenarien verbessert. Umfassende Benchmarks wurden erstellt, um diese Methoden zu evaluieren und ihre überlegene Leistung und ihr Potenzial für praktische Anwendungen im Vergleich zu zahlreichen bestehenden Techniken in diesem Bereich hervorzuheben. Diese Dissertation legt eine starke Grundlage für zukünftige Fortschritte in der Erforschung der generalisierbaren Erkennung menschlicher Aktionen.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Human action recognition is pivotal in numerous applications, including surveillance, healthcare, and sports analytics [174]. In the real world scenario, deep learning-based human action recognition can be used in security systems to automatically detect and alert authorities about strange behaviors, enhancing public safety in real-time scenario [91]. Additionally, it can be applied in healthcare to assist elderly patients, identifying and responding to falls or other critical events [140]. With the rapid development of deep learning in computer vision field, more and more powerful deep learning architectures are developed, which can achieve promising accuracy for the conventional action recognition task [19, 123, 153, 209]. However, these models mostly show limited performance when facing generalizable challenges. Generalizability is a crucial attribute for ensuring adaptable performance across diverse and varying distribution shifts of the data. The lack of generalizability can result in large performance decay of recognition. Additionally, it can have a negative societal impact through providing confident false predictions. The challenges of generalizability has been tackled in computer vision field by defining specific generalizable tasks, *e.g.*, one-shot recognition [13], and open-set recognition [127], where different tasks focus on different distribution changes of the data. Some of those are overlooked in the human action recognition field.

In this thesis, we would like to explore those uncharted challenges in the human action recognition field to find out the limitations of the existing works and propose corresponding solutions. Different action recognition approaches face distinct challenges in achieving generalizability. Existing researches in human action recognition can be categorized into two primary clusters: skeleton-based approaches and video-based approaches.

Skeleton-based approaches use Graph Convolutional neural networks (GCNs) to extract features from 3D body key points [102, 121, 224]. These methods are highly efficient because the data is sparse and GCNs are lightweight, potentially achieving a tenfold reduction in parameters compared to other methods. However, skeleton data is easily disrupted by real-world disturbances, such as occlusions, leading to geometric and temporal discontinuities that can impair recognition per-

formance and discriminative embedding learning. These disruptions are particularly problematic in challenging tasks requiring high generalizability, such as one-shot learning scenario [134, 135] and self-learning scenario [232]. In order to implement those approaches in the real-world, the deep learning skeleton-based human action recognition model should be robust and generalizable to some specific perturbations, where occlusion is the most common disturbance from the environment. How to tackle occlusions on such generalizable tasks remains unexplored in the community.

Apart from the disturbance from the environment, deep learning models also face with generalizable challenge brought by the confidence of the model. Due to the sparsity of the skeleton data, most existing GCN approaches suffer from overfitting issues when using 3D human body poses as input, making the rejection of unseen categories during inference difficult to address [55]. Open-set recognition formalizes this generalizable challenges in the perspective of the calibration of the model confidence by requiring the model to give high confidence scores for categories seen during training and low confidence scores for categories outside the training phase [127]. This capability to reject unseen categories is critical in real-life scenarios to prevent offensive false recognition that could lead to incorrect decisions, which is particularly useful for robots to provide appropriate human assistance according to the action recognition results. However, there is also no existing work focus on the open-set recognition challenge in terms of the skeleton-based human action recognition task.

Video-based approaches utilize 3D convolutions and transformers to extract features from RGB temporal data [19, 123]. These methods offer different advantages, such as being less sensitive to occlusions and providing richer contextual information from the environment, which make video-based human action recognition models face with less overfitting issues. The most essential generalizable challenge for video-based human action recognition approaches lies in making the model to be adaptive to various environments and view point settings. Researchers have proposed domain adaptation challenges by transferring models trained on one specific domain to other variant domains [28, 198]. Most of the existing works for video-based domain adaptation focus on unsupervised domain adaptation using a large-scale unlabelled training set from the target domain and achieving unsupervised domain adaptation within the same modality [63, 84]. We first raise up the question about if few labelled target domain samples can tackle the video-based domain adaptation challenge, named as few-shot domain adaptation, as demonstrated in other fields. however there are very few works focusing on few-shot domain adaptation for video-based human action recognition. The task comparison between few-shot domain adaptation and unsupervised domain adaptation in terms of video-based human action recognition has not been clearly introduced by the community in the past. In this thesis, an analysis of the trade-off effects is conducted for video-based domain adaptation in human action recognition between two scenarios for the target domain: less data with annotations (few-shot domain adaptation) and a large amount of data without an-

notations (unsupervised domain adaptation) when only RGB modality is used. Alongside with the domain adaptation within the same modality, cross-modal domain adaptation is another challenging research direction for video-based domain adaptation in the human action recognition field due to the large discrepancy between different modalities, especially for specific downstream tasks like fall detection, where different users may prefer to use different sensor modalities when considering the privacy supporting issue.

This thesis focuses on overcoming the aforementioned generalizable challenges for both skeleton-based and video-based human action recognition. For each specific challenge, a new benchmark is constructed by considering various of related approaches and datasets, and a new specific method is proposed. By addressing these issues, this thesis aims to develop more generalizable and adaptable human action recognition methods that performs reliably across diverse conditions, maintains robustness to variations, and adapts to unseen changes, enhancing its applicability and longevity in real-world applications. The detailed contributions will be introduced in Section 1.2 and the organization of this thesis will be introduced in Section 1.3.

1.2 Contributions

This thesis addresses five key challenges in human action recognition field, focusing on enhancing the generalizability of the existing deep learning approaches for human action recognition from diverse perspectives.

Firstly, we tackle skeleton-based human action recognition under realistic occlusions by inserting 3D IKEA furnitures into the corresponding 3D coordinate system, contrasting with commonly used random temporal and spatial occlusions. This new proposed occlusion preserves geometric continuity, providing a more realistic and reasonable disturbance scenario for skeleton-based human action recognition methods. We find that existing self-supervised and one-shot skeleton-based methods significantly decline in performance when exposed to both random and realistic occlusions, with the performance of those models on the latter occlusion is much worse.

To address one-shot human action recognition under occlusions, we introduce Trans4SOAR , which uses a transformer architecture for multi-modal feature fusion at the patch embedding level, incorporating human body joints, bones, and velocities from 3D motion data. This method includes a prototype-based latent space consistency loss, enhancing the robustness of learned embeddings and demonstrating superior performance on all benchmarks.

For self-supervised skeleton-based action recognition under occlusions, we propose OPSTL. This method uses three-stream contrastive learning with adaptive spatial masking for data augmentation and a two-stage imputation approach. The imputation process involves KMeans clustering for

grouping similar samples and KNN for imputing missing coordinates, ensuring efficiency and accuracy. OPSTL proves effective across different camera settings, datasets, and occlusion scenarios.

Secondly, we explore the open-set challenges for the deep learning models in terms of skeleton-based human action recognition and construct the first large-scale benchmark. Existing open set recognition methods, designed specifically for RGB images, perform poorly on sparse skeleton data. We propose CrossMax, which combines cross-modal mean-max discrepancy training across joints, bones, and velocities with a channel-normalized distance-based logits calibration approach. This method significantly improves performance across different datasets and evaluation settings, addressing the open-set skeleton-based recognition challenge effectively.

The third part of this thesis focuses on domain adaptation challenges for video-based human action recognition, *i.e.*, particularly few-shot domain adaptation and cross-modality fall detection. We rebenchmark existing approaches on more diverse domains and propose RelaMiX, which enhances model generalizability through temporal relational dropout and snippet-wise attentional fusion. This method enriches target domain features using feature statistics from source and target domains and aligns representations with cross-domain information alignment loss, showing promising performance.

For cross-modal adaptation in terms of fall detection, we contribute a new method that uses domain-agnostic adversarial learning and cross-batch triplet margin loss to learn discriminative embeddings. An intermediate domain module is used to bridge latent spaces from different modalities. The proposed method shows promising performances across various datasets and backbones.

The benchmarks introduced in this thesis highlight the limitations of existing related approaches in dealing with data distribution shifts and underscore the importance of achieving generalizable human action recognition. Each proposed model addresses specific challenges, contributing to more reliable deep learning systems with substantial application value in the field of human action recognition.

1.3 Organization of this Thesis

In this section, the organization of this thesis for the remaining content will be illustrated. In Chapter 2, the background of this thesis is introduced, where the feature learning backbones and the generalizable challenges for each specific kind of the action recognition are presented respectively. In Chapter 3, one-shot and self-supervised skeleton based human action recognition under occlusions are benchmarked separately in Section 3.1 and Section 3.2, where one specific new deep learning solution is proposed for each of these two generalizable challenges. Then, in Chapter 4, open-set challenge for skeleton-based human action recognition is introduced. Furthermore, in

Chapter 5, few-shot domain adaptation challenge with the RelaMiX method is illustrated in Section 5.1 and the cross-modality fall detection method is presented in Section 5.2 for video-based human action recognition. Finally, the conclusions and remarks are discussed in Chapter 6.

2 Related Work

In this section, the related existing researches of the deep learning-based human action recognition backbones and the generalizable tasks will be introduced. Human action recognition methods can be divided into two clusters based on the input modality, *i.e.*, skeleton-based methods and video-based methods. Skeleton-based methods mainly use Graph Convolutional neural Network (GCN) [209] and graph transformer network [224]. The coordinates of human body keypoint are commonly used as input. On the other hand, video-based approaches predominantly rely on 3D convolutional layers [19] or vision transformer layers [123], and use RGB video as input.

Due to the differences between the two leveraged modalities, each kind of human action recognition approach faces unique generalization challenges. Skeleton-based methods, for instance, encounter challenges related to diverse occlusions in one-shot learning and self-supervised learning scenarios. This is because the reliance on keypoint coordinates can be problematic when body parts are obscured or not visible, making it difficult to accurately recognize actions. Moreover, these methods need robust strategies to generalize across various subjects and movements with limited labeled data.

Conversely, video-based human action recognition methods face essential domain adaptation challenges. Since these methods rely on RGB video input, they must effectively handle variations in lighting, background, and camera perspectives to maintain high performance across different domains. Achieving generalizable video-based action recognition in diverse environments and conditions is thereby crucial. In the following, the feature extraction backbones and the generalizable challenges are separately introduced for these two modalities, respectively.

2.1 Skeleton-Based Action Recognition and Challenges

2.1.1 Feature Learning Architectures

Skeleton-based human action recognition is important for real world human action recognition applications requiring high reliability and efficiency. Early research focused on Convolutional

Neural Networks (CNN)-based methods [176, 188, 193], known for hierarchical feature learning. Recurrent Neural Networks (RNN)-based methods [54, 169, 222] were also studied for their ability to model temporal dynamics within sequences, effectively capturing temporal dependencies for action recognition tasks.

To leverage more skeletal geometric information regarding the relationship between different human body joints, Graph Convolution Network (GCN)-based methods [33, 149, 209, 214] emerged, focusing on topology modeling. These methods capture spatial relationships and dependencies between body joints by modeling the human skeleton as a graph, where nodes represent joints and edges represent bones. These approaches allow for efficient action recognition. ST-GCN [209] initially used a fixed graph convolution approach to model dynamic spatial-temporal relationships with a predefined skeleton topology. Liu *et al.* [121] proposed large-kernel attention within a GCN to improve the action reasoning. Peng *et al.* [150] used Neural Architecture Search (NAS) to design a GCN exploring spatial-temporal correlations with multiple dynamic graph modules and multi-hop connections. Dynamic GCN [214] introduced adaptive graph topologies, enhancing modeling flexibility and accuracy. Chen *et al.* [34] proposed a multi-scale spatial-temporal GCN. CTR-GCN [33] refined graph topology at a channel-wise level, improving the capture of spatial-temporal relationships.

Recently, the focus has shifted to graph transformer-based methods [2, 101, 224],since transformer approach is valued for capturing long-range dependencies which show advantages in temporal reasoning. Transformers excel in handling large-scale data and modeling complex relationships across sequences, making them ideal for action recognition tasks with intricate and extended actions [142]. Next, the generalizable challenges in the skeleton-based human action recognition field will be illustrated.

2.1.2 Occlusion Challenges

Occlusions in skeleton data severely impair human action recognition models by causing missing or corrupted joint information, leading to misinterpretations or failures in recognizing actions. Spatial occlusions obscure specific body parts, complicating the accurate identification and tracking of key points necessary for action recognition [165]. Temporal occlusions interrupt movement sequences, which is crucial for understanding action progression. Addressing both spatial and temporal occlusions is critical for developing robust and generalizable action recognition systems that perform well in real-world, dynamic environments.

Most skeleton extraction methods, such as AlphaPose [60, 203], output zeros for occluded joints. To tackle occluded action recognition, some researchers simulate occlusion by randomly setting different body regions to zeros per frame (mimicking spatial occlusion) or by setting randomly selected frames to zero (mimicking temporal occlusion) [4, 50, 68, 103, 172, 173]. Additionally, self occlusion

caused by body movement is considered in [119].

Shi *et al.* [165] propose an occlusion-aware multi-stream fusion graph convolutional neural network to address occlusions using different streams. However, most existing occlusions are randomly generated and do not preserve the geometric continuity of obstacles. This thesis introduces a new approach by inserting 3D Ikea furniture models into the coordinate system to generate more realistic occlusions. This realistic occlusion is verified to be more complicated when the skeleton-based human action recognition models are facing with diverse generalizable challenges.

2.1.3 One-Shot Recognition Challenge

One-shot recognition, a subfield of data-scarce representation learning, aims at recognizing unseen categories with only one reference sample as guidance. This setting is valuable to real world applications in case we want the model to quickly adapt to previous unseen categories by minimal effort. In contrast to one-shot image classification, where meta-learning approaches [11, 77, 79, 187, 205, 233] dominate by re-initializing a new task set every epoch according to the learning-to-learn paradigm, Deep Metric Learning (DML) based approaches [134, 135, 234] have been effectively utilized for Skeleton-based One-shot Action Recognition (SOAR). These DML methods aim to achieve highly discriminative representations and minimize the distance between inter- and intra-category samples in the latent space, as benchmarked by the NTU-120 dataset [115] with predefined reference frames. One-shot action recognition has been extensively studied in several downstream tasks, such as semantic segmentation [220] and video classification [15, 76, 144, 195]. However, research specifically focused on SOAR is much sparser and primarily benchmarked on the NTU-120 dataset [115, 117, 118, 134, 157]. Existing works alleviate the challenge of the SOAR task through learning discriminative embeddings. Memmesheimer *et al.* [134, 135] propose to use image-wise encoding method on the skeleton sequence in order to enable the utilization of the powerful CNN for feature extraction. Their framework is optimized with deep metric learning using a combination of cross-entropy and triplet margin losses, ensuring generalizable and accurate recognition performance. Spatial-temporal adaptive metric learning network is proposed by Li *et al.* [109] to serve as a good solution for SOAR. Multi-scale spatial temporal skeleton matching is proposed by Yang *et al.* [212], which uses more reliable and accurate cross-scale matching to achieve more accurate SOAR. Zhu *et al.* [231] propose adaptive local-component-aware graph convolutional network which can be utilized as effective feature learning backbone for SOAR. In this thesis, we for the first time explore the SOAR task under diverse occlusion challenges.

2.1.4 Self-Supervised Learning Challenges

Self-Supervised Learning (SSL) is a deep learning task where models are asked to learn discriminative embeddings without labeled supervision, which leverages intrinsic data structures for supervision instead of relying on labels. SSL is especially beneficial when labeled data is limited or costly in the real world applications, thereby requires discriminative and generative feature learning without labeled supervision. In skeleton-based action recognition, SSL utilizes unlabeled skeleton sequences to develop meaningful action representations learning, reducing dependency on extensive labeled datasets [83, 120, 138]. Most of the researchers use contrastive learning on SSL of skeleton-based human action recognition approaches to enable discriminative motion embeddings learning. MS2L [112] introduces a multi-task SSL framework involving motion predictions and jigsaw puzzles. SkeletonCLR [106] utilizes momentum updates in contrastive learning on individual streams. CrosSCLR [106] employs a cross-view knowledge mining strategy for knowledge sharing between different streams. AimCLR [71] explores multiple data augmentation methods for contrastive learning. PSTL [229] uses a spatiotemporal masking strategy to learn generalized representations from partial skeleton sequences, addressing the over-reliance on data augmentation. In this thesis, we will explore SSL for skeleton-based human action recognition under occlusion disruptions. Occlusions and missing data disrupt sequence continuity, complicating the SSL learning process as models must infer missing information accurately. Designing effective pretext tasks that capture spatial and temporal dynamics of human actions without labels is particularly challenging, as they must ensure models learn relevant features without being misled by disruptions. Ensuring learned representations generalize well to various actions and subjects is another significant hurdle. SSL models may overfit to specific training patterns, limiting their effectiveness in diverse real-world scenarios [56]. Occlusions have possibility to further amplify this limitation due to the perturbation on the sparse skeleton data. Innovative data augmentation approaches are necessary to simulate various occlusions and movements duting training, enhancing model robustness and generalizability [72]. However, most of the existing SSL works on skeleton-based human action recognition task are verified to be less effective under the occlusion challenges. Developing robust and generalizable pretext tasks and advanced model architectures capable of handling the intricacies of skeleton-based data are essential for overcoming these challenges, which will be handled by this thesis.

2.1.5 Open-Set Recognition Challenge

open-set recognition is a generalizable challenge where the model is required to accurately identify instances from known classes while also detecting when an input belongs to an unknown class not encountered during training [67, 179, 228]. This capability is essential for real-world applica-

tions, as systems often face novel or unforeseen instances that are not from the part of the training categories. Unlike close-set recognition, which operates under the assumption that all test data falls within the known classes, open-set recognition demands that the model strikes a balance between the precise classification of known categories and the effective identification of unknowns.

Addressing the challenges of open-set recognition involves developing models that can generalize well to new and unseen data. This requires sophisticated techniques for distinguishing between familiar and unfamiliar inputs, often involving strategies such as thresholding on confidence scores, incorporating novelty detection mechanisms, or leveraging generative models to estimate the likelihood of inputs [66]. By effectively tackling this task, models can significantly enhance their generalizability and reliability, making them better suited for deployment in dynamic and unpredictable environments where new patterns and classes continuously emerge. Open-set recognition is nearly overlooked by the community for the task of skeleton-based action recognition, related works are mostly conducted in other fields, *e.g.*, image classification and video-based action recognition. We examine the performances of several well-established open-set image classification and open-set video-based action recognition approaches which can be adapted for Open-Set Skeleton-based Human Action Recognition (OS-SAR) by replacing backbone and input data using GCN methods and skeleton sequence data. Shi *et al.* [166] propose an OS-SAR approach using a 3D neural network on joints heat map as the backbone with deep evidential learning, which can be regarded as an implementation of DEAR [10], while no comprehensive OS-SAR benchmark is contributed and the datasets leveraged are not commonly used in skeleton-based action recognition. We implement this approach by substituting the backbone into different GCNS in our benchmark since GCN is the dominant backbone to handle skeleton data. In the field of open-set image classification, numerous works have been presented [24, 25, 65, 74, 127, 143, 179, 217]. Hendrycks *et al.* [74] is the first to use the highest SoftMax score as the open-set probability, paving the way for subsequent reconstruction-based approaches [21, 143, 179, 217]. Recently, prototype-based methods have shown great promise [24, 25, 179]. For instance, reciprocal points distance serves as the open-set probability in works by Chen *et al.* [24, 25], while PMAL [127] is one promising approach in the field of open-set recognition which has superior performance. Cen *et al.* [21] propose a new task for unified few-shot open-set recognition. For our benchmark baselines, we choose to use SoftMax, RPL, ARPL, and PMAL. SoftMax serves as a lower bound for OS-SAR, whereas the other methods, due to their success in open-set image classification, have significant potential to deliver superior performance in OS-SAR.

In the early stages of open-set video classification, Shu *et al.* [168] introduce the Open Deep Network (ODN) by incrementally adding novel classes to the recognition head to achieve awareness of new classes. Following this, Krishnan *et al.* [94] and Subedar *et al.* [178] utilize Bayesian neural net-

works to achieve reliable uncertainty estimation. DEAR [10] establishes a large-scale benchmark for open-set video-based human action recognition and proposed an architecture that employs deep evidential learning, achieving state-of-the-art performance. Humpty Dumpty [53], renamed Humpty in our benchmark, uses clip-wise relational graphical reconstruction error as the open-set probability. Additionally, Monte Carlo Dropout with Voting (MCD-V), proposed by Roitberg *et al.* [155], is used for open-set video-based driver action recognition. Yang *et al.* [213] leverage micro-doppler radar data for open-set recognition, but we do not adapt this model due to its specific architecture tailored for that modality. In this thesis, we observe that the performances provided by the above-mentioned methods are limited due to the sparsity of the skeleton data. We propose to use a novel cross-modal logits calibration method which will be introduced in detail in the following chapters.

2.2 Video-Based Action Recognition and Challenges

2.2.1 Feature Learning Architectures

Supervised video-based human action recognition methods [19, 61, 147, 170, 186, 192, 202] have achieved impressive results with deep learning algorithms in recent years. These video-based approaches can be broadly categorized into Convolutional Neural Network (CNN)-based and transformer-based methods.

The Two-Stream Network [170] is an early and influential model that includes a spatial feature aggregation stream and a temporal feature aggregation stream, capturing information from still video frames and motion across frames, respectively. The final prediction is made by fusing these two streams in late fusion manner. Regarding CNN-based methods, most approaches utilize 3D CNNs combined with various temporal aggregation techniques. For example, Temporal Segment Network (TSN) [192] samples a fixed number of video frames evenly across the video segments and uses these sampled frames as input for a two-stream network. The C3D model [186] employs a full 3D convolutional architecture for spatiotemporal feature extraction. The Inflated 3D ConvNet (I3D) [19] utilizes an inflated Inception v1 model [181], incorporating 3D convolutional layers in each stage. S3D [202] decomposes the lower 3D convolutional layers of I3D into separate spatial and temporal convolution operations. The X3D model [61] systematically expands a small 2D image architecture into a spatiotemporal one by exploring multiple axes such as space, time, channels, and depth, achieving an optimal trade-off between accuracy and complexity. Recently, there has been a shift in backbone architecture design for video-based human action recognition, from CNNs to Transformers. Transformer-based backbones [12, 58] have been increasingly employed.

Multiscale Vision Transformer (MViT) [58] combines convolutional layers and transformer lay-

ers to extract both low-level and high-level features from video frames at multiple scales. TimeSformer [12] explores action recognition with a convolution-free architecture, applying temporal and spatial attention separately within each block of the network and aggregating this information to make predictions. Leveraging the inherent spatiotemporal locality of videos, the Video Swin Transformer [123] approximates full spatiotemporal self-attention using local self-attention, outperforming factorized models in terms of efficiency.

These advancements highlight the evolving landscape of video-based human action recognition, where both CNN-based and transformer-based approaches continue to push the boundaries of performance and efficiency in video-based action recognition. Next, the generalizable challenges in video-based human action recognition field will be introduced.

2.2.2 Domain Adaptation Challenges

2.2.2.1 Few-Shot Domain Adaptation

Domain Adaptation (DA) [22, 28, 64, 107, 158, 201, 204] refers to the scenario where training and test data originate from related but distinct domains. The objective of DA is to adapt a learner to a target domain by jointly leveraging source domain and target domain samples under different labeling conditions. Unsupervised Domain Adaptation (UDA) addresses the inter-domain discrepancy without the labels of target domain samples during training. Image-based tasks have extensively employed UDA methods [87, 99, 107, 124, 125, 201]. For example, DAN [124] and JAN [125] align the marginal distributions of source and target domains by minimizing Maximum Mean Discrepancy (MMD) and Joint Maximum Mean Discrepancy (JMMD), respectively. CAN [87] enhances feature discrimination by leveraging both inter- and intra-class discrepancies.

In the context of video-based UDA, several methods utilize adversarial learning frameworks to handle domain shifts [28, 41]. Beyond adversarial learning, Wei *et al.* [198] employ disentanglement learning to separate content and context information to harvest a better adaptation. CoMix [158] introduce target domain background information into source domain samples during training to reduce domain shift. DANN [63] uses an adversarial learning framework within a 3D-CNN architecture to align source and target domains. TA^{3}N [28] employs a multi-level adversarial framework that attends to, aligns, and learns temporal dynamics across domains.

Recent research in Semi-Supervised Domain Adaptation (SSDA) [104, 159, 210, 216] relax the strict UDA constraint by using a partially annotated target domain training set. Saito *et al.* [159] propose a method using minimax entropy (MME) regularization to learn domain-invariant features. Domain adaptation for few-shot learning [39, 221] tackle UDA where new classes appear in the target domain test set.

Yang *et al.* [210] decompose the SSDA task into intra-domain semi-supervised learning (SSL) and inter-domain UDA tasks, utilizing co-training for complementary information learning. Qin *et al.* [154] reduce decision boundary bias by aligning the conditional distribution between scattered source features and clustered target features. Yoon *et al.* [216] propose a pair-based SSDA method that adapts a model to the target domain using self-distillation to bridge domain discrepancies gradually.

Few-Shot Domain Adaptation (FSDA) [81, 85, 139, 182, 204] provides only a few labeled samples per class for the target domain training set. Unlike SSDA, FSDA-AR does not rely on a large-scale unlabeled training set from the target domain. Research in FSDA for video-based Action Recognition (FSDA-AR) is limited, with only three works targeting this task [64, 108, 204]. PASTN [64] uses an attentive adversarial network for learning domain-invariant features and fine-tunes the model on a small number of labeled target samples after pre-training on a large labeled dataset. FS-ADA [108] integrates category classifier and domain discriminator to extract domain-invariant and category-discriminative features. Xu *et al.* [204] utilizes Timesformer [12] as the backbone and employ prototype-based snippets contrastive learning for FSDA-AR.

However, the benchmark for PASTN [64] is not available, and the datasets used by Xu *et al.* [204] do not encompass different levels of domain shift, nor do they unify the feature extraction backbone like I3D [19], which is widely used in UDA. Additionally, FS-ADA [108] is developed for radar data. For a fair comparison, UDA approaches should be reformulated for the FSDA-AR task to adapt against diverse domain shifts. Further research is needed for video-based FSDA-AR on diverse domain shifts, such as different views of egocentric videos and transitions from synthesized to real videos.

To address this issue, we introduce a novel video-based FSDA-AR benchmark with diverse domain combinations under a unified feature extraction backbone. This benchmark is designed to facilitate fair comparisons and further research in video-based FSDA-AR, promoting the development of more generalizable and adaptable models in this field.

The RelaMiX approach proposed in this dissertation enhances video data generalization through three components: a Temporal Relational Attention Network with Relation Dropout (TRAN-RD) for better temporal embeddings, a Statistic Distribution-Based Feature Mixture (SDFM) method inspired by [211] for diversified latent space targeting the domain, and Cross-Domain Information Alignment (CDIA) loss using contrastive supervision with mixed domain negatives and prototype positives. This method demonstrates strong performance in FSDA-AR across five datasets.

2.2.2.2 Cross-Modal Adaptation for Fall Detection

Existing research on fall detection can be clustered into three major groups based on the sensors employed. The first group focuses on wearable devices, which predominantly use accelerometers to capture signals from various body parts such as the wrist, chest, and waist [1, 7, 16, 111, 206, 215]. For instance, Chen *et al.*[27] utilize a smartwatch worn on the wrist to monitor individuals' movement status. Mehmood *et al.* [133] introduce a novel wearable sensor called SHIMMER, which measures signals from the waist.

The second group of fall detection research leverages Wi-Fi signal networks [20, 23, 48, 78, 191, 197]. In their work, Wang *et al.*[197] propose WiFall, a system that detects falls by analyzing the correlation between radio signal variations and human activities. Similarly, Hu *et al.*[78] develop DeFall, a system comprising an offline template-generating stage and an online decision-making stage, utilizing Wi-Fi features associated with human falls.

The third group involves vision-based approaches [6, 14, 57, 93, 183, 207, 226], which typically use action recognition models as the feature extraction backbone for fall prediction. Several datasets facilitate video-based fall detection, such as the UR Fall dataset [98], Kinetics dataset [89], NTU dataset [114], and UP Fall dataset [131]. For example, Khraief *et al.* [90] introduce a multi-stream deep convolutional neural network employing both RGB and depth modalities for fall detection. Na *et al.* [128] use a combination of 3D-CNN and LSTM as the backbone to extract features from RGB videos for fall detection. Chen *et al.* [32] propose an attention-guided bi-directional LSTM to achieve fall detection in complex backgrounds. Asif *et al.* [6] address privacy concerns by using body skeletons and semantic segmentation masks as input for fall detection, discarding the conventional RGB data.

However, these existing works neglect the potential use case of cross-modal domain adaptation in the fall detection task, which is a downstream task in the filed of human action recognition with great application value. Cross-modal domain adaptation [29, 84] indicates that the source domain and the target domain preserves two distinct modalities. The aim of the cross-modal domain adaptation is to achieve domain knowledge transfer by using labeled source domain training set and unlabeled target domain training set. Cross-modality domain adaptation is crucial because it maximizes the utilization of available data, enhances privacy by transitioning from detailed RGB images to less intrusive modalities like depth images, and improves model generalizability across different sensor conditions. Since fall detection system is highly likely to be employed in smart home environment, achieving more privacy-supporting fall detection system is important. However, there is no existing work conducted for cross-modal adaptation in terms of fall detection, which will be explored by this thesis. On the next several chapters, these generalizable challenges and their corresponding solutions are demonstrated in details.

3 Towards Generalizable Skeleton-Based Human Action Recognition under Occlusions

We begin to introduce of the generalizable challenges under occlusions on the skeleton-based human action recognition in this chapter. The capacity for generalizable skeleton-based human action recognition is critically important, as it ensures models sustain their efficacy across a multitude of environments, user demographics, and situational variances, thereby obviating the exhaustive requirement for model retraining tailored to each distinct context. Such versatility is particularly vital in real-world applications. However, occlusions significantly hinder the effectiveness of skeleton-based human action recognition models by compromising the integrity and continuity of skeletal data. This leads to several issues: a lack of critical movement information, disruptions in spatiotemporal correlations essential for identifying complex actions, and exacerbated data scarcity, especially where labeled data is limited. Additionally, the unpredictable attribute of real-world occlusions complicates the development of adaptable skeleton-based human action recognition models that can handle diverse occlusion types and extents while maintaining generalizable recognition capabilities. In this chapter, we will explore two particular tasks, namely, *one-shot recognition* (Section 3.1) and *self-supervised learning* (Section 3.2) under occlusions, to investigate the intricacies of generalizable skeleton-based human action recognition in the face of data perturbation challenges. Part of the works are from the paper of our publication on IEEE Transactions of Multimedia [146] and our paper [31] which is now under review of IROS.

3.1 Delving Deep into One-Shot Skeleton-Based Action Recognition with Diverse Occlusions

3.1.1 Introduction

In the field of action recognition utilizing skeletal data, significant strides have been made due to the rapid advancements in deep learning technologies. Traditional methodologies in this field have achieved promising performances across numerous benchmark datasets designed for body-pose classification, such as NTU-120 [115] and Toyota Smart Home [47]. These datasets have been meticulously curated to ensure unobstructed visibility of the body by strategically positioning cameras, as documented in studies by Liu *et al* [115], Zhang *et al* [224], Chen *et al* [33], and Yan *et al* [209]. Despite these achievements, the simplifying assumption of clear visibility of the skeleton data becomes less tenable in real-world settings where occlusions frequently impair the quality of input.

The interest in skeleton-based action recognition algorithms, particularly those analyzing 3D body joint coordinates, has surged due to the enhanced accuracy of depth sensors and their capability to respect privacy concerns [8, 33, 36, 132, 152, 153, 171, 209, 224]. However, the challenge of occlusions looms large in this context, as the sparse attribute of skeletal representations means that missing even a minimal number of joints can severely disrupt the geometric and temporal coherence of the data.

The task of learning from a limited subset of labeled examples, defined as one- or few-shot recognition [15, 64, 76, 144, 151], remains a critical challenge in skeleton-based human action recognition. This challenge is exacerbated in scenarios with occlusions, where the limited variety of available data for new categories places a premium on the quality of the few samples provided, complicating the task of accurate recognition.

Our work sets forth to tackle the goal of classifying sequences of unseen 3D actions from a single reference example, despite the presence of missing segments attributable to occlusions. Notably, prior research on one-shot action recognition from skeletal data has not directly addressed the impact of occlusions. To fill this gap, we introduce a novel benchmark that simulates occlusions by manipulating skeletons within three established action recognition datasets, using a library of 3D objects from PIX3D [180] to create REalistic occlusions (RE). These occlusions are applied to the original datasets with a variety of geometric manipulations, including rotation and displacement, aiming to more accurately reflect the complexities encountered in real-world scenarios as opposed to the artificial removal of data points practiced in previous studies [172, 173]. Moreover, we introduce an alternative occlusion method, termed RAndom occlusion (RA), which randomly alters body

joints while preserving spatial and temporal dimensions to enrich the occlusion types involved in this study.

To address these challenges, we propose a pioneering model, Trans4SOAR, grounded in transformer architecture. Prior efforts in Skeleton-based One-shot Action Recognition (SOAR) have largely relied on CNNs in combination with metric or meta-learning techniques [134, 135, 234]. Although transformers have been successfully integrated into conventional video-based action classification [5], their application in encoding body movement signals for SOAR, particularly under conditions of data scarcity and occlusions, remains unexplored.

Our approach involves various methods, specifically Skeleton-DML [134] and SL-DML [135], in the context of our newly formulated occlusion-centric benchmark. Furthermore, we explore the potential of transformer networks for encoding skeleton signals as image-like representations through our Trans4SOAR model, marking a novel application of visual transformers to the SOAR task aimed at overcoming occlusion-related obstacles.

Additionally, we introduce two novel components within Trans4SOAR. The first is the Mixed Attention Fusion Mechanism (MAFM), which integrates various types of data (velocities, bones, and joints) at the patch embedding level while taking into account spatial and temporal proximities. The second is the Latent Space Consistency (LSC) loss, which fosters model robustness against occlusions by ensuring consistency in the model's outputs from the normal branch and the prototype augmented branch. These advancements underscore our commitment to refining the capabilities of skeleton-based action recognition in the face of occlusions. This paper explicitly explores occlusions for SOAR and makes the following contributions:

- We pioneer the study of occlusions in **Skeleton-Based One-Shot Action Recognition**, an unexplored area in community. To address this challenge, we create a new benchmark tailored for this task using three well-known action recognition datasets. This benchmark involves two types of occlusions: random and realistically synthesized occlusions, with a focus on the latter generated using the IKEA 3D furniture models. Our work provides a foundation for understanding and addressing occlusions in SOAR, improving the generalizability of action recognition technologies in practical settings.

- We introduce Trans4SOAR approach, a novel three-stream transformer-based architecture designed to tackle data occlusions in skeleton-based one-shot action recognition. This model enhances resilience to occlusions through two key strategies. Firstly, it integrates diverse input modalities (velocities, bones, joints) using a mixed attention fusion mechanism at the patch embedding level, ensuring a comprehensive representation of action dynamics. Secondly, it augments intermediate representations by iteratively estimating category-specific prototypes

and applying latent space consistency loss. This stabilizes the learning process against occlusions. These novelties position Trans4SOAR as a leading solution for skeleton-based one-shot action recognition under diverse occlusions task.

- We conduct a evaluation of existing skeleton-based one-shot action recognition approaches under diverse occlusions, comparing with our Trans4SOAR method. The evaluation covers four distinct occlusion scenarios. Occlusions degrade performance metrics, highlighting an area for future research. Despite this, Trans4SOAR consistently outperforms existing frameworks across all datasets under occlusive conditions, establishing it as the superior model for this task in environments with diverse occlusions.

- The experimental results of this study reveals that Trans4SOAR excels not only in occlusion scenarios but also in conventional skeleton-based one-shot action recognition without occlusions. It surpasses previous benchmarks by over 2.8% on the challenging NTU-120 dataset. This highlights Trans4SOAR's versatility and generalizability.

3.1.2 Problem Definition

Our study aims to explore the field of SOAR, a challenging yet crucial task that seeks to utilize prior knowledge from data-abundant action classes to effectively categorize new, data-scarce classes, especially in the presence of occlusions within the skeletal data [134]. This investigation rigorously adheres to the one-shot evaluation protocol, as initially defined in the context of the NTU-120 dataset [115], which necessitates the classification of actions based on a single reference example provided for each category. To formally introduce this task, we define C_{base} as the set of $|C_{base}|$ data-rich categories, which are made available during the training phase. The dataset is denoted by $D_{base} = \{(\mathbf{s}^i, l^i)\}_{i=1}^{U}$, where each label l^i belongs to C_{base}, and U represents the total number of samples within D_{base}. The core aim of this study is to accurately classify the $|C_{novel}|$ new action classes encapsulated within the set C_{novel}, each represented by a single reference instance within the support set $D_{supp} = \{\mathbf{s}^i\}_{i=1}^{O}$, with O denoting the count of samples in D_{supp}. It is imperative to note the exclusionary relationship $C_{base} \cap C_{novel} = \emptyset$, thereby ensuring no overlap between previously encountered and new categories. The culminating challenge involves assigning the correct category $l^n \in C_{novel}$ to each instance within the test set D_{test}, which only consists samples from the new categories in C_{novel}.

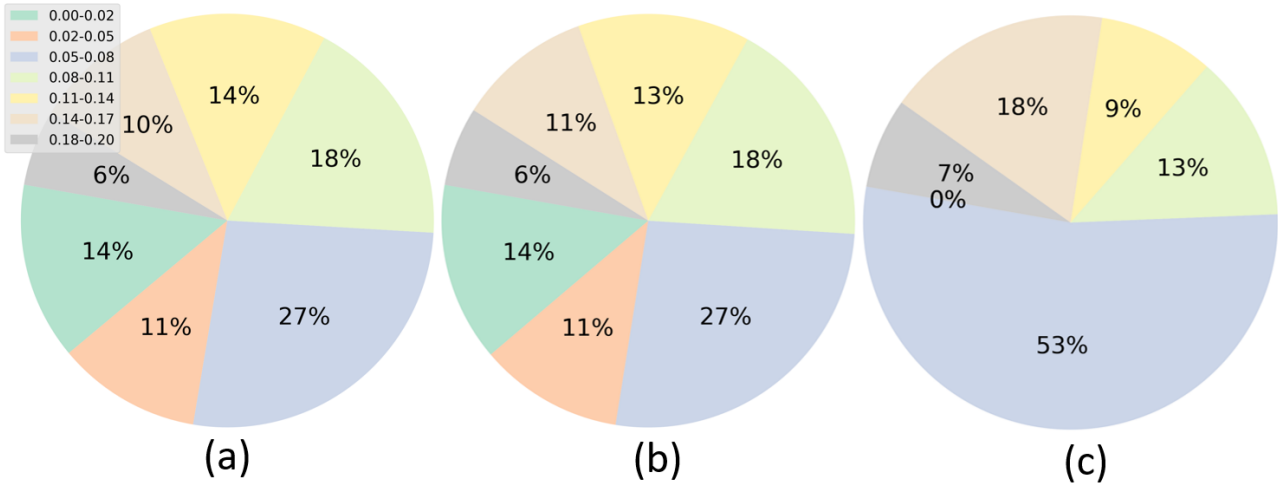


Figure 3.1: An overview of Signal-to-Noise Ratio (SNR) distribution of the realistic synthesized occlusion datasets, where (a), (b) and (c) are for the NTU-120, the NTU-60 and the Toyota Smart Home datasets respectively. The legend indicates the corresponding SNR range. The SNR is calculated using the division of number of occluded skeleton joints per skeleton sequence and the number of joints per skeleton sequence to denote the perturbation level.

3.1.3 Occlusion Types

3.1.3.1 Realistic Synthesized Occlusion

In the existing researches on skeleton-based action recognition, the occlusions, particularly in scenarios not constrained by data scarcity, has predominantly been simulated through the combination of the random temporal or spatial disruptions. This approach is manifest in methodologies that arbitrarily nullify a predetermined quantity of frames or joints, thereby simulating occlusions by directly setting the occluded frames or joints to zero [173]. While this strategy allows for a high degree of control over the occlusion parameters, it fails to adequately replicate the complex and often unpredictable character of real-world occlusions, which are frequently caused by objects that induce missing skeleton points with distinct geometric patterns.

To bridge this gap between the random occlusions without any geometric constraint used in prior research and the multifaceted nature of occlusions encountered in actual environments, our study ventures into the creation of occluded variants of three publicly accessible datasets: NTU-120 [115], NTU-60 [161], and Toyota Smart Home [47]. This endeavor is facilitated by the integration of 3D furniture models from the IKEA furniture collection, as sourced from the PIX3D dataset [180], into the world coordinate system associated with skeleton data. Consequently, the method for augmenting each dataset with realistic synthesized occlusions is tailored to accommodate these differences,

a process that is expounded upon in subsequent sections of this study.

Furthermore, to quantitatively assess the statistics of these realistically synthesized occlusions on each dataset, we provide the Signal Noise Ratio (SNR) for these datasets with the newl occlusion, as shown in Figure 3.1.

3D realistic synthesized occlusion dataset generation. The NTU-120 [115] and NTU-60 [161] datasets encompass diverse camera perspectives recorded at the same time, necessitating the notice of cross-view consistency when integrating occlusions from furniture models into skeleton sequences. A primary challenge in this regard is the absence of calibration data for the cameras involved, which would ideally provide foundational insights for addressing cross-view discrepancies.

Fortunately, our focus on skeleton data presents a unique advantage, as each skeleton frame within these datasets offers comprehensive world coordinate information pertaining to body joints. This information is important to deduce the relative positioning of cameras through the dataset's inherent structure. Specifically, a single sample of a skeleton sequence comprising T frames and J joints furnishes us with a total of $T \cdot J$ coordinate points. With N such samples available, the aggregate number of known world coordinates escalates to $N \times T \times J$, which substantially surpasses the rank of any projection matrix linking two distinct cameras. Consequently, the calibration matrix $\mathbf{F}_{ij}$, delineating the relationship between any two cameras i and j, can be represented as Eq. 3.1,

$$\mathbf{F}_{ij} = (\mathbf{s}_i^T \mathbf{s}_i)^{-1} \mathbf{s}_i^T \mathbf{s}_j^{'}, \tag{3.1}$$

where $\mathbf{s}$ represents the collective body joint coordinates, formatted in homogeneous coordinates, captured simultaneously by the two cameras. This methodology enables the derivation of the projection matrix $\mathbf{F}_{ij}$ as an element of the set $\mathbf{F}$, facilitating cross-view consistency.

A comprehensive algorithmic procedure for generating 3D occlusions is presented in Alg. 1. This process initiates with the random selection of a 3D object model from the extensive IKEA furniture dataset PIX3D [180], which comprises 395 models across 9 categories. The selected model is then augmented using random rotations and translations, to simulate real world randomness.

To maintain cross-view consistency, samples recorded concurrently by different cameras are augmented with the same furniture model, adjusted according to the calibration matrix set $\mathbf{F}$. Subsequently, we determine which body joints are obscured by the occlusion for each camera perspective. This involves projecting the skeleton joint $\mathbf{s} = [s_1, s_2, s_3]$ and the furniture model points $\mathbf{z}^{'}$ along the camera's focal axis to obtain $\mathbf{s}^{\star}$ and $\mathbf{z}^{\star}$. A two-dimensional convex hull is then constructed from the projected furniture points $\mathbf{z}^{\star}$, and the position of skeleton joints within this hull is ascertained as

Eq. 3.2,

$$IsInHull = IsTrue(A \times (\mathbf{s}^{\star T}) \leq Tile(-\mathbf{b}, (1, len(\mathbf{s}^{\star}))), 0), \tag{3.2}$$

where A is the convex hull's boundary equation, $\mathbf{b}$ is the boundary's offset, and $\mathbf{s}^{\star}$ is the point under evaluation. The binary indicator $IsInHull$ discerns whether a point lies inside the hull, facilitating the generation of a mask for each skeleton sample. The above equation checks if a point $(\mathbf{s}^{*})^{T}$ lies inside a convex hull defined by the half-space inequalities $A \cdot x \leq b$. It computes $A \cdot (\mathbf{s}^{*})^{T}$, compares it to the tiled $-\mathbf{b}$, and verifies if all inequalities are satisfied using the IsTrue function. The convex hull is built up based on the boundary key points derived from the furniture model to ensure the precise. This method is widely used in computational geometry and optimization to validate whether a point belongs to a convex set.

Occluded points identified through these binary indicators are then nullified, completing the occlusion simulation process.

2D realistic synthesized occlusion dataset generation. In addressing the unique characteristics of the Toyota Smart Home dataset [47], which is distinguished by its provision of 2D skeletal data within the image plane, we adapt our occlusion generation pipeline to accommodate the specific requirements of 2D data processing. Unlike the procedure for 3D realistic synthesized occlusions, which involves manipulating the 3D furniture model through rotation and translation within the world coordinate system, the adaptation for 2D skeleton data necessitates a different approach due to the inherent planarity of the data.

For the Toyota Smart Home dataset, the process begins with the application of a randomly generated projection matrix. This matrix serves to map the 3D points from the camera coordinate system onto the 2D image plane, effectively projecting the 3D furniture model onto the 2D skeletal data's domain. Following this transformation, we construct a convex hull based on the furniture model's projected points within the 2D image plane, mirroring the methodology applied in the generation of 3D occlusions.

Subsequent to the establishment of the convex hull, an occlusion-aware mask is derived using the $ISInHull$ function. This mask delineates the regions of the image plane obscured by the projected furniture model. With the mask in place, we then proceed to identify and modify the corresponding 2D skeleton joints falling within the occluded regions, setting these points to zero. Through this adapted procedure, we ensure the realistic simulation of occlusions within the 2D skeletal data of the Toyota Smart Home dataset, thereby facilitating a comprehensive examination of occlusion effects across both 2D and 3D skeletal data contexts.

Algorithm 1 The pipeline of the 3D realistic synthesized occlusion generation.

Input: F – the set of projection matrix for each camera pair; S – the set of skeleton data; Z – the collection of 3D furniture models from PIX3D dataset; R and T – random rotation and translation augmentations; S_{Occ} – a empty set for occluded skeleton data; $[a,\ b]$ – predefined occluded signal noise ratio range for the acceptance, where a is the lower limitation and b is the upper limitation.

for all $\mathbf{s} \in S$: **do**
  Set $Accept = False$.
  while $Accept! = True$ **do**
    Set $Found = False$.
    Set $N_{Occ} = 0$.
    {N_{Occ} is the occluded sample number for S_{d+1}.}
    while $Found! = True$ **do**
      Search $\mathbf{S}_d$ *collected simultaneously with* $\mathbf{s}$ *from different views.*
      Extract the calibration set F_d *for* S_d.
      Randomly select $\mathbf{z}$, *where* $\mathbf{z} \in \mathbf{Z}$.
      Obtain augmented $\mathbf{z}$, i.e., $\mathbf{z}^{'}$, *by* $\mathbf{z}^{'} = R(T(\mathbf{z}))$.
      Get Z_d *by applying* $\mathbf{f}_d \in F_d$ *on* $\mathbf{z}^{'}$.
      Define $Z_{d+1} = Z_d \cup \{\mathbf{z}^{'}\}$ *and* $S_{d+1} = S_d \cup \{\mathbf{s}\}$.
      if Z_{d+1} *has no intersection with* S_{d+1} *for each corresponding element:* **then**
        $Found = True$
      end if
    end while
    for $(\mathbf{s}_d, \mathbf{z}_d) \in zip(S_{d+1}, Z_{d+1})$ **do**
      Horizontally project $\mathbf{z}_d$ *and* $\mathbf{s}_d$ *along focus axis of camera* d *into 2D plane as* $\mathbf{z}_d^{\star}$ *and* $\mathbf{s}_d^{\star}$.
      Build up 2D convex hull Φ *of* $\mathbf{z}_d^{\star}$.
      $Mask_d = IsInHull(\Phi, \mathbf{s}_d^{\star})$.
      Calculate $SNR_d = Sum(Mask_d)/len(Mask_d)$ *for* $\mathbf{s}_d$.
      Occlude $\mathbf{s}_d$ *by* $\mathbf{s}_d[Mask_d] = zeros\_like(\mathbf{s}_d[Mask_d])$.
      Append $\mathbf{s}_d$ *into* S_d^{Occ}
      if SNR_d *in* $[a,b]$. **then**
        $N_{Occ} += 1$.
      end if
    end for
    if $N_{Occ} < T_{Occ}$ or $N_{Rep} < T_{Rep}$ **then**
      Set $Accept = False$ *and* $N_{Rep} += 1$.
    else
      Set $Accept = True$.
      Del S_d *from* S *and append the* S_d^{Occ} *into* S_{Occ}.
    end if
  end while
end for

3.1.3.2 Random Occlusion

In our exploration of occlusion variants within the framework of skeleton-based action recognition, we delve into the dynamics of *random* occlusion. This occlusion type combines random temporal and spatial occlusions, reflecting methodologies previously implemented in standard skeleton-based action recognition research, devoid of data scarcity constraints [172, 173]. Random temporal occlusions entail the arbitrary omission of a predetermined number of frames within each skeleton sequence, thereby emulating instances of complete occlusion at specific intervals. Concurrently, random spatial occlusions involve the selective nullification of a specified number of randomly selected joints across every frame in the data stream, introducing a nuanced form of occlusion where visibility of certain joints fluctuates unpredictably across the sequence.

This dual strategy of integrating random temporal with random spatial occlusions presents a scenario that aligns more closely with real-world conditions, characterized by less predictable and less controllable occlusions. For analytical purposes, we represent the skeleton data as $\mathbf{s} \in \mathbb{R}^{T \times J \times B}$, where T denotes the number of frames, J denotes the number of joints, and B denotes the body dimensionality. This data is subsequently flattened into a matrix of dimensions $\mathbb{R}^{(T \times J) \times B}$, facilitating the random selection of data points $\gamma \cdot (T \times J)$ to be occluded, with γ indicating the predetermined SNR that guides this selection process.

While recognizing the mixed model of spatial and temporal occlusions as a more pragmatically viable variant, our study also extends to examining the effects of isolated random spatial and temporal occlusions. This study enables us to establish comparison, thereby enhancing our understanding of the relative complexities introduced by each occlusion type. Next, the proposed new deep learning method to tackle SOAR under scenarios will be introduced.

3.1.4 Trans4SOAR

We introduce Trans4SOAR – a three-stream transformer-based model designed to overcome adverse effects of occlusions (an overview is provided in Fig. 3.2). The proposed Trans4SOAR model is a cutting-edge transformer-based framework specifically designed for Skeleton-based One-Shot Action Recognition (SOAR), addressing the critical challenge of occlusions in skeleton sequences. The architecture employs three distinct data streams—joints, bones, and velocities—each encoded into image-like representations. These streams are processed through patch embedding layers and fused via the Mixed Attention Fusion Mechanism (MAFM), which operates at the patch embedding level to ensure effective integration of complementary information across modalities. MAFM combines spatial-temporal dependencies using a novel Softmax Concentrated Aggregation (SCA) approach, which aggregates query, key, and value features from different streams, enabling robust

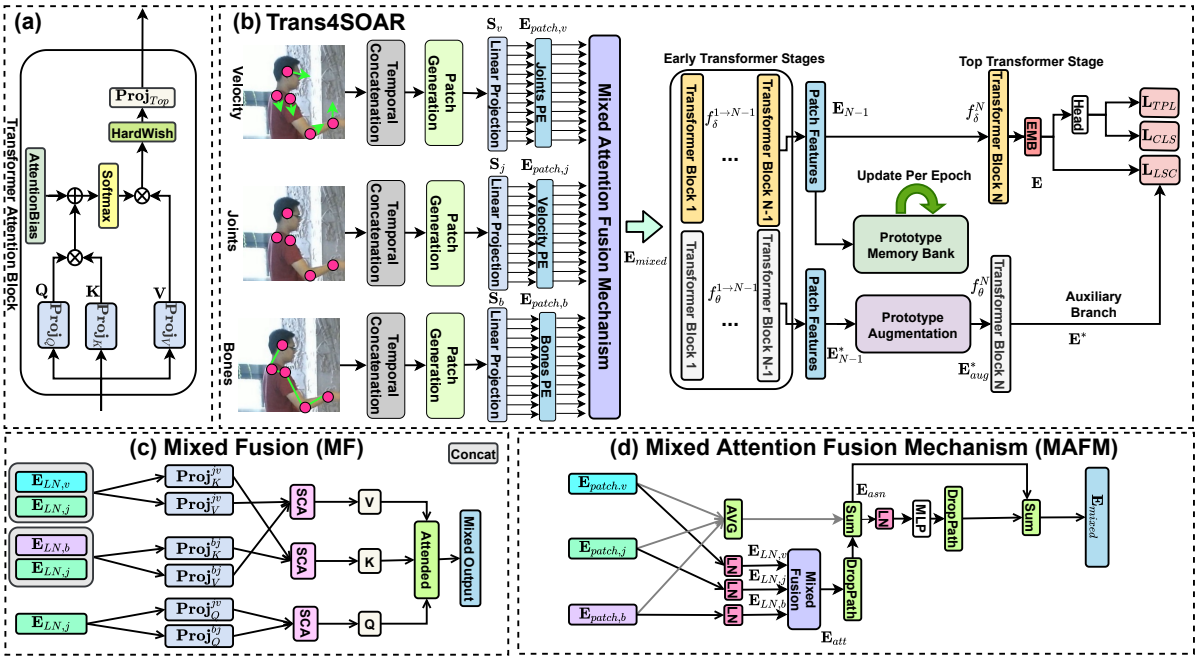


Figure 3.2: An overview of the proposed Trans4SOAR architecture, which is a **Trans**former for **S**keleton-based **O**ne-**S**hot **A**ction **R**ecognition. (a) indicates the transformer block leveraged in Trans4SOAR. This basic transformer attention block is proposed by LeViT [69], which builds up the transformer block in the later stage of our Trans4SOAR architecture through stacking. (b) is the overview of the Trans4SOAR training pipeline. First, the skeleton signals are encoded in three kinds of format, *i.e.*, joints, bones, and velocities. Image-like representations are formulated through the concatenation along the temporal axis of the skeleton data, which are further divided into several patches and fed into its corresponding patch embedding net. Then, the Mixed Attention Fusion Mechanism (MAFM) fuses the embeddings from these three different streams by using Mixed Fusion (MF) to achieve cross-stream aggregation on Key, Query, and Value together with the proposed Softmax Concentrated Aggregation (SCA). The Latent Space Consistency (LSC) loss L_{LSC} integrates an prototype augmented auxiliary branch and adopts cosine similarity loss to encourage the embeddings from the main branch $\mathbf{E}$ and the embeddings from the auxiliary branch $\mathbf{E}^*$ to be more similar. Three losses, *i.e.*, triplet margin loss (L_{TPL}), cross entropy loss (L_{CLS}), and LSC loss (L_{LSC}), are leveraged for discriminative representation learning. EMB indicates embedding generation layers, which are built based on multi-layer perceptrons (MLP). Head indicates a fully-connected (fc) layer based classification head. PE indicates the patch embedding network. (c) shows the workflow of the Mixed Fusion (MF) and (d) shows the Mixed Attention Fusion Mechanism (MAFM), where **Proj** indicates the fc-based projection layer, AVG indicates the average operation and LN indicates layer normalization.

feature extraction even under occlusion-induced distortions.

A distinguishing feature of Trans4SOAR is the Latent Space Consistency (LSC) loss, which enhances robustness to occlusions by leveraging an auxiliary branch enriched with category-agnostic prototypes. These prototypes act as anchors, encouraging consistency between embeddings from the main and auxiliary branches, even when the input data is degraded by occlusions. The LSC loss achieves this through a warm-up phase using self-augmentation and a subsequent prototype-based augmentation phase, refining feature representations by aligning embeddings with prototype distributions. This loss not only improves robustness but also facilitates better generalization to unseen action categories, a core challenge in one-shot learning.

3.1.4.1 Illustration of the Base Components

Input encoding. In the formulation of our approach, we employ the skeleton representation methodology as delineated by SL-DML [135], which facilitates the transformation of sequential skeleton data into image-like formats for processing. Let us denote a given sequential skeleton sample by $\mathbf{s} \in \mathbb{R}^{T \times J \times C}$, where T represents the temporal dimension of the sample, J denotes the total count of joints within the skeleton, and C denotes the dimensional characteristics of the joint coordinates.

To reformulate these skeleton sequences into the regular shape required by the vision transformers, bilinear interpolation is applied to upsample the skeleton data from its original $T \times J \times C$ format to a dimensionality of $H \times W \times C$, thereby aligning with the input specifications typical of image processing models. This preprocessing ensures that our model can interpret skeletal data in a manner akin to image inputs, leveraging the spatial-temporal nuances embedded within the skeletal sequences.

Distinctively, Trans4SOAR is conceptualized as a triple stream architecture before the fusion of the patch embeddings, which enriches its analysis by not solely focusing on joint information but also incorporating the dynamics of bones $\mathbf{b}$ and velocities $\mathbf{v}$. Here, the velocity $\mathbf{v}_t = \mathbf{s}_t - \mathbf{s}_{t-1}$ captures the temporal change for each joint, reflecting the motion dynamics at every timestep t. Similarly, the bone vectors $\mathbf{b}_{i,j} = \mathbf{s}_i - \mathbf{s}_j$, for each pair $(i, j) \in \Omega_{bones}$, encode the structural relationship between joints, embodying the skeletal geometry of the human body.

By mapping these derived modalities into image-wise encoding using bilinear interpolation, we can effectively translate skeletal motion and structure into formats required by CNN and transformer architectures. Consequently, this transformation yields three parallel streams of image-wise inputs: joints, velocities, and bones, each conforming to the shape of $H \times W \times C$.

Patch embedding and transformer blocks. To take advantage of the strengths of CNNs and transformers, we look to the LeViT [69] model. CNNs are great at preserving details, while transformers excel at capturing long-range dependencies. LeViT combines these strengths by using a

four-layer CNNs to create patch embeddings, followed by a series of transformer blocks. This approach blends the two architectures effectively. Building upon this powerful feature learning framework, Trans4SOAR incorporates the core transformer blocks and patch embedding strategy as delineated by LeViT [69]. Self-attention mechanism is the most essential design in transformer block. Query $\mathbf{Q}$, Key $\mathbf{K}$, and Value $\mathbf{V}$ are calculated through dedicated projection layers $\mathbf{Proj}_Q$, $\mathbf{Proj}_K$, and $\mathbf{Proj}_V$ to formulate self attention procedure. The attentive output is subsequently derived using the following Eq. 3.3,

$$\mathbf{Att} = \mathbf{Proj}_{Top}(\text{HardsWish}(\text{SoftMax}(\mathbf{Q} \times \mathbf{K}^T) + \text{Bias}_{att}) \times \mathbf{V}), \tag{3.3}$$

where each projection layer comprises a 1×1 convolution followed by batch normalization, and Bias_{att} signifies the attention bias. HardsWish is designed to optimize the balance between specificity and generalization of the learned embeddings. It modifies the conventional SoftMax operation by incorporating a hard constraint on the attention distribution, improving focus on critical regions of the data. This approach ensures robust attention weights, particularly beneficial in noisy or occluded scenarios, by emphasizing reliable feature correlations while reducing the influence of irrelevant or corrupted data, where its formula is

$$\text{HardSwish}(x) = x \cdot \frac{\text{ReLU}(x+3)}{6}. \tag{3.4}$$

In the adaptation of this architecture within Trans4SOAR, we commence by segmenting the three distinct streams of input, *i.e.*, joints, velocities, and bones, into $N_{patch} = (H/P) \times (W/P)$ patches, given a predefined patch size P.

Subsequent steps involve the construction of patch embedding layers via a series of CNNs, denoted as M_{θ_j}, M_{θ_v}, and M_{θ_b} for each input stream. These networks are tasked with extracting high-dimensional embeddings for each patch sequence, resulting in $\mathbf{E}_{patch,j}$, $\mathbf{E}_{patch,v}$, and $\mathbf{E}_{patch,b}$ for joints, velocities, and bones, respectively. Discarding traditional positional embeddings, we instead adopt the attention bias approach as depicted in Fig. 3.2(a), in accordance with the method proposed by LeViT [69]. The relationship between the input sequences and their respective embeddings is formalized in the Eq. 3.5.

$$\mathbf{E}_{patch,j},\ \mathbf{E}_{patch,v},\ \mathbf{E}_{patch,b} = M_{\theta_j}(\mathbf{s}),\ M_{\theta_v}(\mathbf{v}),\ M_{\theta_b}(\mathbf{b}). \tag{3.5}$$

The derived embeddings are subsequently directed into the MAFM within Trans4SOAR, which achieves the fusion of the multi-modal patch embeddings.

3.1.4.2 Multimodal Fusion at the Patch Embedding Level

Mixed Fusion (MF). Mixed fusion proposed by us is the most important component of the MAFM structure, aimed at enhancing the multi-stream fusion of skeletal data at the patch embedding level. This mechanism is particularly designed to facilitate the transfer of crucial information from the auxiliary streams, *i.e.*, velocities and bones, towards the primary stream of joints, to derive motion cues from different perspectives to alleviate the negative effects brought by the occlusions from the environment. Our approach adopts a novel strategy by utilizing a combination of the Key and Value elements for the purpose of multi-stream fusion, a concept partially inspired by the work in MixFormer [42], which applies a similar principle for template matching tasks.

However, the application of the Key-Value mixture within our MAFM is distinct and non-trivial, tailored specifically for the unique challenges and objectives of SOAR. Contrary to the MixFormer's objective of accentuating similarity cues within the context of template matching, our MF strategy is intricately designed to capture and exploit complementary cross-modality dependencies among the skeletal data streams. This approach aims to foster a multi-stream information emergence, thereby facilitating the learning of discriminative embeddings essential for generalizable action recognition under occlusions.

To achieve this aim, we for the first time propose a three-stream patch-embedding fusion architecture that seamlessly integrates MF and MAFM components. This architecture is strategically developed to not only merge Key and Value elements from different streams but also to ensure that such fusion accentuates the discriminative features necessary for recognizing actions from skeletal data. By doing so, the MAFM architecture uniquely positions itself to address the demands of cross-stream informative embedding fusion within the challenging SOAR task with occlusion disturbance, marking a significant departure from existing methodologies for multi-modal fusion. In the following, we introduce more details toward the proposed MF for multi-stream patch embedding fusion. First, we encode the patch embeddings of the joints $\mathbf{E}_{patch,j}$ through two different linear projection layers, *i.e.*, $\mathbf{Proj}_Q^{jv}$ and $\mathbf{Proj}_Q^{bj}$, as depicted in Eq. 3.6:

$$\mathbf{Q}_{jv},\ \mathbf{Q}_{bj} = \mathbf{Proj}_Q^{jv}(\mathbf{E}_{patch,j}),\ \mathbf{Proj}_Q^{bj}(\mathbf{E}_{patch,j}). \tag{3.6}$$

Then, for Keys and Values of the jv and bj branches, the input embeddings are aggregated together through concatenation (Concat). After this procedure, for each single term, a projection layer is leveraged for the encoding of the embedding. For example, $\mathbf{Proj}_V^{jv}$ is the projection layer for Value of the jv branch. As a result, $\mathbf{V}_{iv}$, $\mathbf{K}_{jv}$, $\mathbf{V}_{bj}$, and $\mathbf{K}_{bj}$ can be obtained after the encoding:

$$\mathbf{V}_{iv} = \mathbf{Proj}_V^{jv}(Concat(\mathbf{E}_{patch,j},\ \mathbf{E}_{patch,v})), \tag{3.7}$$

$$\mathbf{K}_{jv} = \mathbf{Proj}_K^{jv}(Concat(\mathbf{E}_{patch,j},\ \mathbf{E}_{patch,v})), \tag{3.8}$$

$$\mathbf{V}_{bj} = \mathbf{Proj}_V^{bj}(Concat(\mathbf{E}_{patch,j},\ \mathbf{E}_{patch,b})), \tag{3.9}$$

$$\mathbf{K}_{bj} = \mathbf{Proj}_K^{bj}(Concat(\mathbf{E}_{patch,j},\ \mathbf{E}_{patch,b})). \tag{3.10}$$

After the aforementioned procedures, we have harvested Query, Key, and Value for the these two branches. Then these two branches need to be aggregated. We introduce SoftMax Concentrated Aggregation (SCA), which is realized through the following equations to achieve aggregation between $\mathbf{V}_{jv}$ and $\mathbf{V}_{bj}$, $\mathbf{K}_{jv}$ and $\mathbf{K}_{bj}$, and $\mathbf{Q}_{jv}$ and $\mathbf{Q}_{bj}$:

$$\mathbf{V} = (SoftMax(\mathbf{V}_{jv})^T\mathbf{V}_{bj} + SoftMax(\mathbf{V}_{bj})^T\mathbf{V}_{jv})/2, \tag{3.11}$$

$$\mathbf{K} = (SoftMax(\mathbf{K}_{jv})^T\mathbf{K}_{bj} + SoftMax(\mathbf{K}_{bj})^T\mathbf{K}_{jv})/2, \tag{3.12}$$

$$\mathbf{Q} = (SoftMax(\mathbf{Q}_{jv})^T\mathbf{Q}_{bj} + SoftMax(\mathbf{Q}_{bj})^T\mathbf{Q}_{jv})/2. \tag{3.13}$$

After the SCA operation, we obtain the aggregated Query $\mathbf{Q}$, Key $\mathbf{K}$, and Value $\mathbf{V}$, which are fused together by $Att = SoftMax(\mathbf{Q}\mathbf{K}^T/\sqrt{s_k})\mathbf{V}$, where a scale factor s_k is used to avoid the negative influence brought by the dot product on the variance and Att denotes the calculated attention value. MF strategy enhances the feature learning by realizing a major modality agnostic attentional fusion, where the joints modality is chosen as major modality since the other two modalities are derived from joints modality. By fusing the information from the joints-velocity branch and the joints-bones branch regarding the three components for the self attention calculation, more useful cues can be better obtained.

Mixed Attention Fusion Mechanism (MAFM). The proposed MAFM, showcased in the bottom right corner of Fig. 3.2, is a critical component of our Trans4SOAR architecture, designed for advanced aggregation of multimodal inputs. It incorporates Layer Normalization (*LN*), averaged skip connections, and path dropout to enhance fusion efficacy and ensure model generalizability. The process begins with obtaining the attended embedding $\mathbf{E}_{att}$, aimed at unifying the strengths of each input stream, as shown in Eq. 3.14:

$$\mathbf{E}_{att} = MF(LN(\mathbf{E}_{patch,j}),\ LN(\mathbf{E}_{patch,v}),\ LN(\mathbf{E}_{patch,j})). \tag{3.14}$$

As depicted in Fig. 3.2, the original patch embeddings $\mathbf{E}_{patch,j}$, $\mathbf{E}_{patch,v}$, and $\mathbf{E}_{patch,b}$ are firstly averaged and then added with the path-dropped attended embedding $\mathbf{E}_{att}$ to harvest $\mathbf{E}_{asn}$, an embedding after averaging (*AVG*), as shown in Eq. 3.15:

$$\mathbf{E}_{asn} = AVG(\mathbf{E}_{patch,j}, \mathbf{E}_{patch,v}, \mathbf{E}_{patch,b}) + DP(\mathbf{E}_{att}), \tag{3.15}$$

Algorithm 2 An overview of the training pipeline with LSC loss.

Input: $\mathbf{S}$ – a batch in D_{train}; $labels_S$ denotes the label set for batch $\mathbf{S}$; $\mathbf{S}_p$ and $\mathbf{S}_n$ – positive and negative anchor; $f_\delta^{1\to N-1}$ and $f_\theta^{1\to N-1}$ – first N-1 transformer layers of main and auxiliary branches; f_δ^N and f_θ^N – the N-th (last) transformer layer for main and auxiliary branches; EMB – embedding layer; N_e – maximum training epochs; N_t – epoch threshold for the stage changing; $\mathbf{E}$ and $\mathbf{E}^*$ – embedding for main and auxiliary branches; PMB – prototypes memory bank; WarmUpAug and PrototypeAug – warm-up stage and prototype-based feature augmentation stage; L_p is initialized as empty list;

for all $epoch \in Range(N_e)$ **do**
  for all $\mathbf{S} \in D_{train}$ **do**
    $L_p = []$.
    if $epoch > N_t$ **then**
      for all l *in* $label_S$ **do** $Append(PMB[l]) \to L_p$
      end for
      $\boldsymbol{E}_P^* = Concat(L_p)$.
    end if
    if *BaseModel is not None* **then** $\mathbf{S} := BaseModel(\mathbf{S})$
    end if
    $\mathbf{E}_{patch} = PatchEmbeddingAndEncoding(\mathbf{S})$.
    $\mathbf{E}_{N-1} = f_\delta^{1\to N-1}(\mathbf{E}_{patch})$, $\mathbf{E}_{N-1}^* = f_\theta^{1\to N-1}(\mathbf{E}_{patch})$.
    if $epoch < N_t$ **then** $\mathbf{E}_{aug}^* = WarmUpAug(\mathbf{E}_{N-1}^*)$
    else $\mathbf{E}_{aug}^* = PrototypeAug(\mathbf{E}_{N-1}^*, \mathbf{E}_P^*)$
    end if
    $\mathbf{E} = EMB(f_\delta^N(\mathbf{E}_{N-1}))$, $\mathbf{E}^* = EMB(f_\theta^N(\mathbf{E}_{aug}^*))$.
    $\mathscr{L}_{TPL} = TripletMarginLoss(\mathbf{E}, \mathbf{E}_n, \mathbf{E}_p)$.
    $\mathscr{L}_{LSC} = ConsistencyLoss(\mathbf{E}, \mathbf{E}^*) \to$*LSC loss.*
    $\mathscr{L}_{CLS} = ClassificationLoss(Head(\mathbf{E}), label_S)$.
    $BackPropagation(WeightedSum(\mathscr{L}_{TPL}, \mathscr{L}_{CLS}, \mathscr{L}_{LSC}))$.
  end for
  if $epoch > N_t - 1$ **then**
    $CalculatePrototypes(D_{train}) \to Set(\mathbf{E}_{N-1}) \to PMB$.
  end if
end for

where DP indicates the path dropout operation. Then, the final fused embedding $\mathbf{E}_{mixed}$ is harvested through Eq. 3.16:

$$\mathbf{E}_{mixed} = DP(MLP(LN(\mathbf{E}_{asn}))) + \mathbf{E}_{asn}. \tag{3.16}$$

Subsequently, the composite embedding generated by this process is then inputted into a series of transformer blocks for further analysis.

3.1.4.3 Prototype-Based Latent Space Consistency Loss

In the pursuit of achieving discriminative and generalizable embedding learning in one-shot action recognition, we introduce the concept of Latent Space Consistency (LSC) loss. This loss function is designed to ensure that the embeddings derived from the primary branch of the model are

consistent with those obtained from an additional branch that utilizes prototype-based feature augmentation, as delineated in Alg. 2. The core objective of implementing LSC loss lies in bolstering the model's generalizability, compelling it to maintain stable embeddings despite perturbations introduced by feature-level augmentations. This approach is particularly effective in mitigating the impact of occlusions, as validated by our experimental results. Our methodology extends the principles of a recent feature augmentation strategy proposed in the realm of semi-supervised learning [97], further enhanced by the integration of a warm-up self-augmentation phase and architectural modifications, which collectively contribute to notable improvements in model accuracy and generalizability.

Estimation of the prototypes for different action categories. In our approach to feature-level augmentations within the auxiliary branch, we are inspired by FeatMatch [97], an innovative semi-supervised image classification technique that employs weighted combinations of category-specific prototypes to refine intermediate features. For each category $l_i \in C_{base}$ belonging to the dataset's richly annotated action classes, we calculate the category's prototype within the latent space. This is achieved by determining the centroid of all embeddings for a given action class within the training set, specifically utilizing the embeddings obtained just prior to the final transformer block in a sequence of N blocks. Distinct from the FeatMatch approach, which resorts to clustering in the context of semi-supervised learning due to the lack of labels, our method leverages the centroids of categories with label available during the training phase. These prototypes are dynamically updated at the conclusion of each training epoch and are systematically stored in what we term the *Prototype Memory Bank* (PMB), using a category-wise mean averaging process. Consequently, these category prototypes serve a crucial role in facilitating feature augmentations, thereby enabling more discriminative and generalizable embedding learning for SOAR performance improvement.

Prototype-based feature enhancement with self-augmentation warm-up technique. Using prototype-based augmentation within the realm of one-shot learning necessitates additional conceptual modifications. To adapt prototype-based augmentation for one-shot learning, a new approach is employed. Given that the prototypes $\mathbf{E}_P^*$ directly represent specific action categories from C_{base}, we initiate the process by applying SoftMax normalization across the prototype vector's channel dimension. Subsequently, this normalized vector is combined with the feature $\mathbf{E}_{N-1}^*$, derived from $\mathbf{E}_{N-1}^* = f_\theta^{1\to N-1}(\mathbf{E}_{patch})$, where $\mathbf{E}_{mixed}$ denotes the mixed fused patch embedding and f_θ^i represents the i-th transformer stage block in the auxiliary branch. This combination is then projected into an embedding space as $\mathbf{E}_{r,N-1}^* = g_\mu^2(\text{SoftMax}(\mathbf{E}_P^*) \cdot \mathbf{E}_{N-1}^*)$, with N indicating the total count of transformer stage blocks. Concurrently, $\mathbf{E}_{N-1}^*$ undergoes a separate projection as $\mathbf{E}_{l,N-1}^* = g_\mu^1(\mathbf{E}_{N-1}^*)$, and the attention weight $\mathbf{W}$ is calculated via $\mathbf{W} = \text{SoftMax}({\mathbf{E}_{l,N-1}^*}^T \mathbf{E}_{r,N-1}^*)$, setting the stage for aggregating information from $\mathbf{E}_P^*$ into $\mathbf{E}_{N-1}^*$ as depicted in Eq. 3.17:

$$\mathbf{E}^*_{agg,N-1} = g^3_\mu(Concat(\mathbf{W}\mathbf{E}^*_{r,N-1}, \mathbf{E}^*_{l,N-1})), \tag{3.17}$$

the final augmented embedding $\mathbf{E}^*_{aug}$ is then harvested by a residual connection with the original embedding $\mathbf{E}^*_{N-1}$ using $\mathbf{E}^*_{aug}$=$ReLU(\mathbf{E}^*_{N-1}+\mathbf{E}^*_{agg,N-1})$, where g^1_μ and g^2_μ denote two fully-connected (fc) layers without weight sharing, and g^3_μ denotes a stack of two fc layers with ReLU in between.

In our methodology, given that the prototypes correspond directly to known action categories from C_{base} (unlike the unsupervised clustering required in self-supervised tasks), the utilization of category centers during initial training epochs may lead to inaccuracies.

To mitigate this, we incorporate a *warm-up phase*, prioritizing self-augmentation over prototype-based augmentation until a certain degree of model convergence is achieved. Initially, the focus is on self-augmentation, utilizing the embedding $\mathbf{E}^*_{N-1}$ in place of the attended prototype representation. This phase is visually represented at the top of Fig. 3.3, contrasting with the prototype-based augmentation depicted at the bottom. Subsequently, the approach transitions to prototype-based augmentation, introducing zero prototypes for a predetermined period to facilitate decenterization, followed by the implementation of class-agnostic prototype augmentation at the feature level. The augmented embeddings $\mathbf{E}^*$ are derived through $\mathbf{E}^* = EMB(f^N_\theta(\mathbf{E}^*_{aug}))$, where EMB denotes the embedding generation layers based on a multi-layer perceptron. The main branch's final embedding $\mathbf{E}$ is computed as $\mathbf{E} = EMB(f^N_\delta(\mathbf{E}_{N-1}))$, with $\mathbf{E}_{N-1}$ being the output from the $N-1$ stages of transformer blocks, denoted as $f^{1\to N-1}_\delta(\mathbf{E}_{patch})$ for the main branch. Upon obtaining the embeddings from the main branch $\mathbf{E}$ and the auxiliary branch $\mathbf{E}^*$, the LSC loss is harvested as Eq. 3.18, where cos indicates cosine similarity.

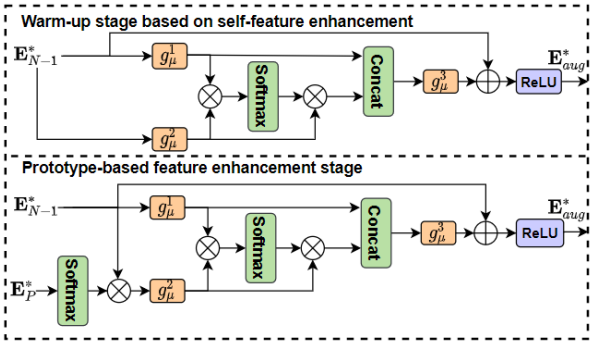


Figure 3.3: An overview of the self-augmentation and prototype-based augmentation mechanisms.

$$\mathscr{L}_{LSC} = 1 - \cos(\mathbf{E}, \mathbf{E}^*). \tag{3.18}$$

3.1.4.4 Deep Metric Learning and Classification Losses

Triplet margin loss. In the protocol we used, a triplet margin loss is utilized to enhance the discriminative ability of the learnt embeddings. This setting involves selecting a triplet of embeddings for each instance: $\mathbf{a}_i$, $\mathbf{p}_i$, and $\mathbf{n}_i$, representing the anchor, a positive anchor sharing the same class as the anchor, and a negative anchor from a different class, respectively. The objective of the triplet

margin loss is designed to minimize the distance between the current anchor and the positive anchor while simultaneously maximizing the distance between the current anchor and the negative anchor, as shown in Eq. 3.19:

$$\mathscr{L}_{TPL} = \sum_{i=1}^{N_B} max\{D(\mathbf{a}_i, \mathbf{n}_i) - D(\mathbf{a}_i, \mathbf{p}_i) + \sigma, 0\}/N_B, \tag{3.19}$$

where the specified margin is represented by σ, which is intended to delineate the minimum desired separation between the anchor-positive and anchor-negative pairs, while $D(\cdot)$ specifies the pairwise distance metric. For calculating the pairwise distance between the current anchor $\mathbf{a}$ and the negative anchor $\mathbf{n}$, the formula $D(\mathbf{a}, \mathbf{n}) = \|\mathbf{a} - \mathbf{n} + \varepsilon\|^2$ is utilized, where ε is maintained at a small value of $1e^{-6}$ to ensure numerical stability, and N_B is the batchsize.

Classification loss. To monitor the training procedure and ensure the discriminative ability of embeddings in the latent space, a cross-entropy loss function is utilized. This function calculates the difference between the model's predicted classifications $\mathbf{p}_i$ and the true labels $\mathbf{y}_i$ for every i-th sample in a given batch. The calculation of the cross-entropy loss is executed as Eq. 3.20,

$$\mathscr{L}_{CLS} = (\sum_{i=1}^{N_B} [-\mathbf{y}_i log(\mathbf{p}_i) + (1 - \mathbf{y}_i) log(1 - \mathbf{p}_i)])/N_B \tag{3.20}$$

3.1.5 Experiments

3.1.5.1 Dataset Introduction

Our research includes extensive experiments in the realm of SOAR across three challenging datasets: NTU-60 [161], NTU-120 [115], and Toyota Smart Home [47]. Adhering to the established SOAR protocol of NTU-120, we create the evaluation protoocls for both Toyota Smart Home and NTU-60 to align with the objectives of our study, focusing on learning from data-scarce setting. Furthermore, we pioneer the introduction of occluded SOAR benchmarks, extending these three foundational datasets. Specifically, the benchmarks for NTU-120, NTU-60, and Toyota Smart Home are structured to include 100/48/24 categories for data-rich training and 20/12/7 categories for data-scarce testing, each with one reference sample per unseen category.

3.1.5.2 Implementation Details

In the training of Trans4SOAR, a preparatory warm-up phase is designated with a threshold $N_t = 20$, followed by an additional span of 10 epochs dedicated to decenterization processes. Optimization of the model employs the AdamW algorithm [126], integrated with a Cosine Annealing

Table 3.1: Experiments for SOAR without occlusion on NTU-120.

Encoder	Accuracy	F1	Precision	Recall
Previously Published Approaches				
AN† [118]	41.0	-	-	-
FC† [118]	42.1	-	-	-
AP† [118]	42.9	-	-	-
APSR [118]	45.3	-	-	-
TCN-OneShot [157]	46.3	-	-	-
SL-DML [135]	50.9	-	-	-
Skeleton-DML [134]	54.2	-	-	-
CNN-based Encoder Optimized by DML				
SL-DML (AlexNet [95])	40.33	39.14	42.42	40.35
SL-DML (SqueezeNet [82])	42.55	40.52	41.88	42.51
SL-DML (ResNet18 [73])	49.19	47.54	49.80	49.23
Transformer-based Encoder Optimized with DML (Ours)				
SL-DML (CaiT [185])	47.86	47.53	50.06	47.94
SL-DML (ViT [52])	48.45	47.40	48.59	48.52
SL-DML (Twins [38])	49.00	48.04	49.30	49.06
SL-DML (ResT [223])	52.58	51.86	53.99	52.61
SL-DML (Swin [122])	53.13	52.09	53.48	53.16
SL-DML (LeViT [69])	53.19	52.22	53.85	53.29
Our Proposed and Extended Approaches (Ours)				
SL-DML (LeViT) + LSC	55.94	54.29	55.80	56.04
Trans4SOAR (Small)	56.27	**56.43**	**58.59**	56.32
Trans4SOAR (Base)	**57.05**	55.90	57.26	**57.12**

Scheduler over a course of 50 epochs. The model's operation utilizes a batch size of 32, facilitated by the computational capabilities of an Nvidia A100 GPU and developed within the PyTorch 1.8.0 framework to achieve optimal performance. An initial learning rate set at $3.5e^{-5}$ supports the balancing of three distinct losses: Triplet Margin Loss ($\sigma = 0.2$), Cross Entropy Loss, and LSC loss, with respective weights of 1.0, 0.4, and 0.1. The architecture of Trans4SOAR (Small) is characterized by a D_{Key} of 1, N_{head} values of $[2,2,2]$, H_{dep} values of $[2,4,4]$, and C_{dim} values of $[384,512,512]$, cumulating to a total of 23M parameters. In contrast, the Trans4SOAR variant is configured with a D_{Key} of 32, N_{head} values of $[6,9,12]$, H_{dep} of $[4,4,4]$, and C_{dim} values of $[384,512,768]$, accounting for a total of 43M parameters. Both models incorporate three principal transformer blocks, denoting a sophisticated approach to feature processing.

To preclude potential data leakage stemming from the augmentation process to the occlusion areas, occlusions are preemptively generated prior to the application of data augmentation techniques. This approach is consistently applied across all the occlusion types leveraged in our work.

Table 3.2: Experiments regarding REalistic synthesized occlusion (RE) and RAndom occlusion (RA) for SOAR.

Encoder	(a) With RE				(b) With RA			
	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
SL-DML [135]	39.82	37.85	39.32	39.86	42.53	42.24	44.79	42.56
Skeleton-DML [134]	49.21	46.82	48.10	49.18	35.15	32.59	34.29	35.22
SL-DML (LeViT [69])	44.22	42.29	44.20	44.31	35.00	33.24	41.45	35.10
SL-DML (Swin [122])	47.19	45.64	46.78	47.29	42.16	39.93	39.67	40.20
SL-DML (LeViT) + LSC	48.28	46.03	47.58	48.31	38.04	35.93	37.87	38.11
Trans4SOAR (Small)	51.64	**50.47**	52.36	51.70	**53.27**	51.33	53.80	**53.35**
Trans4SOAR (Base)	**52.35**	48.79	**52.87**	**52.43**	53.17	**52.89**	**54.50**	53.21

Table 3.3: Experiments on the NTU-60 for SOAR considering different occlusion scenarios.

Encoder	(a) Without Occlusion				(b) With RE				(c) With RA			
	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
Previously Published Approaches												
SL-DML [135]	54.82	54.31	56.72	54.65	36.90	35.86	36.59	37.05	45.28	43.13	45.00	45.42
Skeleton-DML [134]	55.54	50.88	53.13	51.24	42.66	40.90	41.50	42.82	60.43	59.66	61.37	60.54
Transformer-based Encoder Optimized by DML (Ours)												
SL-DML (Swin [122])	56.99	56.24	58.67	56.99	51.71	50.60	52.54	51.82	64.65	63.74	66.57	64.77
SL-DML (LeViT [69])	64.45	64.17	66.35	64.47	52.72	52.19	54.90	52.86	56.73	55.89	57.57	56.85
Our Extended and Evaluated Approached (Ours)												
SL-DML (LeViT) + LSC	67.67	67.87	68.74	67.67	53.79	52.76	54.18	53.88	60.78	58.75	59.97	60.90
Trans4SOAR (Small)	69.74	70.52	72.45	69.82	56.84	55.84	58.27	56.98	67.90	67.32	68.94	68.01
Trans4SOAR (Base)	**74.19**	**74.34**	**75.91**	**74.20**	**59.28**	**58.96**	**59.91**	**59.40**	**72.59**	**71.82**	**73.89**	**72.66**

3.1.5.3 Analyses for SOAR Without Occlusion

Performance analyses of different components. In our investigation, delineated within Table 3.1, we commence by empirically examining the enhancements attributed to the introduction of LSC loss, which is implemented via prototype-based feature augmentation alongside an auxiliary branch. We compare our approach with the baseline SL-DML [134], which employs a data preprocessing methodology congruent with our proposed strategy.

Our methodology incorporates the SL-DML framework, subsequently integrating an auxiliary branch and leveraging transformer-based multi-stream patch embedding level fusion architecture, notably LeViT [69]. This auxiliary branch is dedicated to facilitating attention-driven augmentations through feature-level prototypes. The LSC loss is calculated via the cosine similarity loss between embeddings produced by the main and auxiliary branches. We first delve deeper into the benefits brought by the leveraged transformer architecture. A new variant of SL-DML is introduced by us, where the original CNN architecture is replaced by the LeViT architecture, denoted as SL-DML (LeViT). We observe that, SL-DML (LeViT) outperforms its original CNN version, SL-DML (ResNet18), which signifies the adaptation of the SL-DML pipeline through the substitution of a conventional CNN with transformer architecture due to its promising long-term reasoning ability.

Table 3.4: Experiments on the Toyota Smart Home for SOAR considering different occlusion scenarios.

Encoder	(a) Without Occlusion				(b) With RE				(c) With RA			
	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.	Acc.	F1.	Prec.	Rec.
Previously Published Approaches												
SL-DML [135]	58.98	27.15	27.64	35.00	38.93	25.16	32.93	28.48	53.79	26.28	27.24	29.67
Skeleton-DML [134]	47.31	18.45	18.58	23.80	47.67	24.86	27.93	27.35	48.91	21.60	25.00	21.75
Transformer-based Encoder Optimized by DML (Ours)												
SL-DML (Swin [135])	58.76	28.83	29.17	32.34	35.43	18.48	23.24	23.80	65.50	29.20	30.78	29.69
SL-DML (LeViT [69])	62.22	31.98	37.56	35.16	38.48	22.58	27.66	24.62	61.96	26.42	28.52	29.20
Our Extended and Evaluated Approached (Ours)												
SL-DML (LeViT) + LSC	64.46	31.91	34.07	33.58	41.82	24.34	29.02	26.67	63.77	27.72	29.09	29.90
Trans4SOAR (Small)	66.87	28.08	31.47	34.63	55.12	**26.90**	29.41	30.69	68.47	28.86	29.56	**32.25**
Trans4SOAR (Base)	**70.22**	**33.96**	**37.81**	**35.33**	**60.15**	25.50	**33.12**	**31.86**	**68.91**	**29.27**	**34.15**	31.45

Table 3.5: Ablation study of LSC and MAFM used in the Trans4SOAR on NTU-60 without occlusion.

With LSC	Self-aug. wp	De-centerization	MAFM	Accuracy	F1	Precision	Recall
				64.45	64.17	66.35	64.47
✓	✓	✓		67.67	67.87	68.74	67.67
			✓	71.55	71.85	73.45	71.63
✓			✓	72.69	72.80	74.27	72.73
✓		✓	✓	73.09	73.39	74.54	73.14
✓	✓	✓	✓	**74.19**	**74.34**	**75.91**	**74.20**

Remarkably, the integration of LSC loss into the SOAR task, devoid of occlusion scenarios, yielded substantial improvements in accuracy. Specifically, SL-DML (LeViT) + LSC Loss manifested an enhancement in accuracy by 2.75% on NTU-120 (Table 3.1), 3.22% on NTU-60 (Table 3.3 (a)), and 2.24% on Toyota Smart Home (Table 3.4 (a)), in comparison to SL-DML (LeViT). This variant demonstrated superior efficacy relative to both SL-DML [135] and Skeleton-DML [134] in the context of SOAR tasks without occlusion.

The comparative analysis on the NTU-120 dataset [115] accentuates the superiority of our approach with the LSC loss, which eclipses the achievements of preceding methodologies—namely, SL-DML [135] by more than 5% and Skeleton-DML [134] by 1.74% (Table 3.1).

Furthermore, our detailed ablation studies concerning the pivotal components of the LSC loss, as presented in Table 3.5, underscore the significance of both the warm-up and de-centerization phases. These stages collectively contribute to a performance uplift of 1.5% when compared with the LSC loss absent of the warm-up stage.

Then, the combination of the LSC loss and the MAFM, mixing three streams of input at patch embedding level, further contributes a remarkable performance gain regarding the SOAR without occlusion compared with the existing state-of-the-art works [134]. On the NTU-120 [115], Trans4SOAR (Base) surpasses Skeleton-DML [134] and SL-DML [135] by 2.85% and 6.15% for accuracy while outperforming SL-DML (LeViT) + LSC by 1.11%, indicating an promising performance enhancement which is resultant by the superior discriminative ability of the learned embedding by

Table 3.6: A comparison to other encoder architectures.

Methods	Accuracy	F1	Recall	Precision
SL-DML (CTR-GCN[33])	43.92	41.38	45.21	43.89
SL-DML (STTR[152])	39.56	39.45	41.92	39.58
SL-DML (LeViT) + LSC	55.94	54.29	55.80	56.04
Trans4SOAR (Small)	56.27	**56.43**	**58.59**	56.32
Trans4SOAR (Base)	**57.05**	55.90	57.26	**57.12**

Table 3.7: Experiments without occlusion on NTU-120 under Gaussian noise disruption.

Gaussian Noise	$\sigma = 0.1, \mu = 0$				$\sigma = 0.05, \mu = 0$			
Encoder	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
SL-DML [135]	21.42	11.83	8.50	21.71	21.76	12.23	8.70	21.86
SL-DML (LeViT)	22.31	12.32	8.79	22.40	21.97	12.82	9.69	22.07
SL-DML (LeViT) + LSC	52.54	51.16	51.61	52.65	51.91	50.08	51.67	52.01
Trans4SOAR	**53.09**	**51.89**	**53.05**	**53.15**	**54.74**	**54.65**	**56.33**	**54.83**

MAFM and LSC loss. We also conduct experiments to showcase the individual performance gain brought by LSC and MAFM in Table 3.5. Furthermore, consistent improvements are achieved by Trans4SOAR in the other two datasets, *e.g.*, NTU-60 in Table 3.3 (a) and Toyota Smart Home [47] in Table 3.4 (a) for the SOAR without occlusion. The NTU-60 [161] has less training categories than the NTU-120 [115], thus, it is used to evaluate the generalizability of the leveraged models under the SOAR challenge with less a prior knowledge. In Table 3.3 (a), our Trans4SOAR (Base) surpasses SL-DML [135] and Skeleton-DML [134] by 19.37% and 18.65% for accuracy, indicating that strong capability for harvesting discriminative and generalizable embeddings of our approach under the scenario with less a priori knowledge. Furthermore, the Toyota Smart Home [47] contains 2D skeleton data in image coordinate format, delivering a valuable data format to explore the SOAR task. In Table 3.4 (a), our Trans4SOAR (Base) shows the best performance over all the previous approaches with large margin. Observing the other three metrics, *i.e.*, F1-score, precision and recall, since the first two datasets have balanced distributed samples for different categories, theses three terms do not have large difference compared with the accuracy. However, since the action categories on the Toyota Smart Home [47] is not equal distributed, these three terms are able to showcase whether the true prediction is balanced distributed in the test set or not. Our Trans4SOAR surpasses all the approaches in terms of all metrics on the investigated datasets. In order to ablate the effect of different model scales, we construct Trans4SOAR (Small) with only 23M parameters which pursues both light model structure and high accuracy, and achieves second best performance, showcasing that the LSC loss and MAFM are helpful for learning discriminative features via different model variants. We also conduct experiments in Table 3.6 to compare with graph convolutional approach [33] and skeleton transformer approach [152], however the performance of these two encoder architectures for the SOAR task even without occlusion is not satisfied compared with Trans4SOAR and SL-DML (LeViT), since by using the image-wise encoding and the vision transformer architecture, more in-

formative cues with finer granularity can be captured by the deep learning model, which is very essential in data-scarce scenario, while GCN architectures work directly on sparse data and do not preserve this capability.

Tolerance to noisy inputs. The integrity of skeleton data can be compromised by numerous factors, including sensor inaccuracies and occlusions. An observed disparity in performance between Trans4SOAR and conventional DML methodologies, particularly noted on the Toyota Smart Home dataset [47], which presents a more challenging environment than the comparatively controlled NTU-60 and NTU-120 datasets, suggests a potential robustness of Trans4SOAR against imperfect inputs. To substantiate this hypothesis, we subjected the model to inputs corrupted by varying degrees of Gaussian noise, the results of which are shown in Table 3.7, illustrating Trans4SOAR's promising resilience and generalizability.

Contrary to the deteriorating accuracy observed in conventional DML-based frameworks, the application of LSC loss within SL-DML (LeViT) exhibits notable robustness against degraded input quality, thereby underscoring the efficacy of LSC loss in mitigating the negative effects of Gaussian noise. Specifically, the performance of SL-DML (LeViT) with LSC loss decreases marginally from 55.94% in the presence of clean data to 51.91% when subjected to Gaussian noise with σ=0.05. This degradation is significantly less severe compared to that of SL-DML (LeViT) without LSC loss, which plummets from 53.19% to 21.97% under identical conditions. This phenomenon is attributed to the sophisticated feature-level augmentations executed within the auxiliary branch during the formulation of the LSC loss, which inherently enables the model to maintain performance consistency despite variations in the embedding caused by noise.

Moreover, the Trans4SOAR (Base) model unequivocally outperforms all other baseline methodologies, achieving accuracies of 54.74% and 53.09% against Gaussian noise with σ=0.05 and σ=0.1, respectively. This resilience to noisy inputs is ascribed to the intricate augmentations learned at the feature level through the auxiliary branch. This advantage of LSC loss on dealing with imperfect data is expected to be helpful also on diverse occlusion scenarios. Next, we will conduct the analysis for SOAR on diverse occlusions.

3.1.5.4 Analyses for REalistic Synthesized Occlusion (RE)

In our research, we embark on an exploratory journey to examine the generalizablity of various models under the conditions of realistic synthesized occlusion across three distinct datasets: NTU-120 [115], NTU-60 [161], and Toyota Smart Home [47]. The results of these investigations are systematically presented in Table 3.2 (a), Table 3.3 (b), and Table 3.4 (b), wherein the SNR spans from 0.05 to 0.2, with occlusion implemented on the reference and test sets.

Our initial observations reveal a universal decline in the performance metrics across all evalu-

Table 3.8: Experiments with different realistic synthesized occlusion ratio on the NTU-60.

Model	RE_Range	Accuracy	F1	Precision	Recall
SL-DML [135]	0.05-0.2	36.90	35.86	36.59	37.05
Skeleton-DML [134]		35.15	32.59	34.29	35.22
SL-DML (LeViT [69])		52.72	52.19	54.90	52.86
SL-DML (LeViT) + LSC		53.79	52.76	54.18	53.88
Trans4SOAR (Small)		56.84	55.84	58.27	56.98
Trans4SOAR (Base)		**59.28**	**58.96**	**59.91**	**59.40**
SL-DML [135]	0.05-0.35	39.26	38.71	39.59	39.43
Skeleton-DML [134]		38.52	38.74	39.23	38.64
SL-DML (LeViT [69])		53.17	52.52	54.16	53.34
SL-DML (LeViT) + LSC		53.58	52.75	54.07	53.77
Trans4SOAR (Small)		**61.69**	**61.60**	**64.01**	**61.81**
Trans4SOAR (Base)		58.27	56.63	58.81	58.40
SL-DML [135]	0.05-0.5	34.89	32.63	31.85	35.07
Skeleton-DML [134]		42.83	42.33	42.46	42.93
SL-DML (LeViT [69])		54.84	54.07	57.06	54.99
SL-DML (LeViT) + LSC		55.07	55.01	57.56	55.21
Trans4SOAR (Small)		**59.59**	**59.21**	59.49	**59.70**
Trans4SOAR (Base)		57.52	57.21	**59.61**	57.64

ated methods under the SOAR with RE benchmarks, as opposed to their performance in scenarios devoid of occlusion. This delineates the elevated complexity and challenge posed by the RE task in the realm of discriminative representation learning on the perspective of one-shot skeleton-based human action recognition. Particularly, in Table 3.2 (a), the Trans4SOAR (Base) model shows highest performances, manifesting accuracy of 52.35%, alongside F1-score, precision, and recall of 48.79%, 52.87%, and 52.43%, respectively. These outcomes underscore a balanced performance across the diverse classes within the NTU-120 dataset [115], highlighting the efficacy of the Trans4SOAR (Base) model in navigating the challenges inherent to the RE occlusion.

In our detailed examination, the Trans4SOAR (Small) variant emerges as a strong contender, securing the second-highest performance across all metrics within the NTU-120 with RE benchmark, achieving an accuracy of 51.64%. It is noteworthy that the SL-DML (LeViT) configuration exhibits diminished effectiveness across all datasets subjected to the SOAR with RE evaluation. Specifically, within the NTU-120 [115] dataset under RE conditions, SL-DML (LeViT) records a lower accuracy of 44.22%, trailing behind Skeleton-DML [134], which attains an accuracy of 49.21%. Nonetheless, the variant of SL-DML (LeViT) with LSC loss enhances the accuracy to 48.28%, suggesting that while the LeViT architecture alone struggles with RE challenges, the incorporation of LSC loss mitigates its limitations.

The utilization of MAFM within our Trans4SOAR (Base) formulation marks a significant advancement, exhibiting a superior accuracy of 52.35%. This finding underscores the module's efficacy in counteracting the disturbances introduced by RE through the integration of triplet stream encoding and the proposed mixed fusion mechanism. The results underscore the critical role of MAFM in alleviating the negative effect brought by the disruptions caused by RE, achieved by amalgamating three distinct skeleton encoding formats that inherently provide de-occlusion cues.

Further experiments conducted on the NTU-60 [161] and Toyota Smart Home [47] datasets fur-

Table 3.9: Experiments under different random occlusion ratios on the NTU-60.

Model	RA_ratio	Accuracy	F1	Precision	Recall
SL-MDL [135]	0.1	45.28	43.13	45.00	45.42
Skeleton-DML [134]		60.43	59.66	61.37	60.54
SL-DML (LeViT [69])		56.73	55.89	57.75	56.85
SL-DML (LeViT) + LSC		60.78	58.75	59.97	60.90
Trans4SOAR (Small)		69.74	70.52	72.45	69.82
Trans4SOAR (Base)		**72.59**	**71.82**	**73.89**	**72.66**
SL-DML [135]	0.3	46.39	42.82	46.69	46.54
Skeleton-DML [134]		58.93	56.07	58.45	59.05
SL-DML (LeViT [69])		46.32	43.78	43.94	46.40
SL-DML (LeViT) + LSC		47.82	45.02	48.41	47.91
Trans4SOAR (Small)		66.57	66.26	67.94	66.65
Trans4SOAR (Base)		**72.39**	**72.81**	**74.68**	**72.43**
SL-DML [135]	0.5	43.44	38.46	41.30	43.57
Skeleton-DML [134]		44.69	41.89	45.74	44.79
SL-DML (LeViT [69])		35.77	32.56	36.22	35.94
SL-DML (LeViT) + LSC		40.53	37.38	38.33	40.59
Trans4SOAR (Small)		52.92	50.78	55.13	53.02
Trans4SOAR (Base)		**54.82**	**55.01**	**58.01**	**54.93**

ther validate the superiority of Trans4SOAR (Base) over both Skeleton-DML [134] and SL-DML [135]. Notably, on the NTU-60 dataset, Trans4SOAR (Base) surpasses these models by 16.62% and 22.38% in accuracy, respectively, and on the Toyota Smart Home dataset by 12.48% and 21.22%. Moreover, the Trans4SOAR (Small) variant also exhibits competitive performance metrics. Further experiments exploring various SNR ratio ranges for SOAR with RE, as detailed in Table 3.8, reveal Trans4SOAR's consistent and promising performance, with accuracy exceeding 56% across both Trans4SOAR (Base) and Trans4SOAR (Small) variants for SNR ranges of $0.05-0.2$, $0.05-0.35$, and $0.05-0.5$ on the NTU-60 [161] dataset.

3.1.5.5 Analyses Regarding Random Occlusion (RA)

In this sub section, the effect brought by random occlusion on SOAR task and the corresponding model performances of the baselines and our proposed approach will be introduced. This form of occlusion, characterized by its unpredictability, presents a another challenge in the field of skeleton-based action recognition, as detailed in the corresponding results sections across Table 3.2 (b), Table 3.3 (c), and Table 3.4 (c), under a specified SNR of 0.1 and with the absence of occlusion in the reference set.

Our best model, Trans4SOAR (Base), consistently outperforms existing methods by large margins. Notably, it surpasses SL-DML [135] and Skeleton-DML [134] by 10.64% and 18.02%, respectively, on the NTU-120 dataset. The analysis uncovers that the performance under RA of Skeleton-DML is inferior to that of SL-DML, with the results reversing under recognition with RE, indicating a pervasive lack of resilience among existing frameworks to diverse occlusion scenarios. Contrastingly, Trans4SOAR adeptly addresses this challenge, showcasing superior performance across various occlusion types, with a particular emphasis on RE. This capability is pivotal for the advancement of discriminative representation learning in occluded environments. Central to this achievement is

the MAFM, which demonstrates exceptional efficacy in navigating occlusion challenges by processing inputs through a tripartite stream of skeleton patch embeddings. This innovative approach positions MAFM as the preeminent fusion architecture for confronting the primary occlusions encountered, a comparison further elucidated in Table 3.11. The strategic encoding across three streams, capturing both of the temporal and spatial discrepancies from multi-modal perspectives, facilitates a nuanced understanding of the occluded regions from diverse vantage points to harvest a more generalizable and adaptable human motion reasoning.

In addressing RA across both NTU-60 [161] and Toyota Smart Home [47], both Trans4SOAR (Base) and (Small) variants demonstrate leading performance metrics, underscoring the robustness and versatility of our models. Further explorations into the effects of varying SNR levels, *i.e.*, 0.1, 0.2, and 0.3, under RA perturbation, as delineated in Table 3.9, reveal that Trans4SOAR (erroneously referred to as Trans4DARC) consistently outperforms competing methodologies. Notably, at SNR levels of 0.1 and 0.3, Trans4SOAR (Base) achieves accuracy of 72.59% and 72.39%, respectively, significantly outpacing Skeleton-DML's 60.43% and 58.93%. However, at an SNR of 0.5, the efficacy of Trans4SOAR (Base) experiences a decline, registering an accuracy of 54.82%, yet it still manages to maintain a commendable lead of 10.13% over the most outperforming baseline. These findings not only highlight the generalizability of Trans4SOAR against occlusion but also underscore its potential as a valuable method for future research in the domain of action recognition under occluded conditions.

3.1.5.6 Analyses for Occlusion on Reference Samples

In our ablation study presented within Table 3.10, we conduct an exploration to discern the effects of various occlusions on the reference set of the NTU-60 dataset [161]. The presence of occlusion within this context is quantified by an indicator termed OCCVal, where 'T' signifies the inclusion of occlusion and 'F' denotes its absence. To ensure a balanced comparison, a SNR of 0.1 is maintained for RA, while for RE, the SNR spans from 0.05 to 0.2, thus averaging a comparable SNR across both conditions. Our analysis reveals notable fluctuations in performance metrics for existing methodologies such as SL-DML [135] and Skeleton-DML [134], with observed absolute performance variances in accuracy amounting to 3.46% and 11.13% for RA, and 2.61% and 1.63% for RE, respectively. This variability underscores the differential impact of occlusions on the models' performance, highlighting the critical need for models to exhibit minimal performance deviation across varying occlusion states (OCCVal settings).

Contrastingly, Trans4SOAR demonstrates superior generalizable performance on diverse occlusion settings, exhibiting an absolute performance fluctuation of merely 1.00% for RA and 0.80% for RE. This stability, especially in the face of occlusion within the reference set, is indicative of

Table 3.10: Experiments for reference w/ or w/o occlusions on NTU-60.

Model	OCC	OCCVal	Accuracy	F1	Precision	Recall
SL-MDL [135]	RA	T	48.74	46.46	47.45	48.88
Skeleton-DML [134]			49.30	48.57	49.62	49.45
SL-DML (LeViT [69])			53.47	52.35	54.94	53.63
SL-DML (LeViT) + LSC			53.57	53.73	56.55	53.72
Trans4SOAR (Small)			**72.16**	**72.42**	73.67	**72.23**
Trans4SOAR (Base)			71.59	72.22	**73.95**	71.67
SL-DML [135]	RA	F	45.28	43.13	45.00	45.42
Skeleton-DML [134]			60.43	59.66	61.37	60.54
SL-DML (LeViT [69])			56.73	55.89	57.57	56.85
SL-DML (LeViT) + LSC			60.78	58.75	59.97	60.90
Trans4SOAR (Small)			67.90	67.32	68.94	68.01
Trans4SOAR (Base)			**72.59**	**71.82**	**73.89**	**72.66**
SL-DML [135]	RE	T	36.90	35.86	36.59	37.05
Skeleton-DML [134]			42.66	40.90	41.50	42.82
SL-DML (LeViT [69])			52.72	52.19	54.90	52.86
SL-DML (LeViT) + LSC			53.79	52.76	54.18	53.88
Trans4SOAR (Small)			56.84	55.84	58.27	56.98
Trans4SOAR (Base)			**59.28**	**58.96**	**59.91**	**59.40**
SL-DML [135]	RE	F	39.51	39.64	40.82	39.64
Skeleton-DML [134]			44.29	43.10	44.26	44.46
SL-DML (LeViT [69])			55.12	55.22	57.51	55.26
SL-DML (LeViT) + LSC			55.07	55.01	57.56	55.21
Trans4SOAR (Small)			54.37	52.97	55.08	54.38
Trans4SOAR (Base)			**58.48**	**57.10**	**57.75**	**58.61**

Trans4SOAR's generalizability and its adeptness in managing occlusion-induced variability. Such a trait is immensely valuable for practical applications, where consistency in performance despite the presence of occlusions in the reference samples is essential.

Furthermore, the investigation highlights a discernible trend where both SL-DML and Skeleton-DML manifest inferior performance under RE conditions compared to RA when OCCVal is set to 'F'. This observation indicates that RE scenarios pose a more formidable challenge than RA, thereby necessitating advanced modeling techniques capable of navigating the complexities introduced by occlusion with greater efficacy.

Through this ablation, it becomes evident that Trans4SOAR not only surpasses existing models in terms of stability across different occlusion conditions but also sets a new state-of-the art for conventional SOAR task. This insight into the varying impacts of occlusion types, coupled with the demonstrated superiority of Trans4SOAR, underscores the importance of reasonable model design in the face of environmental variabilities such as occlusions on generalizable challenges of deep learning model.

3.1.5.7 Analyses for Ablation of Fusion Mechanisms

We further provide a series of comparative analyses against other existing fusion methods, as detailed in Table 3.11, to illustrate the significance of our proposed fusion method. This ablation

Table 3.11: Experiments for different fusion techniques on NTU-60 under different occlusion scenarios.

Fusion Method	OCC	Accuracy	F1	Precision	Recall
Single (Joints)	RE	53.79	52.76	54.18	53.88
Single (Bones)	RE	54.22	53.73	54.86	54.33
Single (Velocities)	RE	56.93	56.10	57.97	57.03
Addition	RE	56.37	54.48	55.68	56.51
Multiplication	RE	53.35	51.91	53.69	53.50
Concatenation	RE	58.61	57.21	57.63	58.73
Late Fusion	RE	56.93	56.10	57.97	57.03
MAFM	RE	**59.28**	**58.96**	**59.91**	**59.40**
Single (Joints)	RA	60.78	58.75	59.97	60.90
Single (Bones)	RA	55.15	53.56	56.63	54.16
Single (Velocities)	RA	33.15	30.54	29.67	33.82
Addition	RA	65.09	65.03	66.36	65.18
Multiplication	RA	67.54	67.51	68.65	67.63
Concatenation	RA	68.05	68.54	70.90	68.13
Late Fusion	RA	71.16	71.58	73.16	71.22
MAFM	RA	**72.59**	**71.82**	**73.89**	**72.66**
Single (Joints)	N	67.67	67.87	68.74	67.67
Single (Bones)	N	61.45	61.44	63.50	61.57
Single (Velocities)	N	49.74	50.08	51.31	49.89
Addition	N	67.05	66.88	68.09	67.12
Multiplication	N	64.63	65.05	66.34	64.75
Concatenation	N	67.75	67.79	69.56	67.86
Late Fusion	N	57.15	56.52	57.57	57.26
MAFM	N	**74.19**	**74.34**	**75.91**	**74.20**

study was conducted under specific conditions, including a SNR of 0.1 for RA and a SNR range from 0.05 to 0.2 for RE, to ensure a more generalizable assessment across varied occlusion scenarios.

Among the fusion strategies explored, late fusion emerges as a commonly adopted technique, characterized by its operation at the decision level through the integration of outputs derived from different modality branches. Notwithstanding its prevalence, a notable drawback of late fusion is its substantial demand on model size, effectively tripling the model size to 113M parameters, in stark contrast to alternative approaches maintained at a more economical model scale of approximately 40M parameters.

Prompted by the necessity for a more resource-efficient yet performance-optimized fusion mechanism, we advocate for a fusion strategy at the patch embedding level. This approach endeavors to harmonize the dual objectives of model performance and size efficiency. The variants for patch embedding level fusion encompass addition, multiplication, and concatenation operations, executed immediately subsequent to the generation of patch embeddings for the tripartite modality streams. In comparison, late fusion, albeit incorporating aggregation on the decision perspective, is encumbered by a significantly larger model size.

Our findings highlight MAFM's superior performance relative to both the conventional patch-embedding level fusion baselines and the more resource-intensive late fusion approach across vari-

Table 3.12: Experiments for random temporal and spatial occlusion on NTU-60 dataset.

Model	(a) Random temporal occlusion				(b) Random spatial occlusion			
	Acc.	F1.	Prec.	Rec.	Acc.	F1.	Prec.	Rec.
Experiments on NTU-120 with random temporal occlusion.								
SL-DML [135]	38.15	34.87	38.51	38.11	38.15	35.26	36.76	38.13
Skeleton-DML [134]	27.20	24.43	26.75	27.12	27.93	25.91	28.24	27.93
Trans4SOAR (Small)	51.60	50.73	52.65	50.99	46.99	46.24	49.71	47.07
Trans4SOAR (Base)	**54.11**	**52.93**	**53.85**	**54.21**	**49.43**	**49.08**	**51.35**	**49.48**
Experiments on NTU-60 with random temporal occlusion.								
SL-DML [135]	58.68	58.46	60.20	58.72	52.48	50.59	54.06	52.65
Skeleton-DML [134]	51.81	50.50	53.06	51.95	46.38	43.68	45.94	46.54
Trans4SOAR (Small)	71.45	71.32	72.94	71.51	68.94	69.61	**71.84**	69.01
Trans4SOAR (Base)	**75.01**	**74.75**	**75.76**	**75.06**	**69.08**	**69.18**	71.19	**69.14**
Experiments on Toyota Smart Home with random temporal occlusion.								
SL-DML [135]	53.36	22.97	28.17	24.58	60.36	20.10	24.52	20.89
Skeleton-DML [134]	53.65	23.90	31.54	25.33	41.95	26.36	32.83	27.97
Trans4SOAR (Small)	63.66	29.90	31.76	34.06	**66.76**	31.76	33.14	**35.66**
Trans4SOAR (Base)	**68.48**	**31.11**	**33.81**	**33.80**	64.49	**32.43**	**35.80**	34.29

ous occlusions on the NTU-60 dataset [161], including scenarios with no occlusion (N), RE, and RA. Remarkably, Trans4SOAR integrated with MAFM not only outperforms late fusion by 2.35%, 1.43%, and 17.04% across RE, RA, and N conditions, respectively, but also achieves this with a significantly reduced model size, thus enhancing both inference and training efficiencies.

Moreover, when compared with the most effective patch embedding level fusion technique among the other fusion approaches, Trans4SOAR with MAFM demonstrates a noteworthy advantage, exceeding performance by 0.67%, 4.54%, and 6.44% for RE, RA, and N respectively. This comprehensive comparative analysis underscores the MAFM's pivotal role in achieving a delicate balance between model efficiency and performance, particularly in the challenging context of skeleton-based one-shot action recognition under various occlusion scenarios.

3.1.5.8 Analyses for Random Temporal and Spatial Occlusions

In our ablation aimed at evaluating the efficacy of various models under other different occlusion scenarios, we specifically adopted random *temporal* and *spatial* occlusions as outlined in existing literature [35]. Recognizing the potential interest within the research community concerning the impact of these occlusions, each characterized by their unique temporal and spatial dimensions, we undertook a series of experiments across three distinct datasets. Utilizing the most effective

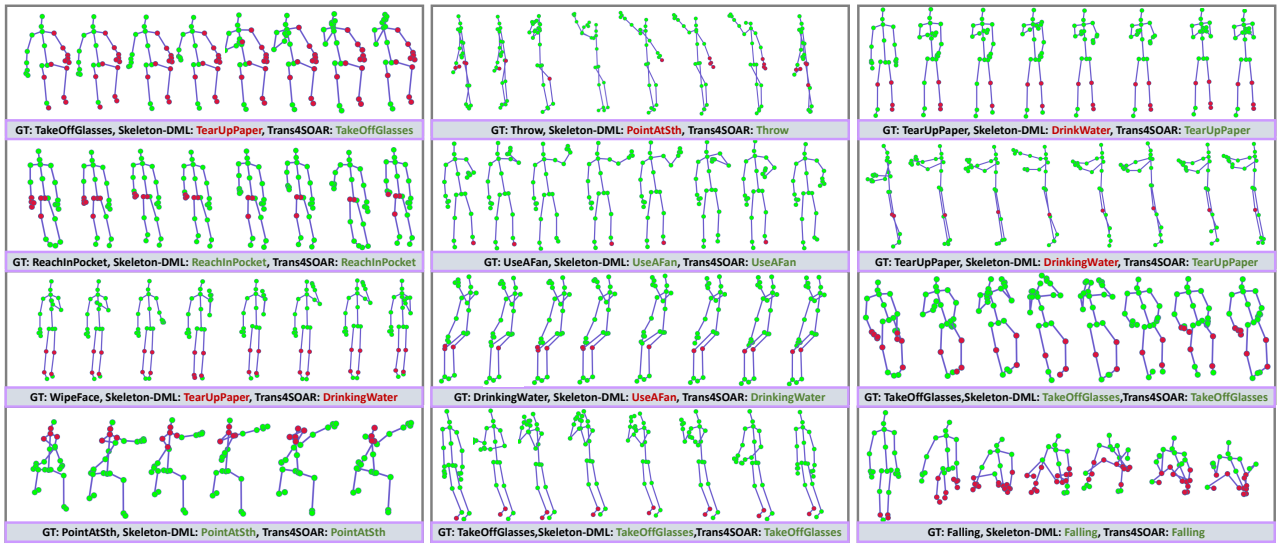


Figure 3.4: An overview of the qualitative experimental results on NTU-60.

methodologies identified in our study, these experiments are systematically cataloged in Table 3.12 (a) and Table 3.12 (b). Herein, we predefined the parameters of occlusion to include a specific number of occluded frames and joints (set at 10 and 5) to facilitate a controlled evaluation environment.

Contrary to the complexities encountered in RE scenarios, the occlusions defined by explicit frame and joint parameters exhibit a reduced level of occlusion reality. Nonetheless, it is crucial to underscore the superior performance of our proposed Trans4SOAR models, both Base and Small variants, which consistently outperformed existing models across all datasets under diverse occlusion types. This noteworthy achievement not only demonstrates the generalizability and robustness of Trans4SOAR against other specific random occlusion challenges but also validates its exceptional efficiency, even when faced with occlusions meticulously delineated by predefined criteria such as the number of occluded frames.

Such empirical evidence accentuates the significance of our model in advancing the field of SOAR, particularly in navigating the intricate landscape of occlusion-induced challenges.

3.1.5.9 Analysis for Qualitative and TSNE Experimental Results

In our study, we extend the analysis beyond quantitative metrics to include a qualitative examination of model performance under occlusion scenarios, specifically RE on the NTU-60 dataset [161]. This qualitative assessment, illustrated in Fig. 3.4, focuses on the efficacy of Trans4SOAR in comparison to Skeleton-DML [134]. The occluded body joints are visually represented as red dots within the figure, facilitating a direct comparison of model predictions in the presence of occlusions. Notably, Trans4SOAR demonstrates superior performance, achieving accurate predictions in three out of four

sample scenarios against Skeleton-DML's two. A critical observation from this analysis is the impact of occlusions on joints pivotal to specific actions, such as arm and hand joints for the action "Take Off Glasses". While Skeleton-DML misinterprets such actions under occlusion, Trans4SOAR consistently delivers true predictions, underscoring its high generalizablility. Nonetheless, the challenge of distinguishing between actions with high visual similarity, such as "Wipe Face" and "Drinking Water", persists as an area for further exploration, with Trans4SOAR occasionally misclassified actions despite offering predictions closer to the true action than those of Skeleton-DML.

Additionally, a TSNE analysis, as depicted in Fig. 3.5, offers a comparative visualization of feature embeddings between Skeleton-DML and Trans4SOAR (Base) under both RA and RE occlusions. This analysis reveals that Trans4SOAR (Base) achieves more discriminative class boundaries in the latent space, indicating a more effective discrimination of action classes despite occlusion. The comparison also shows that Trans4SOAR (Base) maintains consistency in the distribution of embeddings across different occlusion types, in contrast to the more variable embeddings produced by Skeleton-DML. This stability is indicative of Trans4SOAR's generalizability and adaptability in face of varying occlusion scenarios, underscoring the model's advanced capability in learning discriminative features.

3.1.5.10 Analyses for the Model Efficiency.

In our evaluation of the efficiency and performance of various models for the task of SOAR on the non-occluded NTU-120 dataset, a comprehensive comparison was undertaken, the results of which are encapsulated in Table 3.13. This analysis specifically targets the accuracy, the total count of model parameters, and the computational expense measured in GFLOPS during the inference phase. It is pertinent to note that for the initial quartet of models listed under the category of *Previously Published Approaches*, the specifics regarding the number of parameters and GFLOPS remain undisclosed.

A discerning examination reveals that, while CNN-based and GCN-based methodologies tend to exhibit smaller model size in terms of parameter count and GFLOPS, they generally fall short of achieving satisfactory performance levels on the SOAR task. This observation underscores a performance-parameter trade-off inherent to these approaches.

Conversely, the deployment of Visual Transformer-based architectures presents an intriguing paradigm; their enhanced performance on the SOAR task is not invariably linked to an increased model size. For instance, although SL-DML (CaiT) preserves the largest model size and SL-DML (ResT) preserves the highest GFLOPS, neither deliver better performances when compared with SL-DML (LeViT), which delivers an accuracy of 53.19%, alongside a parameter count of 38.9M and 30.4 GFLOPS. This elucidates the notion that superior performance is not solely predicated on the

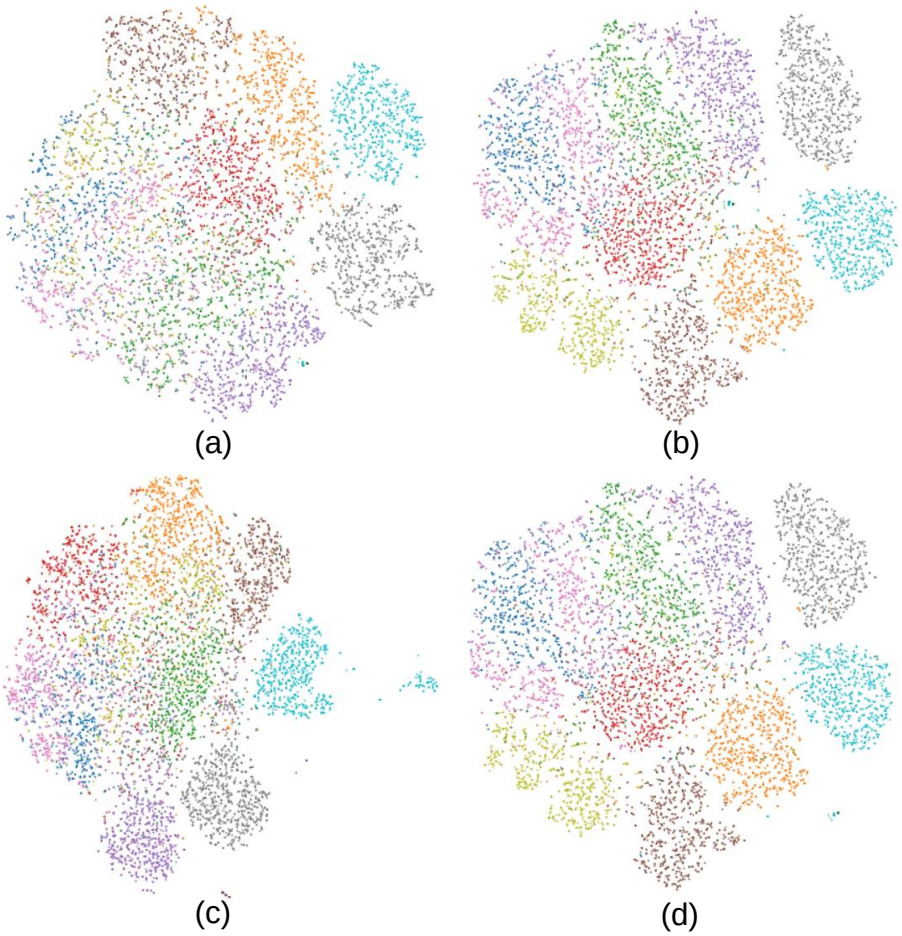


Figure 3.5: TSNE visualizations for (a) Skeleton-DML under RA, (b) Trans4SOAR (Base) under RA, (c) Skeleton-DML under RE and (d) Trans4SOAR (Base) under RE on NTU-60 [161].

increment of the model's scale.

Trans4SOAR takes two advantages together, balancing a competitive parameter count and GFLOPS with superior SOAR perfomance for w/ occlusion and w/o occlusion scenarios. Notably, Trans4SOAR (Small) achieves an impressive 56.27% accuracy, underpinned by a modest architectural size consisting of 23.1M parameters and 34.1 GFLOPS. Given the multi-modality nature of our model, a moderate escalation in both parameters and computational load is anticipated.

Trans4SOAR (Base) achieves efficiency multi-modal fusion, evidencing a reduction of over 70M in parameters relative to the late fusion methodology, while concurrently elevating SOAR performance. This achievement illustrates the preeminence of Trans4SOAR within the multi-modality fusion landscape, affirming its stature as a model that judiciously harmonizes computational efficiency with superior action recognition capabilities.

Table 3.13: The comparison in terms of accuracy, the number of parameters (#Params), and GFLOPs on NTU-120 without occlusion.

Encoder	Accuracy	#Params	GFLOPS
Previously Published Approaches			
AN† [118]	41.0	-	-
FC† [118]	42.1	-	-
AP† [118]	42.9	-	-
APSR [118]	45.3	-	-
TCN-OneShot [157]	46.3	3.5M	8.5
SL-DML [135]	50.9	11.2M	23.8
Skeleton-DML [134]	54.2	11.2M	23.8
CNN-based Encoder Optimized by DML			
SL-DML (AlexNet [95])	40.33	57.1M	9.2
SL-DML (SqueezeNet [82])	42.55	0.7M	9.7
SL-DML (ResNet18 [73])	49.19	11.2M	23.8
GCN-based Encoder Optimized with DML (Ours)			
SL-DML (CTR-GCN [33])	43.92	1.6M	9.2
SL-DML (STTR [152])	39.56	7.0M	37.4
Transformer-based Encoder Optimized with DML (Ours)			
SL-DML (CaiT [185])	47.86	120.8M	53.9
SL-DML (ViT [52])	48.45	53.6M	27.1
SL-DML (Twins [38])	49.00	25.2M	75.1
SL-DML (ResT [223])	52.58	57.8M	61.3
SL-DML (Swin [122])	53.13	87.3M	29.3
SL-DML (LeViT [69])	53.19	38.9M	30.4
Our Proposed and Extended Approaches (Ours)			
SL-DML (LeViT) + LSC	55.94	38.9M	30.4
Trans4SOAR (Small)	56.27	23.1M	34.1
Trans4SOAR (Base)	**57.05**	43.8M	47.9

3.1.6 Discussion

Our findings highlight the critical impact of occlusions on the accuracy and reliability of skeleton-based one-shot action recognition tasks. We discovered that both random occlusions and the proposed realistic synthesized occlusion, such as those caused by everyday objects, significantly impair the performance of current skeleton-based one-shot action recognition models. The proposed realistic synthesized occlusion is proved to be more challenging compared to random occlusion, as it considers the geometric continuity of real-world objects, resulting in more complex and realistic occlusion scenarios. We find that most of the existing skeleton-based one-shot action recognition approaches are facing with more performance decay when suffering from the realistic synthesized occlusion compared with the random occlusion. Occlusion is particularly detrimental to skeleton-based one-shot action recognition because it leads to missing or corrupted joint information, which

disrupts the geometric and temporal continuity essential for accurate action recognition. This disruption can cause models to misinterpret actions or fail to recognize, as skeleton-based one-shot action recognition approaches rely heavily on the visibility and precise positioning of all key joints.

In response to the challenges posed by occlusions, we developed Trans4SOAR, a cutting-edge transformer-based model by using cross-modal patch embedding level fusion and prototype contrastive learning. Trans4SOAR leverages three distinct data streams—joints, bones, and velocities—and integrates them using the newly proposed mixed attention fusion mechanism at the patch embedding level. This mechanism enables the model to capture and fuse diverse types of skeleton information effectively by considering the fusion from the auxiliary modals to the major modal on the query, key, and values aspects, which enables a more communicative fusion manner. Additionally, we introduced a latent space consistency loss, which utilizes category-specific prototypes to enhance the generalizability of the model's embeddings against data disturbance. This loss function ensures that the model maintains consistency in its embeddings, even when features are disrupted by occlusions.

Our experimental evaluations demonstrated the superior performance of Trans4SOAR across multiple datasets, including NTU-120, NTU-60, and Toyota Smart Home. The results showed that while occlusions adversely affect the accuracy of skeleton-based one-shot action recognition models, Trans4SOAR consistently outperforms existing state-of-the-art frameworks. It achieves higher accuracy and generalizability, particularly under diverse occlusion conditions. Furthermore, Trans4SOAR also excels in standard SOAR tasks without occlusions, surpassing the best previously published models by a notable margin. For instance, on the challenging NTU-120 SOAR benchmark, Trans4SOAR improved the accuracy by over 2.8% compared to the prior best model.

These findings affirm the effectiveness of Trans4SOAR in handling different occlusions, making it a versatile solution for skeleton-based one-shot action recognition task under occlusions. The integration of mixed attention fusion mechanism and latent space consistency loss in Trans4SOAR provides a powerful strategy that not only mitigates the impact of occlusions but also enhances overall skeleton-based one-shot action recognition performance, ensuring reliability in diverse and dynamic environments.

3.2 Self-Supervised Skeleton-Based Action Recognition in Occluded Environments

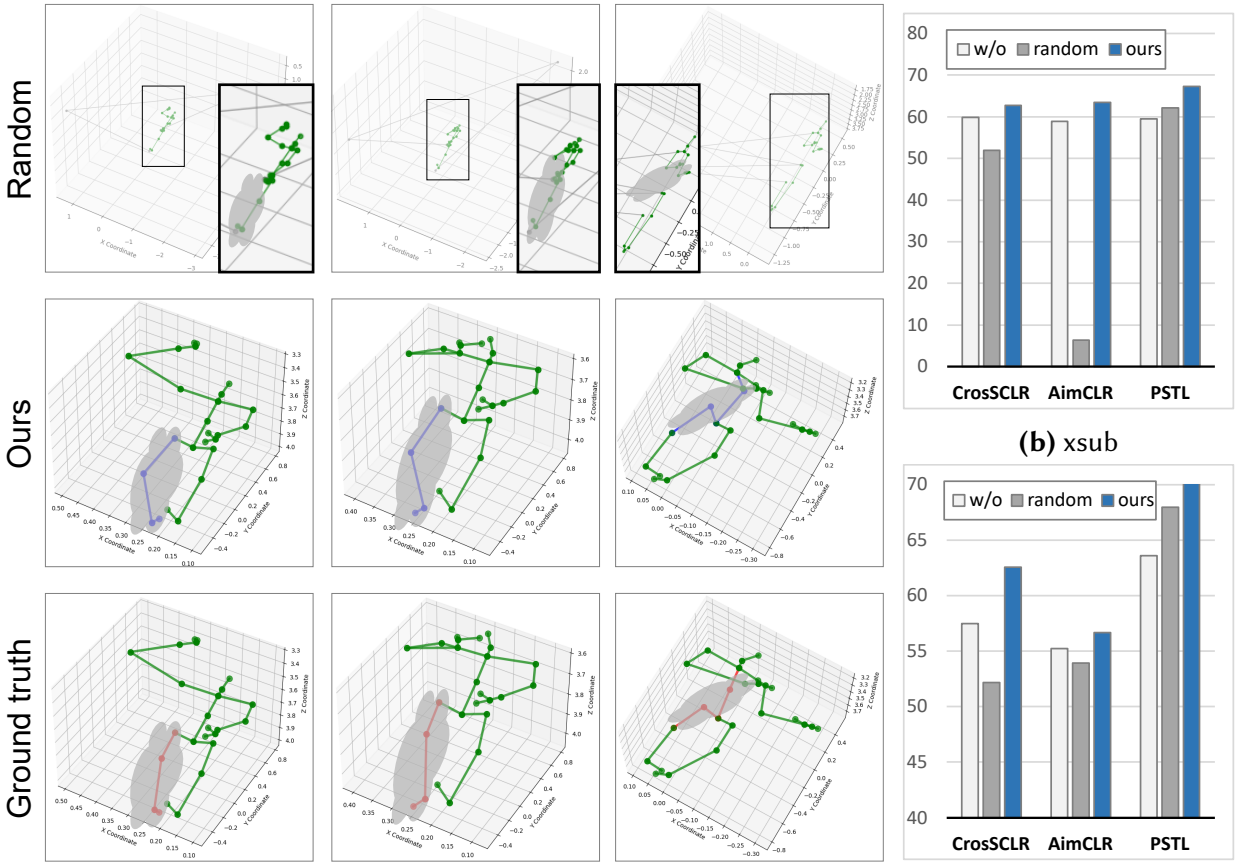
3.2.1 Introduction

The research field of human action recognition plays an important role in the advancement of robotics technology, underpinning significant applications across human-robot interaction, healthcare, industrial automation, security, and surveillance sectors [9, 46, 163, 175, 190]. This multifaceted utility is predicated on the ability of robotic systems to interpret and respond to human actions in a context-aware manner, thereby enabling collaborative human-robot partnerships, augmenting task-specific assistance, and enhancing patient care through vigilant monitoring and support [157, 208].

The integration of human action recognition systems within robotic platforms facilitates the autonomous detection of human intentions and goals, thereby optimizing the timing and relevance of robot interventions in a manner that minimally intrudes upon human activities. Moreover, the application of human action recognition in healthcare robots transcends mere assistance, extending into the realm of patient condition monitoring and rehabilitation support, with the potential to significantly contribute to improved recovery outcomes.

However, the pursuit of effective image- or video-based human action recognition is not without its challenges, including the complexities of dynamic backgrounds, diversity in human physique, variability in camera perspectives [162]. In contrast, skeleton-based human action recognition emerges as a viable alternative, characterized by its resilience to changes in appearance and its operational efficiency. Leveraging sparse 3D skeleton data, skeleton based human action recognition achieves rapid inference and reduced memory demands, rendering it particularly suitable for deployment in mobile robotics where computational resources are limited [17, 80, 225].

In recent years, we have witnessed significant strides in skeleton-based human action recognition field, propelled by technological leaps in depth sensing and pose estimation algorithms. The evolution of these technologies has facilitated the acquisition of high-fidelity skeleton data, thereby enhancing the accuracy and reliability of action recognition systems. Concurrently, the burgeoning field of self-supervised learning within skeleton-based human action recognition has garnered increasing interest within the robotics research community. This paradigm shift towards self-supervised methodologies promises to mitigate the reliance on extensively annotated datasets, thereby streamlining the training process and accelerating the development of advanced human action recognition systems [86, 120]. The confluence of these advancements underscores the potential of skeleton-based human action recognition to redefine the landscape of robotic applications, offering a pathway to more intuitive, label efficient, and context-aware robot-human interactions. In the field of



(a) Comparison of imputations under occlusion

(b) xsub

(c) xview

Figure 3.6: Comparison of different imputation methods. In (a), we compare random imputations (in gray), our imputation results (in blue), and ground-truth skeletons (in red). In (b) and (c), the linear evaluation results of cross-subject (*xsub*) and cross-view (*xview*) settings are tested by using imputation methods across three popular self-supervised action recognition methods (CrossCLR, AimCLR, and PSTL).

self-supervised skeleton-based action recognition, existing research predominantly focuses on data procured from occlusion-free environments, meticulously curated under controlled conditions [71, 106, 112]. However, the practical application of robotics frequently encounters scenarios where occlusions are prevalent, challenging the reliability of even the existing works on the skeleton-based human action recognition. Acknowledging this challenge, our research posits that incorporating occlusion-aware training and evaluation methodologies represents an essential yet underexplored avenue within the discipline.

Addressing the occlusion challenge within the realm of self-supervised skeleton-based action recognition necessitates a bifurcated approach: enhancing model resilience to occlusions through architectural innovations and mitigating data incompleteness by refining the skeleton coordinates. This dual perspective serves as the cornerstone of our pioneering investigation into self-supervised action recognition in occlusion scenarios.

We establish a new benchmark to build up the test bed of this task, to systematically assess the impact of occlusions on the efficacy of existing self-supervised action recognition methodologies. Preliminary evaluations on this benchmark reveal significant performance degradation across all considered methods when subjected to occluded skeleton data. Motivated by these findings, our contribution encompasses a hybrid solution that intertwines model and data-driven strategies to address these identified challenges.

From a model-centric perspective, we introduce the *Adaptive Spatial Masking (ASM)* technique, a novel data augmentation strategy devised by adapting to the distribution patterns of missing joints within the dataset. This approach, inspired by the state-of-the-art PSTL methodology [229], endeavors to optimize the utilization of available data for enhanced feature representation learning.

Simultaneously, we advocate for a data-driven solution aimed at ameliorating the effects of incomplete data. This involves a strategic visualization illustrated in Fig. 3.6, where the conventional method of employing direct K Nearest Neighbour (KNN) searches across the dataset for missing data completion is deemed computationally prohibitive. To circumvent these limitations, our approach employs a two-pronged strategy: initially segmenting the dataset into distinct clusters via KMeans [88] clustering based on features derived from self-supervised learning methods, followed by a targeted KNN imputation within these clusters to efficiently reconstruct missing skeleton coordinates.

This methodology not only streamlines the imputation process by obviating the need for exhaustive KNN searches across the entire dataset but also significantly curtails computational demands, enabling practical implementation.

The cumulative contributions of our work are encapsulated as follows:

1. The introduction of a new benchmark tailored for self-supervised skeleton-based action recognition under occlusions, encompassing the NTU-60 and NTU-120 datasets, aimed at facilitating the assessment of robotic action recognition capabilities within occlusion-prone environments.

2. The development and implementation of a computationally efficient two-stage imputation mechanism employing KMeans clustering and KNN imputation, designed to address the challenges posed by occluded skeleton data. This methodology demonstrates versatility and adaptability across a spectrum of self-supervised action recognition frameworks.

3. The formulation of the Occluded Partial Spatio-Temporal Learning (OPSTL) framework, which incorporates the Adaptive Spatial Masking (ASM) data augmentation to enhance the model's ability to leverage high-quality skeleton data in the presence of occlusions.

4. The substantiation of our method's efficacy through rigorous experimental validation on occluded versions of the NTU-60 and NTU-120 datasets, thereby showcasing the potential of our approach in advancing the field of self-supervised skeleton-based action recognition amidst occlusions.

This research endeavor not only pioneers the exploration of occlusion-aware self-supervised skeleton-based action recognition but also lays the groundwork for future advancements in the field, setting a precedent for subsequent studies to build upon.

3.2.2 Methodology

Our technique, denoted as Occluded Partial Spatio-Temporal Learning (OPSTL), draws inspiration from PSTL [229] for its exceptional efficacy in managing occlusions compared to alternate methodologies. We introduce an ASM on the top of the original Central Spatial Masking (CSM) from PSTL to enhance occlusion handling during the initial phase of self-supervised training (first stage). The process of dealing with occlusions involves data imputation prior to the commencement of the second phase of self-supervised training. In this second stage, embedding clusters, which are formed post the initial training stage alongside the application of KNN, are utilized as illustrated in Fig. 3.7. The KMeans algorithm groups embeddings into finer clusters, followed by the execution of data imputation through KNN search to identify samples analogous to those requiring completion.

It's noteworthy that the data imputation strategy employed in the second stage is applicable across various self-supervised skeleton-based action recognition methods. By filling in the missing data, conducted experiments are poised to witness enhancements in performance across different self-supervised skeleton-based human action recognition methods in comparison to scenarios where no imputation is leveraged.

3.2.2.1 Pre-processing

A skeleton sequence, after undergoing preprocessing, is denoted as $\mathbf{s} \in \mathbb{R}^{T \times J \times C}$, transformed from the initial input $\mathbf{I} \in \mathbb{R}^{T \times J \times C \times M}$. Here, T signifies the number of frames, J indicates the number of joints, C represents the number of channels, and M denotes the number of persons in the sequence. This preprocessing approach is akin to that utilized in CrosSCLR [106], where skeleton coordinates are adjusted to be relative to the skeleton's center joint. For an enhanced application of ASM, it is crucial to establish a boolean matrix for missing joints ($\mathbf{B} \in \mathbb{B}^{N \times V}$) for each V joint per sample. This matrix helps in accurately identifying joints with higher occlusion rates. All absent joint coordinates are marked as "*nan*" to streamline the process of calculating Euclidean distances amid missing values during the data imputation phase.

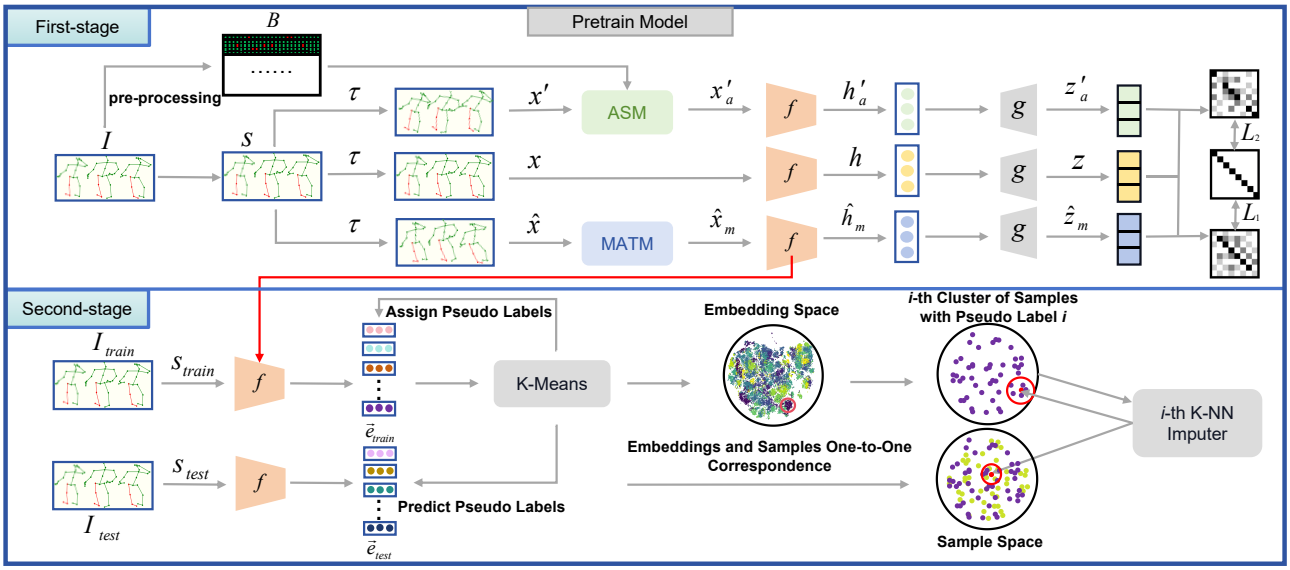


Figure 3.7: Our methodology for the imputation of the occluded skeleton coordinates unfolds in two stages, with the missing skeleton portions in the input I depicted in red.

3.2.2.2 Partial Spatio-Temporal Skeleton Representation Learning

Existing approaches [71, 106] concentrate on the generation of varied views of skeleton sequences for contrastive learning purposes, yet they frequently neglect the intricate local interconnections among different skeleton joints and frames. These local interactions, however, are crucial for practical applications, offering essential context for actions. To address this oversight, PSTL [229] capitalizes on these local relationships through a distinctive strategy of spatiotemporal masking to generate incomplete skeleton sequences. This method employs a triplet stream structure that includes an anchor stream alongside a spatial masking stream that incorporates CSM, and a temporal masking stream equipped with Motion Attention Temporal Masking (MATM). As a result, PSTL demonstrates proficiency in managing occlusions. Barlow Twins framework [219] leverages redundancy reduction to learn effective representation without labels. By maximizing the similarity between differently augmented views of the same skeleton sequence while minimizing the redundancy between feature components, it achieves superior embedding learning. Adopting the Barlow Twins framework, PSTL circumvents the limitations associated with contrastive learning, which demands an extensive collection of negative samples, along with substantial batch sizes and memory banks [71, 106]. This approach is adopted in our study as the self-training backbone.

3.2.2.3 Adaptive Spatial Masking

The self-supervised technique for recognizing skeleton-based actions, PSTL [229], utilizes CSM to bolster the resilience of its joint representations. CSM works by encouraging the contrastive

learning of embeddings from incomplete and complete skeleton data, thus benefiting the encoder in understanding the linkage between masked and exposed joints. The selection of joints for masking under CSM is guided by the centrality within the topology of the human skeleton graph, with joints possessing a higher degree of centrality, indicative of a wealth of neighborhood information, being more prone to masking. The probability of a joint being masked is represented as Eq. 3.21.

$$p_i = \frac{d_i}{\sum_{j=1}^{n} d_j}, \tag{3.21}$$

where d_i signifies the degree of each joint v_i.

This approach, however, does not account for actual occlusion scenarios. To improve upon this, when specific joints are occluded more frequently in the training dataset, selecting these for masking could leverage high-quality data more effectively. Conversely, with higher rates of random occlusion across samples, adopting a more stochastic masking approach could better mimic occlusion distributions. In response, we introduce a dataset-driven ASM method, designed to dynamically alternate between partial and random occlusion scenarios, incorporating CSM within its framework for situations where no occlusions are present, thereby selecting joints for masking. The degree of each joint is recalibrated based on a batch's missing joint boolean matrix (**B**), with the occlusion frequency for each joint $\mathbf{v}_i$, for $i \in (1, 2, \ldots, n)$, computed per batch. The Frequency Degree (FD) for joint $\mathbf{v}_i$ is calculated as Eq. 3.22.

$$FD_i = \lfloor \frac{F_i - \min(F)}{\max(F) - \min(F) + \varepsilon} \times 3 + 1 \rfloor, \tag{3.22}$$

where ε is a negligible constant set to 0.001. Observations indicate that most joints possess a centrality degree around 2, with minimal variance in these degrees. Therefore, the distinction in efficacy between random and centrality-based masking is marginal. To address this, the occlusion frequency for each joint is normalized to a range akin to centrality degrees, specifically $[1, 3]$. This normalization accentuates the disparity in frequency degrees over the skeleton graph's centrality degrees, thereby skewing the preference towards masking joints with higher occlusion frequencies. Here, F denotes the occlusion frequency of joints, calculated from **B** across the batch dimension Eq. 3.23:

$$F_i = \sum_b \mathbf{B}_{b,i}. \tag{3.23}$$

3.2.2.4 Imputation

Drawing inspiration from the KNN imputation technique rooted in traditional machine learning, our objective is to identify samples closely resembling those with missing values for the purpose

of imputation. Given the high dimensionality and vast quantity of sample data, a direct neighbor search within the entire sample space proves to be unfeasible. Consequently, we opt to forgo a comprehensive search in favor of grouping the samples into clusters containing a reduced number of skeleton samples. Through initial pre-training of the self-supervised skeleton-based human action recognition approach in the first stage, KMeans effectively groups samples of identical action types into different clusters.

Initially, features are extracted from the samples using the pre-trained model from the first phase. Pseudo-labels are then assigned to each embedding. Within a cluster holding a pseudo label i, KNN is employed to pinpoint neighboring samples for the one requiring imputation, residing in the same cluster. Given that these neighbor samples might also lack values, the conventional Euclidean distance is inapplicable. Instead, we propose a modified Euclidean distance tailored for missing values [51, 145], expressed as Eq. 3.24

$$D(\mathbf{s}_{ij}, \mathbf{s}_{ik}) = \sqrt{w \times d_{ignore}(\mathbf{s}_{ij}, \mathbf{s}_{ik})}, \tag{3.24}$$

where w denotes a weight reflecting the ratio of total coordinates to the number of existing coordinates, and $d_{ignore}(\mathbf{s}_{ij}, \mathbf{s}_{ik})$ represents the Euclidean distance between samples j and k within the i-th cluster, disregarding any missing values in $\mathbf{s}_{ij}$ and $\mathbf{s}_{ik}$.

This distance metric facilitates the straightforward calculation of distances between sample pairs. The nearest k samples $\mathbf{s}_{ij}^{near}$, with j ranging from 1 to k, are identified based on distance and the location of missing data within the current cluster i for a sample $\mathbf{s}_i^{miss}$ lacking data. Each chosen sample $\mathbf{s}_{ij}^{near}$ must possess a complete coordinate at positions corresponding to where the missing coordinates $\mathbf{c}_i^{miss} = \{c \mid c \in \mathbf{s}_i^{miss}\}$ are found. The imputation equation for a missing skeleton coordinate in a sample with missing data is as Eq. 3.25.

$$\mathbf{c}_i^{miss} := \frac{\sum_{j=1}^{k} r_j \times \mathbf{c}_{ij}^{near}}{\sum_{j=1}^{k} r_j}, \tag{3.25}$$

where r_j is the reciprocal of the modified Euclidean distance, referred to as $dist$, between the sample with missing data $\mathbf{s}_i^{miss}$ and one of the nearest k samples $\mathbf{s}_{ij}$ within the nearest cluster i, as Eq. 3.26.

$$r_j = \frac{1}{D(\mathbf{s}_{ij}, \mathbf{s}_i^{miss})}. \tag{3.26}$$

As depicted in Fig. 3.7, a notable distinction in the imputation process between the training and test sets is the absence of re-clustering for the test set. Instead, the KMeans model trained on the training set predicts pseudo-labels for the test set, and imputed data is generated using training

set clusters that match the test set's predicted pseudo-labels. The test set serves exclusively as an imputation data source.

Although this methodology has enhanced performance in diverse downstream tasks across several models, limitations persist. Specifically, when every sample in a cluster lacks certain joints, these missing sections remain unimputed, thus not ensuring a comprehensive imputation of all absent skeleton coordinates.

3.2.3 Experiments

3.2.3.1 Datasets

NTU-60/120 with occlusion. Derived occluded datasets originate from NTU-60/120. The NTU-60 dataset [161], captured through Microsoft Kinect sensors, consists of 56,578 skeleton sequences across 60 unique action categories. It offers two division schemes [161]: 1) Cross-Subject (xsub), where training and validation data are sourced from distinct individuals, and 2) Cross-View (xview), in which training and validation datasets are obtained from disparate camera perspectives. The NTU-120 dataset [116], an augmentation of NTU-60, includes 113,945 skeleton sequences covering 120 action categories. While maintaining the xsub protocol, NTU-120 introduces the xset protocol for evaluations across different camera setups, as opposed to camera views.

Occlusions are categorized into two types: 1) Synthesized realistic occlusion [146] as mentioned in Section 3.1 utilizes projections of 3D furniture to craft realistic occlusions. 2) Random occlusion, determined by the minimum and maximum coordinate values, where 20% of the coordinates undergo random selection for occlusion.

3.2.3.2 Protocols

Linear Evaluation Method. This approach involves training a supervised linear classifier, which includes a fully connected layer followed by a SoftMax activation, while the encoder remains unchanged.

Semi-Supervised Testing Method. Initially, the encoder is pre-trained with the complete imputed dataset. Then, the full model is fine-tuned using merely 1% or 10% of the labeled data, selected randomly.

Fine-tuning Method. A linear classifier is coupled with the pre-trained encoder, and the entire network is subsequently fine-tuned on the imputed dataset.

3.2.3.3 Implementation Details

In our research, the skeleton based feature learning backbone of each self-supervised learning approach is unified as ST-GCN [209] with 16 hidden channels. The preprocessing methodology is in alignment with those outlined in CrosSCLR [106] and AimCLR [72], entailing the elimination of invalid frames from skeleton sequences, normalization of sequence length to 50 frames through linear interpolation, and conversion of coordinates to a relative format. Additionally, we analyze the distribution of missing joints (**B**) within the dataset. For the training phase, the Adam optimizer [92] is utilized, coupled with a CosineAnnealing scheduler across a span of 150 epochs for both the learning of representations and application to downstream tasks. The training is conducted with a batch size of 128 and an initial learning rate of $5e^{-3}$.

Data Augmentation. Before the feature extraction phase in model training, data augmentation techniques are applied to enhance the variability of skeleton sequences. Different models implement their own distinct combinations of data augmentation strategies. For example, SkeletonCLR [106] and CrosSCLR apply one form of spatial augmentation (Shear) along with a temporal augmentation technique (Crop). Conversely, PSTL employs a trio of spatial augmentations (Shear, Rotate, Spatial Flip) and a single temporal augmentation (Crop). Our model, OPSTL, adopts the identical data augmentation scheme as that of PSTL.

Self-Supervised Pre-training Protocol. To facilitate a fair comparison with PSTL, identical parameter settings are employed. As illustrated in Fig. 3.7, the transformation τ applies to the incomplete skeleton sequence $\mathbf{s}$, producing three distinct views $\mathbf{s}$, $\mathbf{s}'$, and $\hat{\mathbf{s}}$. Both $\mathbf{s}'$ and $\hat{\mathbf{s}}$ undergo ASM and MATM processes, respectively, resulting in partial skeleton sequences $\mathbf{s}'_a$ and $\hat{\mathbf{s}}_m$. Using ST-GCN as the backbone, we extract 256-dimensional features $\mathbf{f}$, $\mathbf{f}'_a$, and $\hat{\mathbf{f}}_m$, which are further transformed into $6,144$-dimensional embeddings $\mathbf{z}$, $\mathbf{z}'_a$, and $\hat{\mathbf{z}}_m$ via the projector $\mathbf{g}$. The cross-correlation matrices between $\mathbf{z}$ and $\mathbf{z}'_a$, and between $\mathbf{z}$ and $\hat{\mathbf{z}}_m$, are computed to understand the relationships between masked and unmasked joints. The loss is calculated from these matrices with loss parameter λ set to $2e^{-4}$, incorporating a warm-up period of 10 epochs. The weight decay parameter is $1e^{-5}$. For ASM, 9 joints are selected for masking, and for MATM, 10 frames are masked.

Imputation Strategy. Our proposed imputation strategy addresses occlusion challenges. For clustering in the imputation phase, KMeans is utilized with 60 clusters for NTU-60 and 120 clusters for NTU-120 datasets, specifically for handling realistic occlusions. The KNN method, with $k = 5$, identifies neighboring samples for imputation.

Table 3.14: Linear evaluation results on **NTU-60** with synthesized realistic occlusion, randomly imputed values, and imputed values by our proposed method. "Δ" represents the difference compared to the non-imputed NTU-60. **J** and **M** represent the joint stream and the motion stream.

Method	**Stream**	**Occluded** (%)		**Randomly imputed** (%)				**Our imputed** (%)			
		xsub	xview	xsub		xview		xsub		xview	
		acc.	acc.	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ
SkeletonCLR[106]	**J**	56.74	53.25	47.12	↓9.62	58.09	↑4.84	57.61	↑0.87	64.43	↑11.18
2s-CrosSCLR[106]	**J+M**	59.88	57.47	51.96	↓7.92	52.18	↓5.29	62.76	↑2.88	62.54	↑5.07
AimCLR[71]	**J**	58.90	55.21	6.36	↓52.54	53.91	↓1.30	63.40	↑4.50	56.68	↑1.47
PSTL[229]	**J**	59.52	63.60	62.18	↑2.66	67.97	↑4.37	**67.31**	↑7.79	71.10	↑7.50
OPSTL (ours)	**J**	**61.11**	**65.55**	**65.63**	↑4.52	**68.01**	↑2.46	67.11	↑6.00	**71.39**	↑5.84

Table 3.15: Linear evaluation results on **NTU-120** with synthesized realistic occlusion, randomly imputed values, and imputed values by our proposed method. "Δ" represents the difference compared to the non-imputed NTU-120. **J** and **M** represent the joint stream and the motion stream.

Method	**Stream**	**Occluded** (%)		**Randomly imputed** (%)				**Our imputed** (%)			
		xsub	xset	xsub		xset		xsub		xset	
		acc.	acc.	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ
SkeletonCLR[106]	**J**	44.93	42.78	44.42	↓0.51	40.12	↓2.66	48.63	↑3.70	45.06	↑2.28
2s-CrosSCLR[106]	**J+M**	49.63	48.14	39.11	↓10.52	33.77	↓14.37	49.58	↓0.05	54.43	↑6.29
AimCLR[71]	**J**	44.58	48.93	0.86	↓43.72	1.16	↓47.77	52.50	↑7.92	52.83	↑3.90
PSTL[229]	**J**	54.18	51.90	56.12	↑1.94	52.66	↑0.76	57.05	↑2.87	57.94	↑6.04
OPSTL (ours)	**J**	**55.65**	**54.18**	**56.43**	↑0.78	**53.90**	↓0.28	**59.29**	↑3.64	**58.25**	↑4.07

3.2.3.4 Evaluating Against Non-imputed NTU-60/120

To validate the efficacy of our imputation technique, it is benchmarked on the original NTU-60/120 datasets with realistic occlusion. As demonstrated in Tables 3.14, 3.15, and 3.16, the performance across nearly all three downstream tasks for all evaluated methods exhibits improvements on the imputed versions of the NTU-60/120 datasets. We observe that most of the existing self-supervised skeleton-based human action recognition approaches can not work well under the disturbance from occlusions, while our OPSTL approach with ASM structure can outperforms all of the leveraged baselines, illustrating the importance by using the proposed adaptive spatial masking strategy in three branch contrastive learning framework. OPSTL delivers 61.11% and 65.55% accuracy on NTU-60 xsub and xview and 55.65% and 54.18% accuracy on NTU-120 xsub and xview by using the linear evaluation protocol, respectively.

Table 3.16: Finetune and semi-supervised results on the imputed NTU-60/120 with synthesized realistic occlusion."Δ" represents the difference compared to the non-imputed NTU-60/120 with synthesized realistic occlusion. **J** and **M** represent the joint stream and the motion stream.

Method	Stream	Imputed NTU-60 (%)				Imputed NTU-120 (%)			
		xsub		xview		xsub		xset	
		acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ
Finetune:									
SkeletonCLR[106]	**J**	70.58	↑3.22	80.76	↑4.42	63.17	↑4.06	62.12	↑1.20
2s-CrosSCLR[106]	**J+M**	72.94	↑1.32	80.34	↑0.09	65.06	↑0.39	67.45	↑2.43
AimCLR[71]	**J**	70.53	↑0.44	75.52	↓3.21	67.08	↑5.25	66.62	↑1.91
PSTL[229]	**J**	75.16	↑2.48	85.24	↑2.05	69.10	↑1.20	**69.42**	↑2.71
OPSTL (ours)	**J**	**75.43**	↑2.41	**86.01**	↑1.92	**70.89**	↑2.21	69.14	↑1.89
Semi 1%:									
SkeletonCLR[106]	**J**	31.99	↑13.47	31.18	↑10.16	20.45	↑3.13	16.24	↑2.23
2s-CrosSCLR[106]	**J+M**	32.66	↑4.69	31.18	↑10.35	19.38	↑0.21	20.12	↑7.59
AimCLR[71]	**J**	34.44	↑5.28	27.04	↑8.92	22.59	↑6.13	20.68	↑5.38
PSTL[229]	**J**	**40.81**	↑7.99	**39.61**	↑13.06	27.43	↑5.92	**25.52**	↑6.52
OPSTL (ours)	**J**	40.07	↑6.48	38.65	↑10.76	**27.90**	↑4.69	24.57	↑4.71
Semi 10%:									
SkeletonCLR[106]	**J**	55.97	↑2.98	60.83	↑9.37	44.37	↑2.33	42.68	↑7.72
2s-CrosSCLR[106]	**J+M**	59.17	↑3.16	59.01	↑3.86	46.89	↑2.07	48.24	↑8.44
AimCLR[71]	**J**	59.64	↑2.30	54.34	↓0.36	48.38	↑6.25	48.96	↑3.63
PSTL[229]	**J**	63.04	↑4.41	68.89	↑7.54	53.26	↑2.80	53.42	↑4.14
OPSTL (ours)	**J**	**64.04**	↑5.26	**70.04**	↑6.22	**54.71**	↑2.90	**53.50**	↑3.38

3.2.3.5 State-of-the-art Comparisons

Our evaluation of OPSTL includes comprehensive comparison experiments. According to the results presented in Tables 3.14, 3.15, and 3.16, OPSTL surpasses the previously leading PSTL in linear evaluations across both non-imputed and imputed variants of NTU-60/120. Specifically, OPSTL exhibits enhancements of 1.59% and 1.95% for xsub and xview protocols on the NTU-60 dataset with realistic occlusion, respectively. Moreover, it secures gains of 1.47% and 2.28% on xsub and xset protocols of the NTU-120 dataset, also under conditions of realistic occlusion. These improvements are evident not just on the non-imputed dataset but extend to the imputed versions of NTU-60/120 as well. Compared with OPSTL w/o imputation, our imputation method can deliver 3.64% and 4.07% accuracy improvements while limited benefits are delivered by randomly imputation, indicating the effectiveness of the skeleton completion. The superior performance brought by our proposed imputation method demonstrating its promising capability of completing missing skeleton body joints

which are disrupted by occlusions.

3.2.3.6 Ablation Analysis

Through ablation studies, we underscore the efficacy of our proposed ASM and imputation strategy. Evaluating the imputation strategy involved testing with random imputation on NTU-60/120 datasets featuring realistic occlusion. The comparative analysis, as seen in Table 3.14 and Table 3.15, reveals that our imputation method significantly surpasses random imputation in linear evaluation. Methods such as SkeletonCLR, 2s-CrosSCLR, and AimCLR experience performance drops with random imputation, with AimCLR showing the most drastic decrease, where accuracy plunges by nearly 1%. This indicates that random imputation adversely affects the integrity of skeleton data, which is critical for action recognition methods that rely on complete skeleton information.

Conversely, for approaches that engage partial skeleton sequences in learning representations, like PSTL, accuracy still slightly increases even with random imputation applied. Nevertheless, with our ASM approach, any improvement is marginal or sometimes negative. For example, on the xset protocol of randomly imputed NTU-120, OPSTL sees a decline of 0.28% in accuracy compared to its non-imputed counterpart, yet it remains superior to the state-of-the-art PSTL. This demonstrates ASM's capacity to harness high-quality data effectively for learning representations.

3.2.4 Discussion

This study makes significant contributions to self-supervised skeleton-based action recognition by addressing the critical challenge of occlusions in real-world environments. We introduce a large-scale benchmark by incorporating realistic occlusions and random occlusions on self-supervised skeleton-based human action recognition approaches. This benchmark highlights the substantial performance degradation of existing self-supervised skeleton-based human action recognition methods under occlusions, emphasizing the need for more generalizable solution. Current self-supervised methods often rely on contrastive learning to understand human motion sequences, expecting similar sequences to have smaller latent space distances. However, occlusions disrupt this similarity by introducing noise and missing information, particularly harming contrastive learning's ability to accurately group similar actions.

To address these challenges, we propose the OPSTL framework. This framework introduces an adaptive spatial masking data augmentation technique, which masks joints based on their occlusion frequency, leveraging high-quality skeleton data to enhance feature learning. Unlike traditional methods, adaptive spatial masking dynamically adjusts based on observed occlusion patterns, providing a more realistic and robust training process. Additionally, a two-stage imputation approach

completes missing skeleton data efficiently. First, KMeans clustering groups semantically similar samples from sequence embeddings. Second, KNN within each cluster imputes missing coordinates based on the closest sample neighbors, preserving geometric and temporal continuity essential for action recognition.

Experimental results demonstrate OPSTL's superior performance across multiple self-supervised skeleton-based action recognition models. On occluded versions of the NTU-60 and NTU-120 datasets, OPSTL mitigates occlusion effects and enhances accuracy and robustness. In linear evaluation protocols, OPSTL consistently outperforms existing methods, achieving higher accuracy in cross-subject and cross-view settings. For example, on the NTU-120 dataset with realistic occlusions, OPSTL improves accuracy by 1.47% in cross-subject and 2.28% in cross-set protocols. In semi-supervised evaluation protocols, OPSTL shows significant gains, particularly with a small percentage of labeled data, highlighting its efficiency and effectiveness.

Overall, our study emphasizes the importance of addressing occlusions in self-supervised skeleton-based action recognition. By combining advanced data augmentation techniques with efficient imputation methods, the OPSTL framework enhances the performance and reliability of action recognition systems in dynamic and unpredictable environments.

4 | Towards Open-Set Skeleton-Based Action Recognition

Apart from the challenge of occlusion on the tasks requiring discriminative feature learning in the field of skeleton-based human action recognition, open-set recognition is a more challenging and still unexplored area in the skeleton-based human action recognition, which requires the model to deliver low confidence score on the unseen categories. In this chapter, we will for the first time open the vistas to the task open-set skeleton-based action recognition. Part of the content of this chapter is from our publication [149] in the thirty-eighth AAAI conference on artificial intelligence.

4.1 Introduction

Utilizing skeleton sequences for recognizing human actions offers numerous advantages, including enhanced privacy, reduced data size, and improved adaptability to unfamiliar human appearances. Contemporary approaches based on skeletons [230] are static in their predictive capabilities post-training. A scenario closer to reality involves models encountering *open sets*, which include both known and previously unseen action categories [136]. Actions not within the model's learned categories often lead to incorrect classifications as known actions, potentially causing considerable issues, especially when these outputs have influence on decision-making processes, such as in the context of assistive robotics. The necessity for advancements in open-set, skeleton-based action recognition is highlighted by previous research [62, 137], driving the motivation behind our study.

While several methods have been developed for open-set action recognition *in videos* [10], the challenge of identifying novel actions from *skeleton* data has not been adequately addressed. Despite pursuing similar objectives, the two tasks differ significantly due to the lack of visual background information and the sparse nature of skeleton sequences, posing unique challenges in managing out-of-distribution actions. Addressing the absence of an appropriate benchmark, we establish a comprehensive benchmark for **O**pen-**S**et **S**keleton-based **A**ction **R**ecognition (OS-SAR), incorporating three notable backbones for skeleton-based human action recognition, which are CTRGCN [33],

HDGCN [100], and Hyperformer [49]. This benchmark is constructed on three publicly available datasets, *i.e.*, NTU-60 [161], NTU-120 [114], and Toyota Smart Home [43], and includes defined open-set splits and evaluation protocols. Effective open-set recognition methods should sustain consistent performance across various dataset and backbone combinations. Adopting practices from open-set image classification [127], we compute performance averages over five random splits of unseen classes. However, traditional open-set recognition techniques often falter in delivering consistent OS-SAR outcomes, with recognition accuracy varying significantly across different backbones and evaluation settings. This variation underscores the limitations of current approaches in addressing OS-SAR challenges, which can not disentangle the samples from seen and unseen categories well on the perspective of the prediction confidence.

To address these issues, we introduce a novel strategy for OS-SAR. Our multimodal method utilizes three streams: joints, velocities, and bones, facilitating an exchange of distribution-wise information in their latent spaces through a Cross-Modality Mean Max Discrepancy (CrossMMD) suppression mechanism. Additionally, we tackle the issue of overconfidence in SoftMax-normalized probability estimate when dealing with mixed distributions by leveraging a distance-based confidence measure, the **C**hannel **N**ormalized **E**uclidean distance (CNE-distance), relative to the nearest training set embeddings in latent space. While this approach markedly enhances open-set recognition, it does not perform as well in close-set scenarios compared to traditional SoftMax. To merge the strengths of both approaches, we propose a cross-modality distance-based logits refinement technique, combining modality-averaged logits with CNE-distances. This integrated method, named CrossMax, incorporates both CrossMMD for training and cross-modality distance-based refinement for testing, setting new state-of-the art performances on OS-SAR task across datasets, backbones, and evaluation settings.

Our primary contributions are summarized as follows:

- The development of a large-scale benchmark for Open-Set Skeleton-based Action Recognition (OS-SAR), encompassing three datasets for skeleton-based human action recognition classification, seven open-set recognition baselines, and three established backbones for skeleton data streams.
- A multimodal OS-SAR strategy that leverages joints, velocities, and bones streams, with a Cross-Modality Mean Max Discrepancy (CrossMMD) suppression mechanism for inter-stream information exchange.
- The introduction of a Channel Normalized Euclidean distance (CNE-distance) as a confidence measure to mitigate overconfidence issues in SoftMax-normalized probabilities, thereby improving open-set recognition accuracy.

- The proposal of the CrossMax approach, which merges CrossMMD training and a cross-modality distance-based logits refinement technique, achieving generalizable superior OS-SAR performance in various evaluations.

4.2 Methodology

4.2.1 Task Introduction

In this work we focus on open-set recognition problem. Open-set recognition addresses the challenge of classifying skeleton sequences from both known and unknown classes. The confidence sore is usually measured by the SoftMax probability of the correct/wrongly predicted class within the open-set realm. Formally, let $\mathscr{X}$ denote the input space and $\mathscr{Y} = \{y_1, y_2, \ldots, y_C\}$ represent the set of known classes, where C is the number of classes present during training. Given a training dataset $\mathscr{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ consisting of N labeled samples, where $x_i \in \mathscr{X}$ and $y_i \in \mathscr{Y}$, the goal is to learn a classifier $f : \mathscr{X} \to \mathscr{Y}$ that can accurately classify test samples from the known classes $\mathscr{Y}$.

In the OSR setting, however, the test set $\mathscr{D}_{\text{test}} = \{x_j\}_{j=1}^{M}$ includes samples from both the known classes $\mathscr{Y}$ and a set of unknown classes $\mathscr{Y}_{\text{unknown}}$, such that $\mathscr{Y}_{\text{unknown}} \cap \mathscr{Y} = \emptyset$. The classifier must not only assign correct labels to samples from the known classes but also identify and reject samples from the unknown classes, typically by assigning them to a special "unknown" class label or by outputting a low confidence score for such samples.

To evaluate the performance in this open set scenario, we consider two main objectives:

- **Closed Set Accuracy**: The accuracy of the classifier in labeling samples from the known classes $\mathscr{Y}$.
- **Open Set Detection**: The ability of the classifier to correctly identify and reject samples from the unknown classes $\mathscr{Y}_{\text{unknown}}$. This is often measured by metrics such as the Area Under the Receiver Operating Characteristic Curve (AUROC) or the Area Under the Precision-Recall Curve (AUPR).

In a typical OSR scenario, the model is trained on a set of known classes, but during inference, the models are required to provide low confidence scores on the previous unseen categories during the training procedure.

4.2.2 Benchmark

We present OS-SAR, a comprehensive benchmark designed for **O**pen-**S**et **S**keleton-based **A**ction **R**ecognition, incorporating CTRGCN [33], HDGCN [110], and Hyperformer [49] as foundational

models to assess cross-model adaptability. This benchmark is developed upon the NTU-60 [161], NTU-120 [114], and Toyota Smart Home [43] datasets, which are adapted for skeleton-based human action recognition under open-set scenarios. The selection of backbones and benchmarks will be elaborated upon, with the dataset details to be discussed in the experiments section.

4.2.2.1 Backbones for Skeleton Representation Learning.

CTRGCN [33] utilizes the Channel-wise Topology Refinement Graph Convolution (CTRGC) mechanism for dynamically learning unique topologies, thereby enabling efficient feature aggregation across various channels. *HDGCN* [110] employs a Hierarchically Decomposed (HD) GCN structure. It takes advantage of an HD-Graph that segregates nodes into several groups, facilitating the capture of both structurally adjacent and semantically relevant distant connections. *Hyperformer* [49] adopts a transformer-based model, integrating bone connectivity through graph distance embeddings. These models were chosen to examine the adaptability of open-set approaches across different architectures, given their proven effectiveness in standard benchmarks for skeleton-based human action recognition and their diverse foundational technologies.

4.2.2.2 Existing Open-Set Recognition Baselines.

Our benchmark incorporates open-set recognition methodologies from image and video classification fields, due to the absence of open-set recognition methods tailored for skeleton data. These methods are adaptable to various skeleton-based architectures. *From image classification open-set baselines:* We adopt principal-point distance-based methods such as RPL [25] and ARPL [24], along with the prototype learning method PMAL [127], which stands as the current frontrunner in open-set image classification, and the foundational SoftMax score [74] for comparison. *From video-based action recognition open-set baselines:* We select DEAR [10], employing deep evidential learning for estimating open-set probabilities, alongside Monte Carlo Dropout + Voting (MCD-V) [155], and Humpty [53], which utilizes temporal graph reconstruction for open-set probability assessment. These open-set recognition baselines are applied in conjunction with our chosen skeleton representation backbones to ensure equitable evaluation, replacing the conventional image/video backbones with our specified skeleton-based models.

4.2.3 CrossMax

To exploit the distinct cues offered by various skeleton data modalities (*e.g.*, joints, bones, and velocities) for open-set action recognition, we employ three ensembled backbones to perform feature extraction. To align the learned latent spaces across modalities, we first introduce a cross-modality

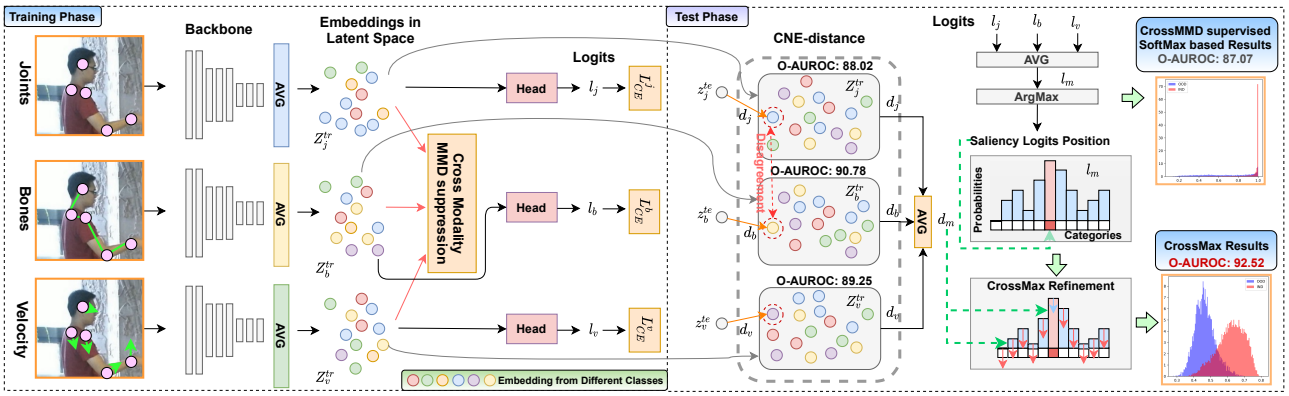


Figure 4.1: An overview of CrossMax method. During training, we utilize the Cross-modality Mean Maximum Discrepancy (CrossMMD), to better align the latent spaces across different modalities. At test-time, for each modality, we calculate the Euclidean distance to the closest training set sample and combine this with the averaged logits from the three branches. This combination undergoes a refinement process based on the cross-modality distance, which is conducted differently on the salient and not-salient logits. The refined logits are then processed through SoftMax, a better confidence estimate for both in- and out-of-distribution samples, while keeping the accurate close-set classification capability inherent to the standard SoftMax.

mean-max discrepancy suppression approach (CrossMMD) in training phase to ensure consistency in the learned information, while using conventional SoftMax score as open-set probability. However, we observe that effectively disentangling the in-distribution and out-of-distribution data remains an open challenge for open-set recognition through vanilla SoftMax score. To tackle this concern, we then propose a novel solution using the channel-wise L2 normalized nearest Euclidean distance to the train set embeddings, in a cross-modality setting to serve as distance-based open set probability score and considering different modalities. Nonetheless, it is worth noting that the probability predicted by SoftMax, operating on logits, offers valuable probability-like values for each category, facilitating intuitive probability interpretation and clear decision-making by selecting the class with the highest probability, where these benefits are not preserved for distance-based approaches. To capitalize on the benefits of both the proposed distance-based probability prediction schema and the conventional SoftMax probability prediction schema, we introduce a novel logits refinement approach during the test phase. This approach refines the predicted logits employed before SoftMax by using the averaged cross-modality distance. CrossMax combines these techniques while enhancing the overall predictive capabilities, providing more robust and accurate discrimination between in-distribution and out-of-distribution samples in terms of the distribution style on the predicted open-set probability while preserving the accurate close-set classification. The two major components of our CrossMax approach will be introduced in the following. This allows us to calibrate the aggregated cross-modality SoftMax score after the cross-modality mean-max discrepancy suppression, facilitating the disentanglement of uncertainty predictions for both in-distribution and out-of-distribution samples during the test phase. The two major components of our CrossMax

approach will be introduced in the following.

4.2.3.1 CrossMMD.

We introduce CrossMMD during training phase to enhance the information interchange across different modalities. Utilizing the concept of MMD, which measures the disparity between probability distributions [70], CrossMMD is applied as a loss function to foster greater alignment among distributions from diverse modalities, addressing the paucity of MMD research in cross-modal contexts.

Our aim is to mitigate the significant variance between the latent spaces of different modalities, capitalizing on the unique open-set discriminative features inherent to each modality for effective information sharing based on distribution characteristics. To accomplish the CrossMMD, we employ the Gaussian kernel within the Reproducing Kernel Hilbert Space (RKHS) for this purpose.

Given two batches of embeddings, Ω_x and Ω_y, from distinct modalities, viewed as separate distributions, we first concatenate them to create $\Omega_z = Concat(\Omega_x, \Omega_y)$. Subsequently, we calculate the pairwise L2 Norm distance among all samples, denoted as d_z. The kernel function's bandwidths, determined by Eq. 4.1 based on the sum of distances and the number of samples,

$$BW = \frac{\sum(d_z)}{(N_z)^2 - N_z}, \tag{4.1}$$

where N_z represents the number of samples. With N_k indicating the number of kernels, we derive a bandwidth list L_{BW} as $\{BW * (\alpha)^i | \ i \in [0, N_k]\}$, where α is a scaling parameter. Smaller bandwidths are employed to detect fine-grained differences among embeddings, beneficial for distributions with complex local structures, whereas larger bandwidths are used for capturing broader, global discrepancies.

The kernel matrix for the embeddings is formulated as Eq. 4.2,

$$\mathbb{H}_k = \{exp(-\frac{d_z}{\beta}) | \beta \in L_{BW}\}. \tag{4.2}$$

For each kernel $\mathscr{K} \in \mathbb{H}_k$, the intra-source differences are calculated by Eq. 4.3,

$$\text{Intra}(\mathbf{z}_j^{tr}, \mathbf{z}_b^{tr}, \mathbf{z}_v^{tr}) = \mathbb{E}\left[\sum_{\mathscr{K} \in \mathbb{H}_k} \mathscr{K}(\mathbf{z}_j^{tr}, \mathbf{z}_j^{tr})\right] + \mathbb{E}\left[\sum_{\mathscr{K} \in \mathbb{H}_k} \mathscr{K}(\mathbf{z}_b^{tr}, \mathbf{z}_b^{tr})\right] + \mathbb{E}\left[\sum_{\mathscr{K} \in \mathbb{H}_k} \mathscr{K}(\mathbf{z}_v^{tr}, \mathbf{z}_v^{tr})\right], \tag{4.3}$$

where $\mathbb{E}$ signifies the empirical mean, and $\mathbf{z}_j^{tr}$, $\mathbf{z}_b^{tr}$, and $\mathbf{z}_v^{tr}$ are the training embeddings from the three modalities, as demonstrated in Fig. 4.1. Conversely, the inter-source differences among modalities

are determined by Eq. 4.4,

$$\text{Inter}(\mathbf{z}_j^{tr}, \mathbf{z}_b^{tr}, \mathbf{z}_v^{tr}) = \mathbb{E}\left[\sum_{\mathscr{K} \in \mathbb{H}_k} \mathscr{K}(\mathbf{z}_j^{tr}, \mathbf{z}_v^{tr})\right] + \mathbb{E}\left[\sum_{\mathscr{K} \in \mathbb{H}_k} \mathscr{K}(\mathbf{z}_j^{tr}, \mathbf{z}_b^{tr})\right] + \mathbb{E}\left[\sum_{\mathscr{K} \in \mathbb{H}_k} \mathscr{K}(\mathbf{z}_b^{tr}, \mathbf{z}_v^{tr})\right]. \quad (4.4)$$

The CrossMMD loss, as defined in Eq. 4.5,

$$\text{CrossMMD}(\mathbf{z}_j^{tr}, \mathbf{z}_b^{tr}, \mathbf{z}_v^{tr}) = \text{Intra}(\mathbf{z}_j^{tr}, \mathbf{z}_b^{tr}, \mathbf{z}_v^{tr}) - \text{Inter}(\mathbf{z}_j^{tr}, \mathbf{z}_b^{tr}, \mathbf{z}_v^{tr}), \quad (4.5)$$

which is formulated to minimize intra-class variance and maximizes the inter-class variance, enabling efficient cross-modal information sharing at multi-scale levels through Gaussian kernels. In addition to the $\mathscr{L}_{MMD}$ loss, our training process incorporates the cross-entropy loss for each of the three modalities, as shown in Eq. 4.6,

$$\mathscr{L}_{\text{overall}} = \mathscr{L}_{\text{CE}}^{j} + \mathscr{L}_{\text{CE}}^{b} + \mathscr{L}_{\text{CE}}^{v} + \lambda \cdot \mathscr{L}_{\text{MMD}}, \quad (4.6)$$

where λ is a predetermined constant to ensure the gradients of the two types of loss are on the same scale. The terms $\mathscr{L}_{\text{CE}}^{j}$, $\mathscr{L}_{\text{CE}}^{b}$, and $\mathscr{L}_{\text{CE}}^{v}$ represent the cross-entropy losses for the joint, bone, and velocity branches, respectively.

4.2.3.2 Refinement of Logits Based on Cross-Modality Distance.

Employing the averaged logits from the three modalities, adjusted by CrossMMD, the model calculates a preliminary open-set probability using the maximum score derived from SoftMax applied to the logits. While this method enhances the accuracy of predicting open-set probabilities, we observed that relying solely on SoftMax for these predictions leads to inadequate differentiation between in- and out-of-distribution samples with respect to their open-set probability distributions, hindering further advancements in OS-SAR performance.

To address this challenge, we introduce the concept of Channel-Normalized Euclidean distance (CNE-distance). This approach allows for the creation of Gaussian-like probability distributions, facilitating a clearer separation between in- and out-of-distribution samples. Initially, embeddings for the training samples across the three modalities are extracted, yielding the embedding sets Ω_a^{tr}, where a encompasses j, b, v. This extraction process is replicated for the test sample embeddings, denoted as $\mathbf{z}_a^{te}$. For every test sample, three distances are computed relative to the closest training set embedding within Ω_a^{tr}, specifically d_j, d_b, and d_v. The process begins with L2 normalization applied to each embedding along the channel dimension, ensuring the feature values are scaled between 0 and 1. Subsequently, the Euclidean distance to the nearest training set embedding is calculated,

serving as the measure for open-set probability. Thus, the CNE-distance is formulated as in Eq. 4.7,

$$d_j, d_b, d_v = D\left[\mathscr{N}_C(\mathbf{z}_j^{te}), \mathscr{N}_C(\Omega_j^{tr})\right], D\left[\mathscr{N}_C(\mathbf{z}_b^{te}), \mathscr{N}_C(\Omega_b^{tr})\right], D\left[\mathscr{N}_C(\mathbf{z}_v^{te}), \mathscr{N}_C(\Omega_v^{tr})\right], \tag{4.7}$$

where te and tr refer to the test and training datasets respectively, and $D[\cdot]$ represents the Euclidean distance. $\mathscr{N}_C(\cdot)$ denotes channel normalization. The mean distance is then derived as per Eq. 4.8,

$$d_m = Mean(d_j, d_b, d_v). \tag{4.8}$$

Our experiments reveal the effectiveness of the CNE-distance in producing more reliable probability estimates under open-set conditions, especially when differentiating between in- and out-of-distribution samples. Yet, when using the CNE-distance to determine the class among the known classes, as in Fig. 4.1, the close-set accuracy are sub-optimal.

To address this, we introduce a novel refinement methodology. This approach refines the averaged logits utilizing the CNE-distance, addressing the disparities among modalities and improving the close-set classification. By incorporating the averaged CNE-distances among modalities, our method seeks to strike a balance between effective open-set probability estimation and good close-set classification. We first acquire the position with the highest logit value of the averaged logits by Eq. 4.9,

$$M_P = ArgMax((l_j + l_b + l_v)/3), \tag{4.9}$$

where l_j, l_b, and l_v denote the predicted logits for joints, bones, and velocities branches through classification heads. Then we refine the predicted averaged logits l_m by using Eq. 4.10 considering a given sample, where the salient logit position is indicated by a one-hot mask M_P. We use this formula to achieve a separated calibration of the predicted categories and the none-predicted categories, where the distribution of the confidence score of the seen categories and the unseen categories can be well disentangled.

$$l_m[M_P] := Log(exp(l_m[M_P] * d_m^2)(\frac{1}{d_m} - 1)). \tag{4.10}$$

While the not salient positions are indicated by mask M_{NP}, the not saliency logits are refined by Eq. 4.11,

$$l_m[M_{NP}] := l_m[M_{NP}] * d_m^2. \tag{4.11}$$

Then, we get the refined full logits l_m, which will be passed through SoftMax further to get the classification and the open-set probability. The final predicted open-set probability is as follows,

$$P_{prob} = Max(SoftMax(l_m)), \tag{4.12}$$

Table 4.1: Experiments on NTU-60/120 datasets, where CS, CV, and B indicate **C**ross-**S**ubject/**V**iew evaluations and **B**ackbone. The results are averaged for five random splits. RPL [25], ARPL [24], PMAL [127], the vanilla SoftMax score (SoftMax) [74], Monte Carlo Dropout + Voting (MCD-V) [155], DEAR [10], and Humpty [53] are chosen as open-set baselines to construct the benchmark.

B	Method	NTU-60						NTU-120					
		O-AUROC		O-AUPR		C-ACC		O-AUROC		O-AUPR		C-ACC	
		CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV
CTRGCN	SoftMax	83.68	87.77	67.37	76.38	90.56	93.83	82.37	83.10	91.84	91.88	90.37	91.04
	RPL	84.02	88.06	67.86	76.75	90.82	95.38	82.06	83.40	91.55	92.05	90.40	90.96
	ARPL	84.13	88.37	68.24	76.58	91.00	95.45	81.93	83.03	91.54	91.80	90.12	91.16
	PMAL	82.72	88.06	64.99	73.31	90.74	95.09	80.46	81.75	90.55	90.93	89.61	90.14
	DEAR	83.11	87.54	63.07	75.52	84.14	95.41	81.98	82.66	91.51	91.67	90.11	90.61
	Humpty	82.08	85.82	62.05	69.09	89.17	93.75	82.12	83.35	90.78	91.06	89.89	90.54
	MCD-V	81.31	85.58	61.88	69.99	90.14	94.72	78.83	79.17	89.60	76.93	88.12	88.10
	Ours	**90.62**	**94.14**	**80.32**	**88.07**	**93.68**	**97.51**	**85.44**	**85.42**	**93.67**	**93.36**	**91.43**	**92.94**
HDGCN	SoftMax	81.52	86.95	63.62	73.89	89.14	94.67	81.34	82.90	91.49	91.83	89.92	90.21
	RPL	82.92	88.38	66.06	76.27	91.92	95.32	82.00	83.05	91.59	91.83	89.77	90.77
	ARPL	83.92	87.19	67.76	74.49	90.65	94.90	82.06	82.80	91.51	91.74	90.08	90.68
	PMAL	82.41	83.57	64.53	66.98	90.26	93.33	80.68	81.89	90.71	91.22	89.53	90.75
	DEAR	83.87	87.92	67.76	76.15	90.65	95.15	81.89	82.78	91.38	91.63	89.85	90.68
	Humpty	81.91	87.47	61.49	71.32	88.70	94.64	82.38	83.26	90.72	85.78	89.40	89.93
	MCD-V	82.51	86.74	64.24	72.70	90.04	94.88	80.55	80.24	90.26	90.27	89.80	89.00
	Ours	**89.57**	**93.14**	**78.82**	**86.48**	**93.30**	**96.88**	**83.76**	**84.46**	**92.84**	**93.07**	**90.82**	**91.67**
HyperFormer	SoftMax	83.40	87.11	66.29	74.38	90.46	94.90	81.16	82.74	91.40	91.60	90.69	90.95
	RPL	79.97	83.96	60.15	68.52	88.39	92.46	81.26	82.20	91.19	91.30	89.65	90.31
	ARPL	82.37	84.88	64.38	69.74	89.87	93.99	82.08	82.06	91.25	91.53	90.19	90.46
	PMAL	82.43	85.80	64.29	70.89	90.33	94.79	81.95	81.90	91.63	89.13	90.65	90.42
	DEAR	81.47	85.22	62.87	70.33	89.94	94.26	81.00	81.90	90.96	91.15	89.51	90.15
	Humpty	71.72	73.66	55.21	60.95	89.98	94.67	70.67	69.28	86.93	86.29	89.92	89.40
	MCD-V	82.52	79.69	65.00	59.46	93.05	88.80	80.21	81.17	90.24	90.64	88.87	89.80
	Ours	**88.98**	**92.73**	**77.75**	**85.94**	**93.24**	**96.71**	**83.67**	**83.70**	**92.84**	**92.62**	**91.30**	**92.50**

while the open-set novelty score can be obtained by $1 - P_{prob}$. By using this refinement method, the accurately predicted class from the SoftMax score computed on averaged logits can be preserved while the predicted open-set probability can achieve a distance-controllable disentanglement. This disentanglement ability benefits the OS-SAR a lot, as observed in our experiments. We refer to our full pipeline combining CrossMMD and the proposed distance-based refinement as CrossMax.

4.3 Experiments

4.3.1 Metrics

To assess open-set performance, we employ the Area Under the Receiver Operating Characteristic (O-AUROC) and the Area Under the Precision-Recall curve (O-AUPR), which provide insights into

Table 4.2: Experiments on Toyota Smart Home [43] dataset, where CS and CV indicate **C**ross-**S**ubject/**V**iew evaluations. The results are averaged for five random splits.

Method	Toyota Smart Home																	
	O-AUROC		O-AUPR		C-ACC		O-AUROC		O-AUPR		C-ACC		O-AUROC		O-AUPR		C-ACC	
	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV
	CTRGCN						HDGCN						Hyperformer					
SoftMax	70.04	65.18	70.10	69.02	70.41	78.52	72.88	54.47	71.16	61.07	78.37	75.10	74.25	72.26	74.30	74.94	78.68	81.40
RPL	56.74	51.90	60.46	59.46	74.42	75.41	74.26	61.93	73.97	63.61	78.35	77.02	73.24	74.30	72.62	75.84	78.62	82.23
ARPL	74.11	64.22	73.80	67.04	78.55	79.53	73.00	64.53	72.93	68.73	78.75	77.62	72.73	72.77	72.99	73.98	78.60	82.67
PMAL	57.80	51.73	61.27	52.94	74.06	67.50	64.64	74.72	69.20	73.41	77.23	78.39	73.48	51.89	73.68	47.26	78.01	69.97
DEAR	76.19	60.54	75.42	74.52	78.49	65.50	75.03	59.25	75.10	63.41	78.41	78.54	72.86	74.54	72.70	76.09	78.20	82.87
Humpty	65.10	59.17	68.71	62.43	77.76	75.19	62.41	57.12	66.78	66.32	77.83	80.12	72.32	62.70	71.26	62.88	78.32	80.23
MCD-V	69.61	67.92	71.12	71.68	77.74	76.41	72.29	64.64	72.57	69.20	79.90	78.93	61.69	53.71	65.17	62.70	74.15	48.20
Ours	**83.99**	**84.00**	**86.74**	**87.37**	**80.25**	**80.51**	**84.32**	**83.70**	**86.57**	**86.44**	**80.41**	**81.29**	**82.23**	**80.76**	**84.28**	**81.46**	**79.58**	**83.54**

the model's performance with varying focus on category balance. Close-set classification accuracy (C-ACC) is also measured to determine if the open-set method maintains good close-set classification performance. Following PMAL [127], O-AUROC and C-ACC are primary metrics, with O-AUPR added due to the imbalance in the ToyotaSmartHome dataset.

4.3.2 Implementation Specifications

Our model is implemented using PyTorch1.8.0, trained with an SGD optimizer (learning rate of 0.1), a step-wise learning rate scheduler (decay rate of 0.1 at steps $35, 55, 70$), a weight decay of 0.0004, and a batch size of 64 over 100 epochs on four Nvidia A100 GPUs with an Intel Xeon Gold 6230 processor. Parameters λ, N_k, and α are set to 0.1, 5, and 2.0, respectively, resulting in model sizes of 4.29 MB for CTRGCN, 5.04 MB for HDGCN, and 7.8 MB for Hyperformer.

4.3.3 Benchmark Insights

An extensive analysis of existing open-set recognition methods within our benchmark reveals principal point distance-based methods like RPL [25] and ARPL [24] offer good O-AUROC improvements on CTRGCN and HDGCN compared to SoftMax [74] on the NTU-60 (CS). Yet, their performance lags on the Hyperformer backbone for NTU-60, highlighting the challenge of cross-backbone adaptability, as shown in Table 4.1. Examining cross-dataset adaptability, RPL and ARPL show superiority over SoftMax using HDGCN and Hyperformer on NTU-120 but underperform on the Toyota Smart Home (CS) with CTRGCN as shown in Table 4.2. The prototypical learning method PMAL [127] generally falls short in OS-SAR task. Similarly, video-based open-set approaches like DEAR [10], MCD-V [155], and Humpty [53] demonstrate limited efficacy on OS-SAR task, likely due to the significant differences between skeletal and RGB image/video data, as well as the sparse character of skeletal data. The reliance on GCNs rather than CNNs for feature extraction in skeletal data suggests the need for a specific OS-SAR method that is effective across various datasets and

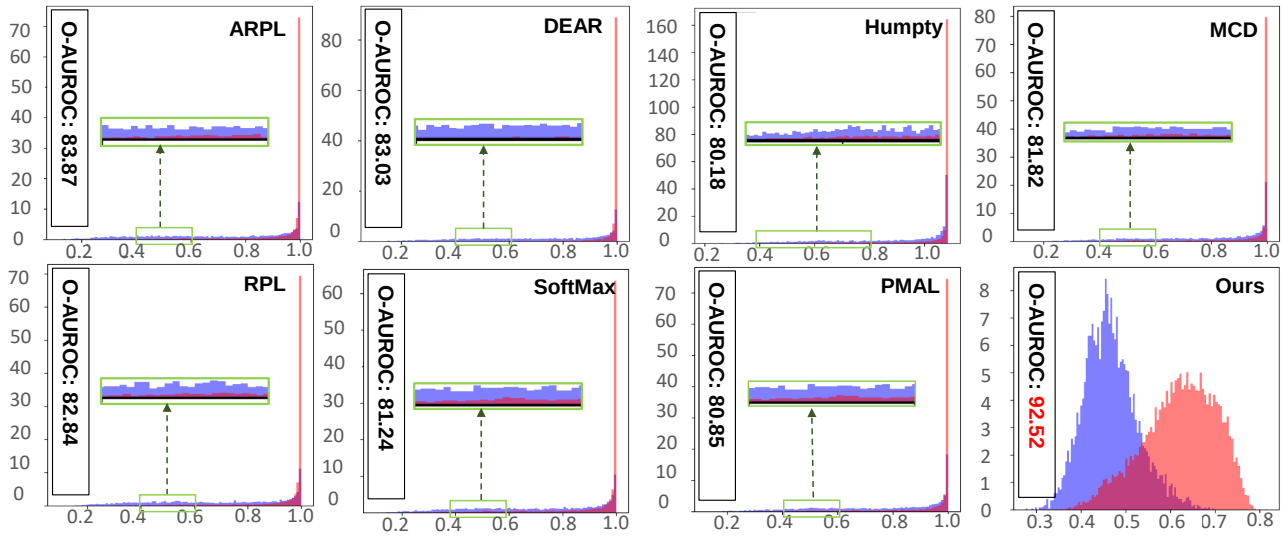


Figure 4.2: A comparison of open-set probability estimated using HD-GCN on the NTU-60 (CS) dataset across one randomly selected split.

backbones.

To handle the underlying issue for existing open-set recognition approaches, we analyze the disentanglement capability between the in- and out-of-distribution samples considering the open-set probability in Fig. 4.2, where open-set probability tends to 1.0 when the prediction is quite certain. We observe that most of the baselines can not well disentangle in- and out-of-distribution samples according to their predicted open-set probabilities, which serves as a critical reason for the undesired performance on OS-SAR. Keeping this issue in mind, we propose CrossMax by using CrossMMD in the training phase and cross-modality distance-based logits refinement in the test phase. CrossMax delivers superior disentanglement in terms of the open-set probability considering the in- and out-of-distribution samples. CrossMax achieves 6.94%, 8.05%, and 5.58% O-AUROC improvements and 12.95%, 15.20%, and 11.46% O-AUPR improvements on CTRGCN, HDGCN, and Hyperformer backbones within NTU-60 cross-subject evaluation compared with vanilla SoftMax, while consistent performances can be found for different backbones, datasets, and settings, demonstrating the importance of the superior disentanglement ability for open-set probability between in- and out-of-distribution samples.

4.3.4 Analysis of Observations and Ablations

4.3.4.1 Advantages of Implementing CrossMMD

Through t-SNE visualizations comparing models trained without CrossMMD (labeled as Ensemble) and with CrossMMD, as shown in Fig. 4.3, it's evident that CrossMMD significantly enhances

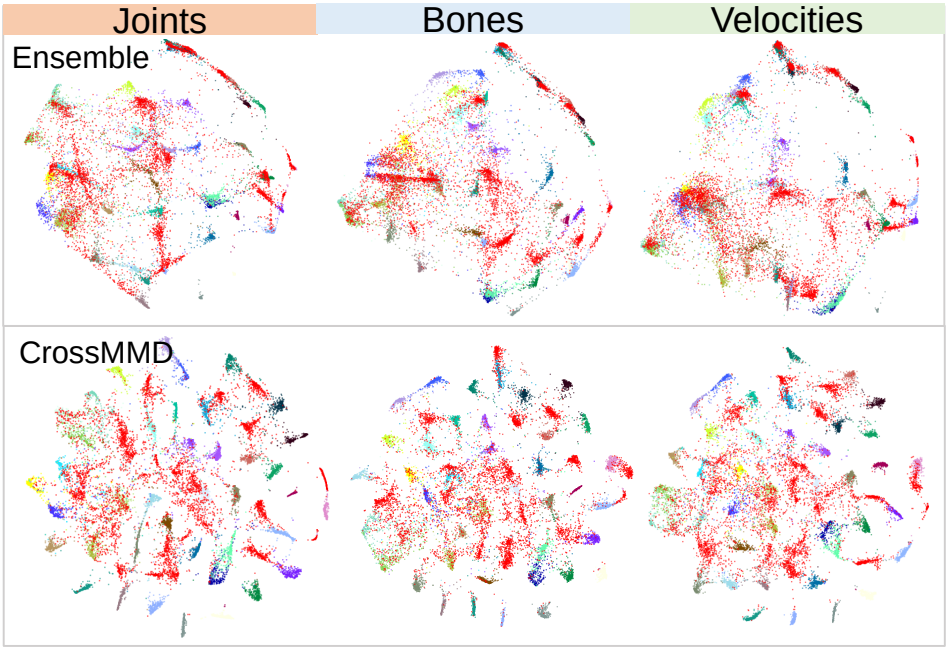


Figure 4.3: T-SNE [130] visualizations on NTU-60 (CS) using CTRGCN. Out- and in-of-distribution samples are marked by red and other colors.

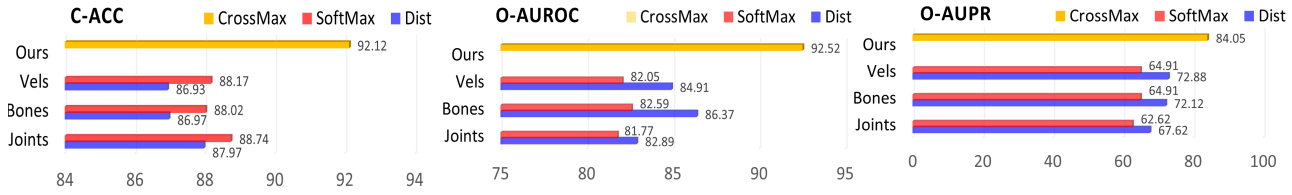


Figure 4.4: Comparison of SoftMax, CNE-distance, and CrossMax using HDGCN on NTU-60 (CS) on one split.

the discriminative ability and structure of latent spaces for both in- and out-of-distribution samples. This improvement aligns with the performance gains detailed in Table 4.3, where both Ensemble and CrossMMD configurations utilize the SoftMax score for estimating open-set probabilities.

4.3.4.2 CNE-distance Versus Conventional SoftMax

Comparative analysis reveals that the open-set recognition performance on the corresponding metrics, *i.e.*, O-AUROC and O-AUPR, are substantially higher for the CNE-distance across all modalities than for the conventional SoftMax without logits refinement, as demonstrated in Fig. 4.4. The CNE-distance method surpasses the vanilla SoftMax by 2.86%, 3.78%, and 1.12% in O-AUROC for joints, bones, and velocities, respectively. Despite its strengths, the CNE-distance method falls short in ensuring satisfactory performance in close-set classification accuracy. The proposed logits refinement method leverages the strengths of both the vanilla SoftMax and CNE-distance to enhance both

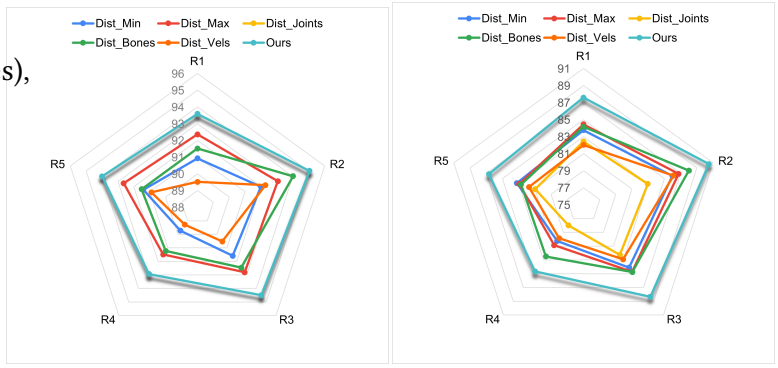
Table 4.3: Module ablation on NTU-60 (CS) on CTRGCN, where the results are averaged among five random splits.

Method	O-AUROC	O-AUPR	C-ACC
Ensemble	86.23	71.35	93.31
CrossMMD (Ours)	88.31	74.80	93.68
CrossMax (Ours)	90.62	80.32	93.68

open- and close-set recognition performances.

4.3.4.3 Logits Refinement Versus CNE-distance Evaluation

Addressing inquiries about the superiority of the proposed cross-modality distance-based logits refinement over CNE-distance variations in predicting open-set probabilities, we present a comparative analysis in Fig. 4.5 from five random splits (R1 to R5). The comparison includes CNE-distance variations based on joint-modality (Dist_joints), bone-modality (Dist_bones), velocity-modality (Dist_velocities), and both minimum (Dist_min) and maximum (Dist_max) aggregation across modalities. A regular pentagon shape in the curve signifies consistent and generalizable performance across different splits. The logits refinement approach exhibits the most stable and superior performance among all, indicating that this method not only provides more stable estimations of open-set probabilities but also maintains exceptional close-set classification performance.



(a) O-AUROC **(b)** O-AUPR

Figure 4.5: Comparison of open-set recognition performances using CTRGCN backbone on NTU-60 cross-view evaluation for five different random splits.

4.3.4.4 Ablation of Each Module.

We use CrossMMD during training while using cross-modality distance-based logits refinement during test. We show the benefits from different modules in Table 4.3, where Ensemble indicates using ensembled modalities and vanilla SoftMax, CrossMMD indicates using CrossMMD and vanilla SoftMax, and CrossMax indicates using CrossMMD and the proposed logits refinement method. CrossMMD achieves 2.08%, 3.45%, and 0.37% improvements for O-AUROC, O-AUPR, and C-ACC

Table 4.4: Ablation study for OS-SAR under different open-set ratios using CTRGCN on NTU-60 [161] dataset for cross-view and cross-subject evaluations, where the results are averaged on five random splits.

Method	O-AUROC		O_AUPR		C-ACC		O-AUROC		O_AUPR		C-ACC	
	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV
	❶ Case1.						❷ Case2.					
SoftMax	83.68	87.77	67.37	76.38	90.56	93.83	83.10	85.58	91.54	96.05	**95.10**	95.76
RPL	84.02	88.06	67.86	76.75	90.82	95.38	83.72	87.58	95.34	96.51	92.20	95.54
ARPL	84.13	88.37	68.24	76.58	91.00	95.45	83.72	87.52	95.34	96.49	92.20	95.49
DEAR	83.11	87.54	63.07	75.52	84.14	95.41	83.00	86.13	95.00	96.15	91.30	95.28
Ours	**90.62**	**94.14**	**80.32**	**88.07**	**93.68**	**97.51**	**94.61**	**96.20**	**98.63**	**99.04**	94.17	**96.90**

compared with the ensemble variant, while CrossMax preserves the superior C-ACC of CrossMMD and delivers improvements by 2.31% and 5.52% of O-AUROC and O-AUPR, showing the importance of using both.

4.3.4.5 Ablation for Different Open-Set Ratios

In this ablation study, we evaluate the performance of four leading models, *i.e.*, SoftMax [74], RPL [25], ARPL [24], DEAR [53], and our own model, CrossMax, on the NTU-60 dataset, using CTRGCN as the backbone framework. The results, presented in Table 4.4, cover both cross-subject and cross-view scenarios under varying open-set conditions. Specifically, *Case1* utilizes 40 classes for training as known classes and designates 20 classes as unknown, whereas *Case2*, which presents a more challenging scenario, trains on only 10 classes and leaves 50 as unseen due to the significantly reduced prior knowledge available for training. The outcome reveals a minor decline in O-AUROC scores for the baseline models in *Case2*, yet CrossMax maintains robust performance across different open-set ratios, marking improvements of 10.89% and 8.68% over ARPL in O-AUROC for cross-subject and cross-view evaluations, respectively. Moreover, when compared to its performance in *Case1*, CrossMax exhibits gains of 3.99% and 2.06% in O-AUROC and 18.31% and 10.97% in O-AUPR for the two evaluation scenarios, respectively, underscoring its exceptional capability in handling challenging open-set contexts. Additionally, the distribution of predicted open-set probabilities for both in- and out-of-distribution samples, shown in Fig. 4.6, highlights CrossMax's superior disentanglement capability in terms of the open-set probability on seen and unseen categories on a different open-set ratio. Performance stability is further evidenced by consistent top-tier results across different random open-set splits (R1 to R5), as detailed in Fig. 4.7.

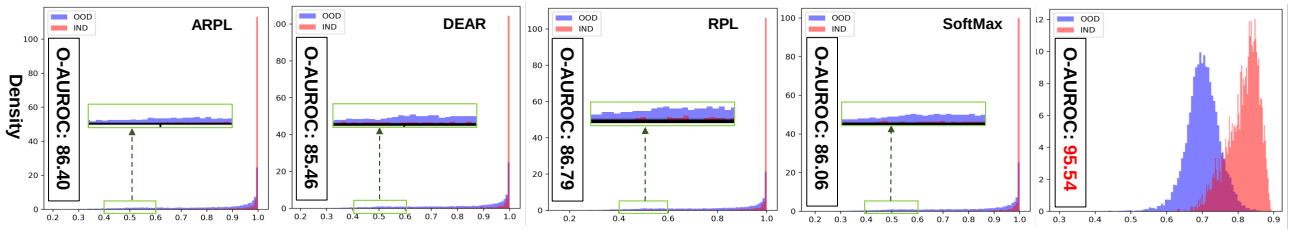
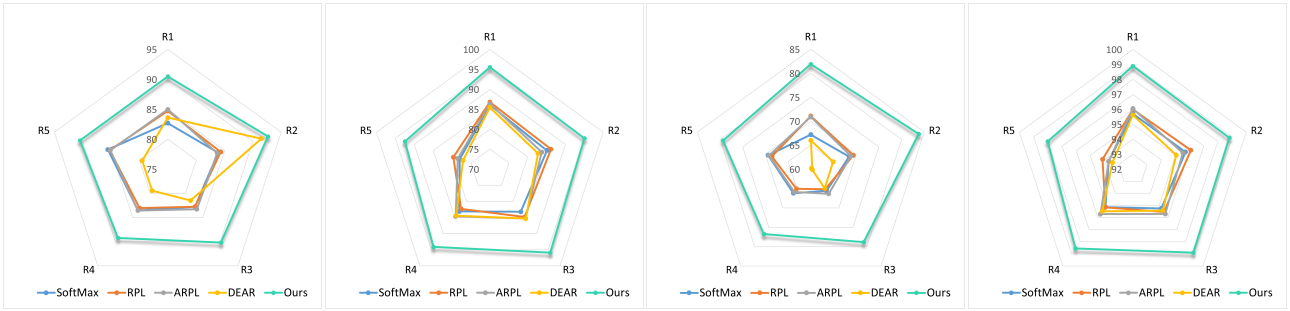


Figure 4.6: A Visualization of the open-set probability comparison using the CTRGCN [33] backbone on the NTU-60 [161] dataset for a cross-subject evaluation on Run1 for a specific open-set ratio scenario (Case2).



(a) O-AUROC for Case1 **(b)** O-AUROC for Case2 **(c)** O-AUPR for Case1 **(d)** O-AUPR for Case2

Figure 4.7: Comparison of open-set recognition performances using CTRGCN [33] backbone on NTU-60 [161] cross-subject evaluation for five different random splits on two different open-set ratios (case 1 and case 2), where R1 to R5 indicates the five random splits.

Table 4.5: Ablation study for OS-SAR under Gaussian noise disturbance using CTRGCN on NTU-60 [161] dataset for cross-view and cross-subject evaluations, where the results are averaged on five random splits.

Method	O-AUROC		O_AUPR		C-ACC		O-AUROC		O_AUPR		C-ACC	
	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV
	❶ **Without Noise.**						❷ **Gaussian Noise Disturbance.**					
SoftMax	83.68	87.77	67.37	76.38	90.56	93.83	72.76	76.36	52.04	57.63	56.44	57.68
RPL	84.02	88.06	67.86	76.75	90.82	95.38	74.21	77.79	53.31	59.49	71.74	75.36
ARPL	84.13	88.37	68.24	76.58	91.00	95.45	74.99	78.66	54.99	60.55	82.76	88.26
DEAR	83.11	87.54	63.07	75.52	84.14	95.41	73.65	76.83	52.51	57.32	82.14	86.63
Ours	**90.62**	**94.14**	**80.32**	**88.07**	**93.68**	**97.51**	**79.94**	**83.36**	**66.03**	**73.91**	**85.93**	**89.31**

4.3.4.6 Ablation for Noise Disturbance

In this ablation study, we explore the performance of our CrossMax model along with four prominent baselines, *i.e.*, SoftMax, DEAR, ARPL, and RPL, on the NTU-60 dataset [161], employing CTRGCN [33] as the feature extraction backbone, under conditions of Gaussian noise perturbation, as summarized in Table 4.5. This investigation follows a similar methodology to our prior ablation study, focusing on the same leading baselines. A skeletal sequence is represented as $\mathbf{s} \in \mathbb{R}^{3\times T\times N_j}$. Gaussian noise is then introduced, denoted by $\mathbf{n} \in \mathbb{R}^{3\times T\times N_j}$, generated from a normal distribution.

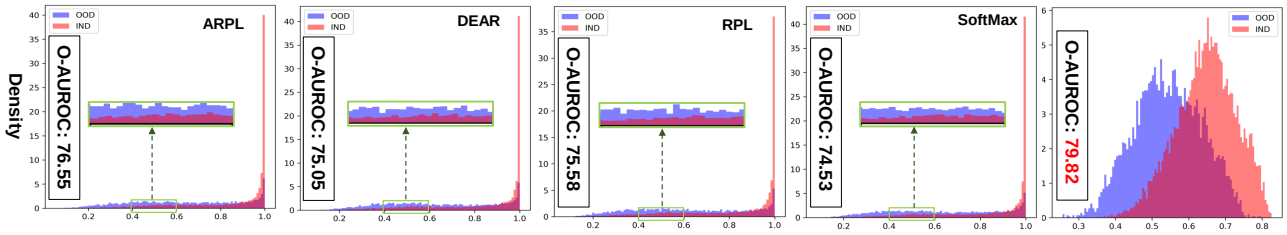
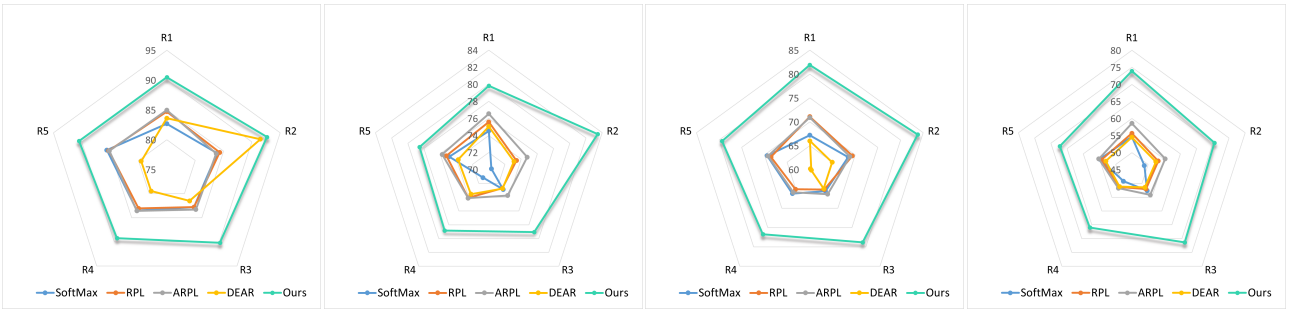


Figure 4.8: Comparison of the open-set probabilities using CTRGCN [33] as the backbone on NTU-60 [161] for cross-subject evaluation for Run1 under Gaussian noise disturbance.



(a) O-AUROC w/o Noise **(b)** O-AUROC w/ Noise **(c)** O-AUPR w/o Noise **(d)** O-AUPR w/ Noise

Figure 4.9: Comparison of open-set recognition performances using CTRGCN [33] backbone on NTU-60 [161] cross-subject evaluation for five different random splits for w/ noise and w/o noise scenarios, where R1 to R5 indicates the five random splits.

The noise-interrupted skeletal sequence is expressed as $\mathbf{s}_n = \mathbf{s} + \gamma * \mathbf{n}$, with γ set to 0.3. This Gaussian noise is applied to both the training and testing datasets.

The introduction of Gaussian noise leads to a noticeable decrease in performance across all models, indicating the adverse impact of such disturbances on OS-SAR results. Despite this, CrossMax demonstrates a relatively minor decline in effectiveness, maintaining its lead in performance amidst noise challenges. The method continues to achieve superior results under noise conditions. Furthermore, we present the predicted probabilities for both in- and out-of-distribution samples in the face of noise in Fig. 4.8, where CrossMax's exceptional ability to disentangle these sample types under various open-set ratios is evident. The model's performance across various random splits (R1 to R5) is detailed in Fig. 4.9, where CrossMax consistently provides stable and leading performance across different open-set divisions for OS-SAR tasks on the NTU-60 dataset [161], utilizing the CTRGCN [33] backbone for cross-subject evaluations.

4.3.4.7 Ablation under Occlusions

This ablation presents the outcomes of the OS-SAR experiments conducted with random occlusions on the NTU-60 dataset [161], utilizing CTRGCN [33] as the feature extraction backbone, as

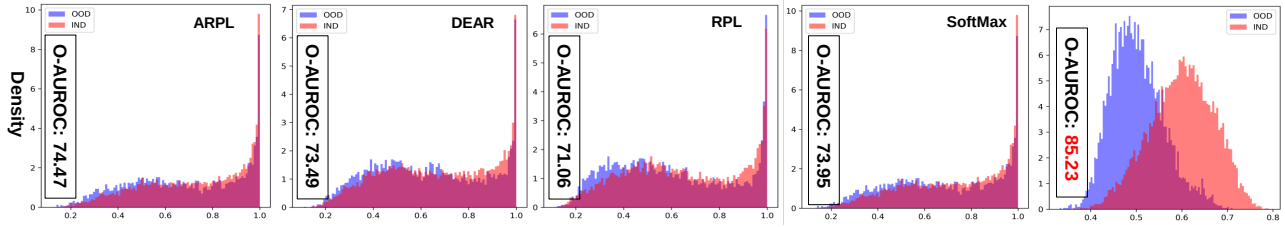
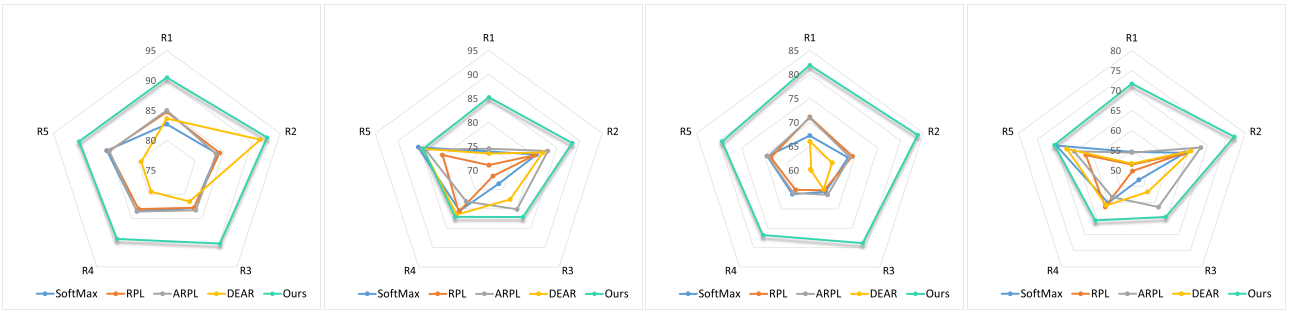


Figure 4.10: Comparison of the open-set probabilities using CTRGCN [33] as the backbone on NTU-60 [161] for cross-subject evaluation for Run1 under random occlusion disturbance.



(a) O-AUROC w/o Occlusion **(b)** O-AUROC w/ Occlusion **(c)** O-AUPR w/o Occlusion **(d)** O-AUPR w/ Occlusion

Figure 4.11: Comparison of open-set recognition performances using CTRGCN [33] backbone on NTU-60 [161] cross-subject evaluation for five different random splits for w/ occlusion and w/o occlusion scenarios, where R1 to R5 indicates the five random splits.

Table 4.6: Ablation study for OS-SAR under random occlusion disturbance using CTRGCN on NTU-60 [161] dataset for cross-view and cross-subject evaluations, where results are averaged on five random splits.

Method	O-AUROC		O_AUPR		C-ACC		O-AUROC		O_AUPR		C-ACC	
	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV	CS	CV
	❶ Without Occlusion.						❷ With Random Occlusions.					
SoftMax	83.68	87.77	67.37	76.38	90.56	93.83	77.34	80.09	58.88	63.67	67.32	71.24
RPL	84.02	88.06	67.86	76.75	90.82	95.38	76.72	78.96	57.71	61.63	74.39	78.97
ARPL	84.13	88.37	68.24	76.58	91.00	95.45	79.92	79.55	61.55	63.35	87.18	88.01
DEAR	83.11	87.54	63.07	75.52	84.14	95.41	79.79	81.45	60.45	65.23	87.44	90.75
Ours	**90.62**	**94.14**	**80.32**	**88.07**	**93.68**	**97.51**	**84.44**	**86.30**	**69.89**	**74.07**	**88.69**	**92.66**

detailed in Table 4.6. To simulate random occlusions, we applied a random occlusion rate θ chosen from the set $\{10\%, 20\%, 30\%\}$, setting θ percentage of the coordinates in a skeleton sequence to zero. This method emulates the effects of random occlusion, which poses significant challenges for OS-SAR by inducing geometric discontinuities and exacerbating the sparsity of skeletal data.

Comparative analyses, against experiments without occlusion mentioned in Table 4.6, reveal a marked decline in the performance of all baseline models when faced with occluded data. This decline underscores the detrimental impact of occlusions on OS-SAR model efficacy. An examination

of the predicted open-set probabilities for in- and out-of-distribution samples, illustrated in Fig. 4.10, demonstrates increased overlap between these sample types under occlusion conditions. Nonetheless, our CrossMax model retains its excellent capability for disentangling these sample categories.

In terms of quantitative performance, CrossMax achieves scores of 84.44%, 69.89%, and 88.69% for O-AUROC, O-AUPR, and C-ACC, respectively, in the cross-subject evaluation. For the cross-view evaluation, it records 86.30%, 74.07%, and 92.66% for O-AUROC, O-AUPR, and C-ACC, respectively. Performance metrics under random occlusion across different data splits are showcased in Fig. 4.11. Here, while all considered OS-SAR baseline models display significant performance variations, CrossMax demonstrates the most consistent and superior performance across these evaluations.

Table 4.7: Comparison with our implemented MM-ARPL on NTU-60 [161] cross-subject evaluation on CTRGCN backbone, where the results are averaged among five random splits.

Method	O-AUROC	O-AUPR	C-ACC
SoftMax [74]	83.68	67.37	90.56
ARPL [24]	84.13	73.27	91.00
Ensemble (MM-SoftMax)	86.23	71.35	93.31
MM-ARPL	87.60	73.27	93.67
CrossMMD (Ours)	88.31	74.80	93.68
CrossMax (Ours)	**90.62**	**80.32**	**93.68**

4.3.4.8 Comparison with ARPL under Three Modalities

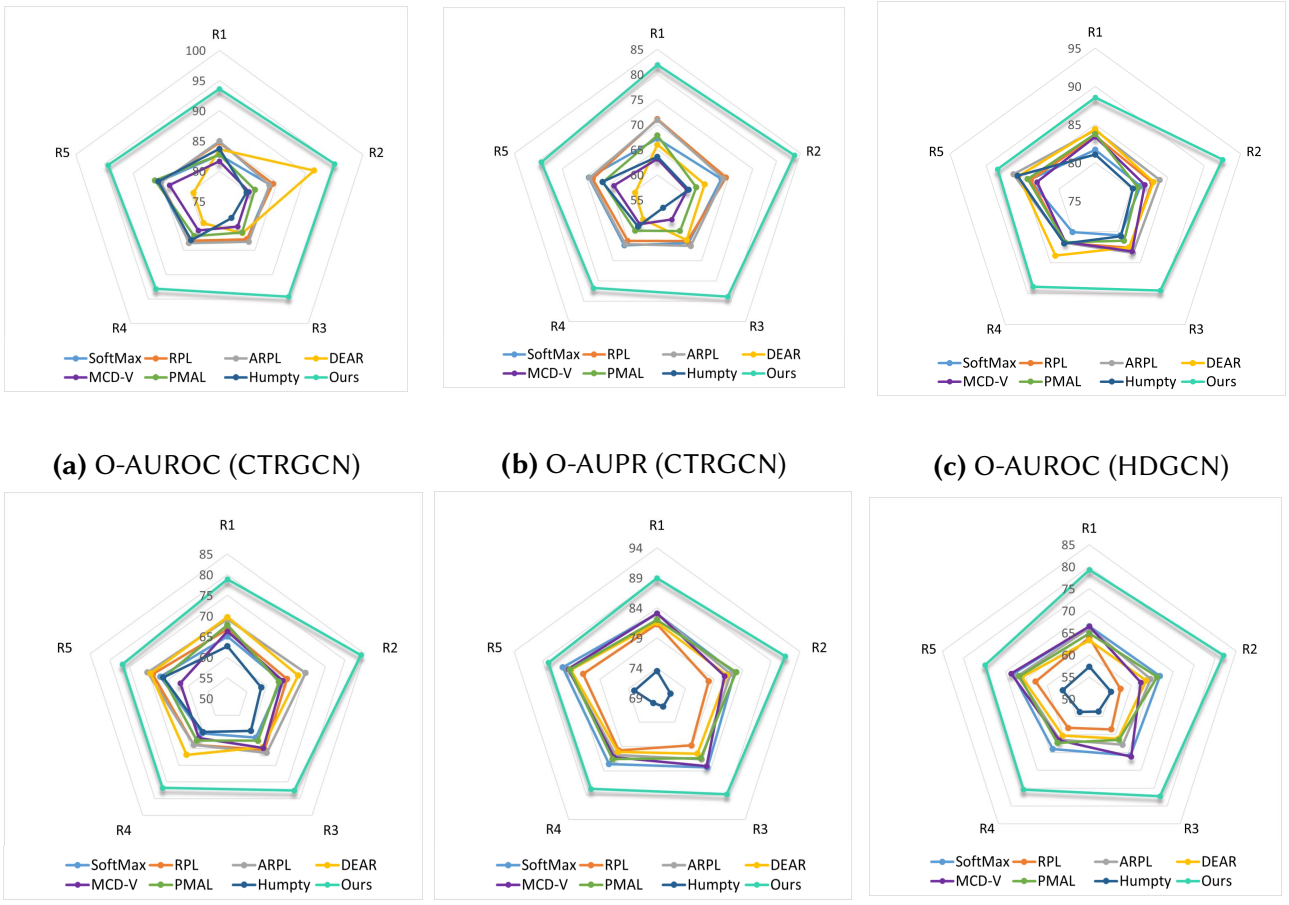
In this evaluation, we delve into the performance comparison of our CrossMax method against the top-performing baseline, ARPL [24], within a multi-modality framework on the NTU-60 dataset [161], using CTRGCN [33] as the backbone for cross-subject analysis. Due to ARPL's exemplary performance with the CTRGCN architecture on the NTU-60 dataset, it has been selected to accomplish a multi-modality variant, where ARPL is trained separately across three modalities, *i.e.*, joints, bones, and velocities, with the outputs averaged prior to calculating the final open-set probability. This multi-modality version of ARPL is referred to as MM-ARPL in Table 4.7. Additionally, the performances of a multi-modal adaptation of SoftMax, labeled as Ensemble, alongside the standard versions of SoftMax [74], and ARPL [25] are documented in Table 4.7.

Compared to the standard ARPL model, the MM-ARPL variant exhibits improvements of 3.47%, 5.03%, and 2.67% in O-AUROC, O-AUPR, and C-ACC metrics, respectively. These enhancements highlight the significance of incorporating multiple modalities in OS-SAR task. Moreover, our CrossMax method not only continues to outperform MM-ARPL by margins of 3.02% and 7.05% in O-AUROC and O-AUPR, respectively, but also demonstrates advantageous close-set performance (C-

ACC), underscoring our approach's superior design in addressing open-set challenges. It is worth noting that within our benchmark, O-AUROC and O-AUPR are prioritized as key indicators of open-set model performance, whereas C-ACC serves as a secondary metric for evaluating close-set classification efficacy.

4.3.4.9 Stability for Different Splits across Backbones

This section discusses the results of OS-SAR experiments conducted using various data splits on the NTU-60 dataset for cross-subject evaluation, comparing different models across backbones, as depicted in Fig. 4.12. The evaluations focus on O-AUROC and O-AUPR metrics. Our model, CrossMax, consistently outperforms other methods, achieving the highest scores while displaying consistent performance across all splits, indicating its superior generalization capabilities.



(a) O-AUROC (CTRGCN) **(b)** O-AUPR (CTRGCN) **(c)** O-AUROC (HDGCN)

(d) O-AUPR (HDGCN) **(e)** O-AUROC (Hyperformer) **(f)** O-AUPR (Hyperformer)

Figure 4.12: Experimental results for all five random splits on NTU-60 [161] dataset under cross-subject evaluation on HDGCN [110].

4.3.5 Evaluation Protocols

In this section, more details of the evaluation protocols are introduced. The challenge level of OS-SAR varies with different data splits. To ensure equitable comparisons, we generate five random splits for each dataset, maintaining consistent seen and unseen class divisions across comparisons.

For the NTU-60 dataset, the classes designated as unseen are listed in Table 4.8, according to the NTU-60 [161] class index scheme. The remaining classes are considered seen. The NTU-120 dataset follows a similar approach, with seen classes detailed in Table 4.9, based on the NTU-120 [114] class index system, and the remaining classes categorized as unseen. For the Toyota Smart Home dataset, unseen classes are enumerated in Table 4.10, with the rest deemed seen.

Table 4.8: Unseen classes for five random splits on NTU-60 [161] dataset.

NTU-60	Unseen classes
Run1	50, 40, 30, 37, 12, 48, 45, 49, 8, 29, 58, 13, 1, 39, 27, 47, 14, 52, 3, 44
Run2	41, 21, 52, 6, 12, 36, 24, 56, 35, 57, 15, 26, 39, 53, 19, 4, 27, 25, 17, 47
Run3	46, 10, 47, 39, 55, 14, 58, 53, 13, 40, 24, 9, 45, 23, 27, 3, 7, 54, 33, 17
Run4	21, 55, 11, 43, 41, 3, 52, 39, 46, 59, 47, 15, 17, 54, 40, 33, 9, 38, 31, 57
Run5	56, 14, 17, 7, 40, 52, 37, 50, 36, 6, 44, 11, 41, 9, 47, 24, 53, 2, 10, 58

Table 4.9: Seen classes for five random splits on NTU-120 [114] dataset.

NTU-120	Seen classes
Run1	0, 37, 52, 70, 96, 92, 91, 4, 39, 12, 46, 81, 87, 31, 72, 48, 16, 62, 42, 102, 112, 68, 56, 49, 22, 11, 88, 107, 93, 43
Run2	17, 90, 47, 80, 79, 48, 27, 82, 61, 53, 96, 117, 62, 35, 23, 85, 8, 98, 104, 77, 51, 75, 56, 105, 54, 25, 18, 44, 40, 109
Run3	76, 9, 57, 59, 5, 51, 83, 104, 73, 27, 92, 72, 42, 111, 100, 67, 105, 4, 101, 12, 84, 119, 15, 33, 78, 62, 82, 24, 65, 108
Run4	48, 12, 26, 63, 20, 109, 80, 33, 79, 67, 100, 6, 24, 11, 76, 61, 10, 59, 0, 99, 19, 4, 90, 58, 28, 88, 44, 95, 72, 18
Run5	45, 0, 44, 13, 100, 14, 32, 72, 101, 17, 39, 63, 20, 56, 105, 71, 78, 73, 8, 99, 19, 115, 23, 54, 12, 109, 15, 37, 88, 18

4.4 Discussion

4.4.1 Societal Contributions

In this study, we introduce the first large-scale benchmark for open-set skeleton-based action recognition task, termed the OS-SAR benchmark. This benchmark encompasses a wide range of

Table 4.10: Unseen classes for five random splits on ToyotaSmartHome [43] dataset.

TYT	Unseen classes
Run1	'Drink.Fromcup', 'Cook.Cleandishes', 'Laydown', 'Enter', 'Takepills', 'Walk', 'Usetablet', 'Cook.Usestove', 'Leave', 'Eat.Snack', 'Maketea.Boilwater', 'Cook.Cut', 'Pour.Frombottle', 'Drink.Fromglass', 'Uselaptop', 'WatchTV', 'Pour.Fromkettle', 'Usetelephone'
Run2	'Leave', 'Usetelephone', 'Maketea.Boilwater', 'Cook.Usestove', 'Eat.Snack', 'Cook.Cleanup', 'Pour.Fromkettle', 'Cook.Stir', 'Walk', 'Usetablet', 'Pour.Frombottle', 'Drink.Fromglass', 'Getup', 'Makecoffee.Pourgrains', 'Drink.Fromcup', 'Takepills', 'Makecoffee.Pourwater', 'Cutbread'
Run3	'Usetelephone', 'Makecoffee.Pourwater', 'Cook.Usestove', 'Maketea.Insertteabag', 'Uselaptop', 'Enter', 'Maketea.Boilwater', 'Cutbread', 'Pour.Frombottle', 'Drink.Fromcan', 'Cook.Stir', 'Laydown', 'Cook.Cleanup', 'Drink.Fromcup', 'Readbook', 'Drink.Frombottle', 'Leave', 'Pour.Fromcan'
Run4	'Cutbread', 'Usetelephone', 'Drink.Frombottle', 'Walk', 'Usetablet', 'Cook.Cleanup', 'Drink.Fromcan', 'Drink.Fromglass', 'Drink.Fromcup', 'Pour.Fromcan', 'Makecoffee.Pourgrains', 'Maketea.Boilwater', 'Leave', 'Cook.Stir', 'Makecoffee.Pourwater', 'WatchTV', 'Laydown', 'Eat.Attable'
Run5	'Enter', 'Eat.Attable', 'Pour.Frombottle', 'Eat.Snack', 'Cook.Cleanup', 'Takepills', 'Pour.Fromkettle', 'Sitdown', 'Makecoffee.Pourgrains', 'WatchTV', 'Uselaptop', 'Drink.Frombottle', 'Drink.Fromcan', 'Cook.Cut', 'Readbook', 'Cutbread', 'Maketea.Boilwater', 'Maketea.Insertteabag'

backbones, datasets, and evaluation methodologies. To select reasonable baselines, seven open-set recognition methods, including vanilla SoftMax [74], RPL [25], ARPL [24], DEAR [10], Humpty Dumpty [53], PMAL [127], and MCD-V [155], originally developed for image classification and video-based action recognition, are adapted to skeleton-based human action recognition approaches due to the lack of related researches in OS-SAR direction. Our evaluations reveal that these existing methods generally underperform on the OS-SAR benchmark, attributed to the significant differences between the rich image/video data and the comparatively sparse skeleton data.

According to the observation on the limitations faced by all evaluated methods, we identified a lack of effective separation between in- and out-of-distribution samples based on their open-set probabilities. Addressing this, we developed CrossMax, which employs a cross-modality mean max discrepancy during training and a cross-modality logits refinement during testing. This approach significantly improves the disentanglement capability between in- and out-of-distribution samples in terms of open-set probability. CrossMax achieves state-of-the art performances on the constructed benchmarks for open-set skeleton-based action recognition across various backbones, datasets, and evaluation protocols. Its superior performance are validated on both of the open-set metrics and the close-set metrics.

Despite its advancements, CrossMax is susceptible to misclassifications and may inadvertently propagate biased content, leading to incorrect predictions with potential adverse societal impacts.

4.4.2 Limitations

The CrossMax methodology necessitates the tripling of the model due to its reliance on ensemble modalities, covering joints, bones, and velocities. Despite this, the memory footprint remains manageable, especially when considering the relatively compact size of GCN models tailored for skeleton data.

4.4.3 Future Works

Our empirical analysis reveals that the incorporation of multi-modality data yields enhanced open-set performance within the OS-SAR benchmark. As a consequence, we identify a compelling avenue for future research stemming from our devised OS-SAR benchmark. This prospective direction pertains to the optimal utilization of multi-modality data to further improve the efficacy of OS-SAR.

5 Towards Video Based Domain Adaptation for Human Action Recognition

Unlike skeleton-based action recognition where only sparse body joints are available, video-based human action recognition approaches can make use of video data which has rich appearance and background information and serves as another option to accomplish human action recognition task. However, video-based domain adaptation has its own shortcomings, *i.e.*, high sensitivity to the background change and the appearance change, which make generalizability to different domains very essential to the video-based human action recognition methods. In the following part of this thesis, few-shot video-based domain adaptation for general human action recognition (illustrated in Section 5.1) and cross-modal RGB2Depth unsupervised domain adaptation for human fall detection (illustrated in Section 5.2) will be explored, which are separately important for efficient domain adaptation when we take data-label trade-off into consideration, and for real-wold privacy supporting applications, respectively. Part of the content in this chapter comes from our submission in ACM Multimedia [148] and our publication on IEEE Sensors Journal [200].

5.1 Exploring Few-Shot Domain Adaptation for Video Based Human Action Recognition

5.1.1 Introduction and Motivation

Domain shifts, *i.e.*, distribution discrepancies between source domain data and target domain data, are inevitable in real-world applications. Most existing works in the field of human action recognition are conducted within a single domain [30, 129, 164, 167, 184, 199]. The performance of human action recognition models can be significantly impacted by factors such as changes in

sensor types and placements, variations in room layouts, or transitions from synthetic to real-world settings [45, 156].

The majority of current domain adaptation research focuses on Unsupervised Domain Adaptation (UDA) [28, 37, 63, 87, 158, 198] and Semi-Supervised Domain Adaptation (SSDA) [159, 210, 216] settings for human action recognition. These paradigms facilitate the transfer of recognition capabilities from the source domain to the target domain. The advantage of UDA lies in minimizing the labor-intensive labeling of large-scale data in the target domain. However, both UDA and SSDA paradigms require a substantial number of target domain samples, with SSDA additionally requiring a few labeled target domain samples. Our work focuses on another perspective to achieve domain adaptation, *i.e.*, **Few-Shot Domain Adaptation for Action Recognition (FSDA-AR)**, while considering various domains. FSDA-AR minimizes reliance on extensive target domain examples, instead requiring only a few or even a single annotated sample per class in the target domain. These labeled target domain samples act as knowledge support to bridge the domain gap. Unlike pixel-wise tasks, such as semantic segmentation [59], human action recognition assigns a single label to a video sample. Consequently, annotating a few target domain samples for FSDA-AR is less time-consuming, given the substantial human effort required for acting or video surveillance. Collecting video samples demands significant human effort, making few-shot domain adaptation a more practical approach. Despite this advantage, research on FSDA-AR remains sparse.

Among the published works most relevant to our task, one addresses FSDA-AR in videos but does not provide a publicly available benchmark [64], while another [108] focuses solely on a specific data type (radar-based action recognition), which is not ideal for general action recognition. Xu *et al.* [204] investigate FSDA for sports and daily living scenarios, which have limited domain diversity, and the feature extraction backbone is not as standardized as in the UDA task.

To further explore the applicability of FSDA-AR in diverse and challenging domain adaptation scenarios, we evaluate our approach using five publicly available video-based human action recognition datasets. This evaluation encompasses a variety of conditions, including both small and large domain gaps. Specifically, we investigate the adaptation from Sims4Action [156] to Toyota Smart Home (TSH) [47], different scenarios inside EPIC-KITCHENS [45], from HMDB [96] to UCF [177], and vice versa.

The selected datasets represent a range of scenarios, such as transitioning from movie-based to real-world third-person datasets, adaptation among egocentric kitchen action recognition datasets, and synthetic-to-real domain adaptation with significant domain differences. The benchmark includes various baseline approaches, such as UDA methods, few-shot action recognition techniques, and statistical methods, all reformulated to fit the FSDA-AR framework.

Our observations reveal that existing baselines often fail to deliver consistent performance across

the diverse domains involved in FSDA-AR tasks, particularly with challenging domain adaptation datasets. This highlights the necessity for a novel method specifically designed for few-shot domain adaptation, capable of adapting to a wide range of domains. Consequently, we introduce our innovative approach for FSDA-AR in this study.

Our approach is structured around three core objectives: (1) Enhancing temporal data generalization: We aim to improve the model's ability to generalize when handling temporal data by developing mechanisms to extract transferable temporal patterns and features across different domains. (2) Leveraging statistical distributions: We utilize the statistical properties of source-domain samples and a small number of labeled target-domain samples to enrich features within the latent space, thereby enhancing robust and discriminative feature learning. (3) Establishing a unified embedding space: We aim to create a shared embedding space that integrates both source and target domains, enabling cohesive operation across various domains and promoting cross-domain knowledge transfer.

To achieve these objectives, we propose RelaMiX. RelaMiX is constructed by three components, which are TRAN-RD, SDFM, and CDIA loss. The Temporal Relational Attention Network with Relation Dropout (TRAN-RD) is designed to enhance the generalizability of temporal features. TRAN-RD captures nuanced neighborhood information by considering diverse snippet levels, relational attention, and relation combinations, with relation dropout enhancing the representativeness of each combination. Statistical Distribution-based Feature Mixture (SDFM) Mechanism increases feature diversity within the aligned latent space. By computing the covariance and empirical mean for each temporal snippet, we construct Gaussian distributions for latent space features from the source domain. We then generate mixed-domain features through empirical mean transformations and interpolation techniques. During training, we fine-tune the temporal aggregation network with features from both source and target domains, as well as mixed features, fostering cross-domain knowledge transfer. Cross-Domain Information Alignment (CDIA) aligns the source domain with target domain centers and distances them from negative anchors, applying a similar strategy to mixed features with temporally augmented positives and random mixed negatives. This alignment bridges the domain gap and enhances feature transfer using few-shot samples.

These innovations collectively utilize diverse temporal data, enhance feature diversity, and ensure domain alignment, leading to RelaMiX achieving top performance on the benchmark. Our contributions are summarized as follows:

- We tackle the task of Few-Shot Domain Adaptation for Action Recognition (FSDA-AR) by establishing a new benchmark that addresses diverse and challenging domains. These domains include transitions from movie data to real-world third-person data, cross-person egocentric perspectives, and synthetic data to real-world data.

- We introduce the novel RelaMiX approach, which consists of three pivotal components: the Temporal Relational Attention Network with Relational Dropout (TRAN-RD) for enhanced temporal generalization, the Statistic Distribution-based Feature Mixing (SDFM) mechanism to enrich the shared latent space, and the Cross-Domain Information Alignment (CDIA) to effectively bridge significant domain gaps.
- Our method sets a new standard by achieving state-of-the-art results on the FSDA-AR benchmark across the 1-, 5-, 10-, and 20-shot settings. Notably, compared to UDA solutions, RelaMiX for FSDA-AR demonstrates comparable performance.

5.1.2 Method

5.1.2.1 Problem Formulation

The FSDA-AR task operates under the assumption of having a fully labeled large-scale source domain training set $D_{source} = (\mathbf{v}_i^{source}, l_i^{source})_{i=1}^{N_{source}^{train}}$ and a small target domain training set $D_{target} = (\mathbf{v}_i^{target}, l_i^{target})_{i=1}^{N_{target}^{train}}$, which contains only a few labeled samples per class. Our aim is to train a model utilizing both D_{source} and D_{target} to ensure that the model performs well on the target domain test set, denoted as $T_{target} = (\mathbf{v}_i^{target}, l_i^{target})_{i=1}^{N_{target}^{test}}$.

Here, N_{source}^{train}, N_{target}^{train}, and N_{target}^{test} refer to the number of samples in the source domain training set, target domain training set, and target domain test set, respectively. The index i indicates the sample number, while $\mathbf{v}$ and l represent the video sample and their corresponding label.

5.1.2.2 Baselines on FSDA-AR Benchmark

For our FSDA-AR benchmark, we construct a range of diverse baselines by incorporating three FSDA-AR specific approaches: FS-ADA [108], SSA^{2}Lign [204], and PASTN [64]. Additionally, we adapt established UDA approaches for FSDA-AR, including TA^{3}N [28], TranSVAE [198], CoMix [158], and CO^{2}A [41]. We also include few-shot human action recognition methods reformulated for FSDA-AR, such as TRX [151] and HyRSM [194]. Furthermore, we evaluate statistical baselines including random chance (Random), K-Nearest Neighbors (KNN), Nearest Neighbor (NN), and Nearest Center (NC).

For consistency, all baselines utilize the I3D backbone [19] pre-trained on Kinetics400 [89], as I3D is a standard choice for video feature extraction in UDA tasks [198].

Statistical baselines. Video features are extracted with I3D, omitting the final classification layer, and these features are used by all statistical baselines except for the Random method. The Random baseline randomly assigns a class to test samples, setting a performance lower bound. In the KNN

method, labels for test samples are determined by averaging outcomes from the 3-, 5-, and 10-NN methods, using neighbors from the source domain. The Nearest Center method assigns labels based on the nearest class center in the source domain, while the Nearest Neighbor method uses the closest source domain sample for labeling. These methods maintain consistent performance across different shot settings by using the same source domain. They serve as statistical baselines to highlight Domain Generalization (DG) performance, emphasizing the advantage of incorporating few-shot samples from the target domain.

Unsupervised domain adaptation baselines. To enrich our FSDA-AR benchmark and facilitate comparisons with existing domain adaptation frameworks, we implement and adapt several established video-based UDA methods for the few-shot domain adaptation task. This adaptation involves incorporating supervised classification loss on the target domain training set and converting the unsupervised contrastive loss into triplet margin loss, using only a few-shot samples from the target domain.

Specifically, we leverage four prominent methods: TA^{3}N [28], TranSVAE [198], CO^{2}A [41], and CoMix [158]. TA^{3}N and CO^{2}A utilize domain adversarial learning, TranSVAE focuses on domain disentanglement, and CoMix aims to bridge domain gaps by mixing background information from both domains.

Few-shot learning baselines. To evaluate the performance of few-shot learning approaches in the few-shot domain adaptation task, we leverage two representative works in video-based few-shot learning: TRX [151] and HyRSM [194]. These methods are employed to assess how effectively few-shot learning techniques can be adapted to the FSDA-AR context.

Few-shot domain adaptation baselines. We utilize three existing approaches for FSDA-AR. The first approach is FS-ADA [108], which addresses radar-based FSDA-AR using adversarial domain adaptation. The second approach, PASTN [64], introduces a pairwise attentive, adversarial spatio-temporal network. The third approach, SSA^{2}Lign [204], employs attentive alignment of snippets to bridge the domain gap.

To explore FSDA-AR against more diverse domain shifts and achieve the unification of the feature backbone for a fair comparison with UDA tasks, we establish a novel benchmark using publicly accessible datasets that encompass diverse domain styles. We replicate the performance of these methods by unifying the video extraction backbone as I3D [19], ensuring a fair comparison with UDA methods.

5.1.2.3 Introduction of RelaMiX Method

In this section, we introduce the key concepts of our proposed RelaMiX approach, illustrated in Fig. 5.1.

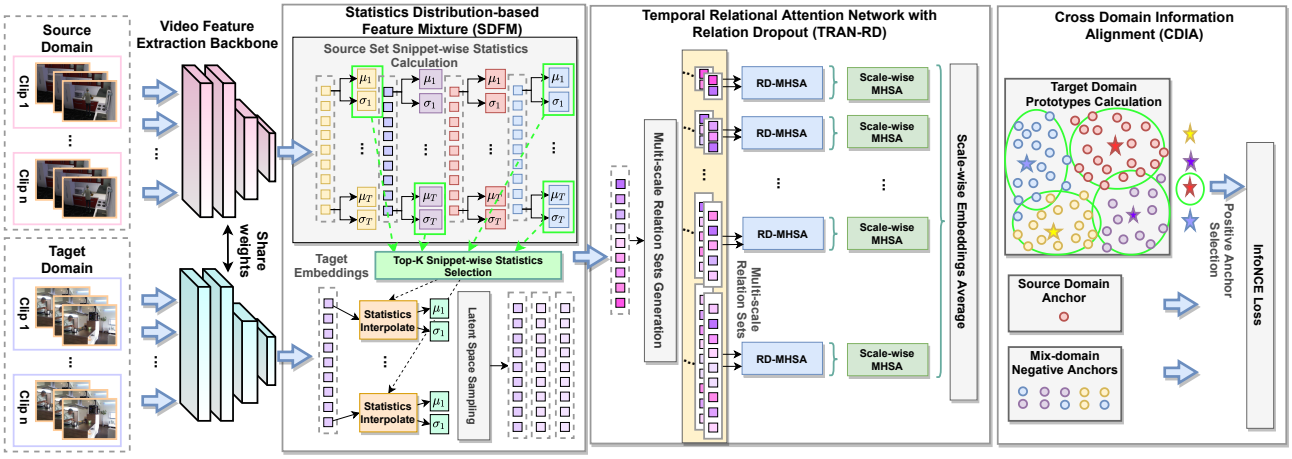


Figure 5.1: An overview of the RelaMiX architecture. The input video is first separated into overlapped snippets extracted through a fixed-size temporal sliding window and then fed into the video feature extraction backbone to extract snippet-level features. Then, we calculate the statistical empirical mean and covariance of each class for each snippet considering all the samples from the source domain training set. The mean and covariance of the Top-K nearest centers corresponding to a snippet from the given query are chosen to generate the synthesized cluster center of the samples of target-domain latent space. We use the generated mean and covariance to formulate Gaussian distributions for each temporal snippet and sample more latent space features according to the few available shots from the target domain. Next, temporal relation sets are built, we make use of Relation-Dropout Multi-Head Self-Attention (RD-MSHA) to learn representative features within each relation set while using scale-wise Multi-Head Self-Attention (Scale-wise MSHA) to aggregate features across different relation scales. Finally, alongside cross-entropy losses, Cross-Domain Information Alignment (CDIA) loss is leveraged to bridge the domain gap by using target space batch-wise prototypes.

RelaMiX begins by utilizing I3D [19] to extract snippet-wise features. It then integrates a statistic distribution-based feature mixture technique to enhance the information in the latent space shared across domains. Additionally, it employs a temporal relation attention network with relation dropout to achieve more generalizable temporal information aggregation. Lastly, a cross-domain information alignment loss is applied to facilitate representative feature learning and bridge the domain gap.

Snippet-wise video feature extraction. Similar to our baselines, we adopt I3D [19] as our backbone to obtain video representations and use a temporal sliding window to extract snippet features for a given video. Consider a video sample represented as $\mathbf{v} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{N_T}\}$, where $\mathbf{v}_i$ denotes the i-th frame of the video, N_w is the window size of the sliding temporal window, and N_T is the number of frames. We can derive a set of video snippets, denoted as Eq. 5.1,

$$\{\mathbf{s}_i\} = \{\{\mathbf{v}_i, \ldots, \mathbf{v}_{i+N_w}\} \mid i \in [1, N_T]\}, \tag{5.1}$$

where zero padding is used at the start and end of the video.

Next, we extract snippet-wise features. By inputting these snippets into the I3D-based feature

extractor $\mathbf{H}_\alpha$, we obtain the set of clip features for a sample, denoted as Eq. 5.2,

$$\mathbf{f} = [\mathbf{H}_\alpha(\mathbf{s}_1), \ldots, \mathbf{H}_\alpha(\mathbf{s}_{N_T})]. \tag{5.2}$$

Statistic distribution-based feature mixture (SDFM). To better leverage the information provided by the few-shot samples from the target domain, we propose a new method, SDFM, to synthesize more target-domain embeddings using the statistics calculated from the snippet features of the source domain training set. The details of the SDFM are as follows. We first calculate the statistical centers and covariance matrix as shown in Eq. 5.3 and Eq. 5.4.

$$\mu_c^{(t,source)} = \frac{\sum_{i=1}^{N_c} \mathbf{f}_{(i,c)}^{(t,source)}}{N_c}, \tag{5.3}$$

$$\sigma_c^{(t,source)} = \sqrt{\frac{\sum_{i=1}^{N_c} (\mathbf{f}_{(i,c)}^{(t,source)} - \mu_c^{(t,source)})^2}{N_c - 1}}, \tag{5.4}$$

where $\mu_c^{(t,\text{source})}$ and $\sigma_c^{(t,\text{source})}$ represent the mean and covariance of the embeddings of the c-th category and t-th snippet from the source domain. Here, N_c denotes the number of samples in category c, and $\mathbf{f}_{(i,c)}^{(t,\text{source})}$ indicates the embeddings from the i-th sample within the c-th category, considering the t-th snippet.

For a sample from the few-shot target domain training set with embeddings $\hat{\mathbf{f}}^{(t,\text{target})}$ for the t-th snippet, we first calculate the Top-K nearest cluster centers for each snippet. This calculation is based on the distance to the mean of the snippet embeddings from the source domain, as shown in Eq. 5.5.

$$I_c^t = TopK_{c\in\Omega_c}(e^{(1-D(\mu_c^{(t,source)}, \hat{\mathbf{f}}^{(t,target)}))}), \tag{5.5}$$

where I_c^t indicates the categories which are selected. $D(\cdot)$ indicates the Euclidean distance. Ω_c indicates the set of categories. Then we calculate the synthesized empirical mean for this target domain embeddings according to Eq. 5.6,

$$\hat{\mu}^t, \hat{\sigma}^t = \frac{\hat{\mathbf{f}}^{(t,target)} + \sum_{k\in I_c^t} \mu_k^{(t,source)}}{K+1}, \frac{\sum_{k\in I_c^t} \sigma_k^{(t,source)}}{K} + \alpha, \tag{5.6}$$

where α is a fixed factor and K is the number of selected centers.

Next, we construct a multivariate normal distribution based on the synthesized empirical mean and covariance, as shown in Eq. 5.7.

$$\hat{\mathbf{f}}_{new}^{(t,target)} \sim \frac{1}{\hat{\sigma}^t\sqrt{2\pi}} e^{-\frac{(\mathbf{x}-\hat{\mu}^t)^2}{2(\hat{\sigma}^t)^2}}. \tag{5.7}$$

Additional embeddings can be derived by leveraging the established normal distributions for each temporal snippet within the target domain. The newly generated features $\hat{\mathbf{f}}_{new}^{(t,\text{target})}$ share the same class as $\hat{\mathbf{f}}^{(t,\text{target})}$.

This process utilizes statistical parameters obtained from the source domain in conjunction with the provided few-shot samples from the target domain. The aim of this approach is to enhance diversity within the latent space while simultaneously integrating information from both the source and target domains.

Temporal relational attention network with relation dropout (TRAN-RD). When adapting the model to another domain using only a few samples, a highly generalizable temporal aggregation approach is essential for capturing important cues from different temporal relations. To address this, we propose a new temporal aggregation mechanism, namely the TRAN-RD. This mechanism incorporates two major concepts: Relation-Dropout based Multi-Head Self-Attention (RD-MHSA) and Scale-wise Multi-Head Self-Attention (Scale-wise MHSA), which focus on different temporal relational granularities and learn representative and generalizable features.

RD-MHSA is applied to all the embeddings from the source domain, the target domain, and the generated embedding set of the target domain. We first generate multi-scale relational index sets as shown in Eq. 5.8,

$$\begin{aligned}\Omega_r = &\{(i,...,k,...,j)|i \leq ... \leq k \leq ... \leq j, \\ &(i,...,k,...,j) \in [1,N_T]^r, \ and \ r \in [2,N_T]\},\end{aligned} \tag{5.8}$$

where we can capture relational indexes according to different scales. The selected snippets preserve the temporal order. The variable r is used to define the selected scales of the desired relational set.

RD-MHSA is first used to aggregate the temporal features within each snippet. The relational attended snippet embedding $\hat{\mathbf{f}}_s$ for a snippet $\mathbf{f}_s$ from the given snippet relation set $\Omega_s \subseteq \Omega_r$ can be calculated using Eq. 5.9.

$$\mathbf{f}_a = LN\left[SoftMax\left[\frac{\mathbf{P}_Q(\mathbf{f}_s)\cdot\mathbf{P}_K(\mathbf{f}_s)}{\sqrt{d_k}}\right]*\mathbf{P}_V(\mathbf{f}_s)\right], \tag{5.9}$$

$$\hat{\mathbf{f}}_s = LN\left[\mathbf{f}_s + \sum_{h=1}^{N_h}\mathbf{f}_a^h + FFN(\mathbf{f}_s)\right], \tag{5.10}$$

where LN denotes layer normalization, and FFN denotes a multi-layer-perceptron (MLP) based Feed-Forward Network. The term d_k is a scale factor. $\mathbf{P}_Q$, $\mathbf{P}_K$, and $\mathbf{P}_V$ are linear projections. $\mathbf{f}_a^h$ indicates the relational attention obtained through N_h heads.

Following this process, we obtain the self-attended relation set $\hat{\Omega}_s$. We then apply dropout to

the snippets within each attended relation set as shown in Eq. 5.11.

$$\hat{\Omega}_s^{DP} = DropOut(\hat{\Omega}_s, \beta), \tag{5.11}$$

where β is the pre-defined dropout ratio for the snippets within one relation set. The RD-MHSA aims to learn representative features even when some snippets are randomly unavailable during the training procedure.

Scale-wise MHSA is then used to aggregate the information within each scale. First, we concatenate all snippets inside one relation set after the relation dropout along the temporal dimension, denoted as $\hat{\mathbf{f}}_s^{\Omega_s}$. The temporal dimension is then treated as the token dimension for MHSA. Scale-wise MHSA is then performed as shown in Eq. 5.12 and Eq. 5.13,

$$\hat{\mathbf{f}}_a^{\Omega_s} = LN\left[SoftMax\left[\frac{\hat{\mathbf{P}}_Q(\hat{\mathbf{f}}_s^{\Omega_s}) \cdot \hat{\mathbf{P}}_K(\hat{\mathbf{f}}_s^{\Omega_s})}{\sqrt{\hat{d}_k}}\right] * \hat{\mathbf{P}}_V(\hat{\mathbf{f}}_s^{\Omega_s})\right], \tag{5.12}$$

$$\widetilde{\mathbf{f}}_a^{\Omega_s} = LN\left[\hat{\mathbf{f}}_s^{\Omega_s} + \sum_{h=1}^{N_h} \hat{\mathbf{f}}_a^{(\Omega_s,h)} + FFN(\hat{\mathbf{f}}_s^{\Omega_s})\right], \tag{5.13}$$

where all the projections $\hat{\mathbf{P}}_Q$, $\hat{\mathbf{P}}_K$, and $\hat{\mathbf{P}}_V$ are linear projections. The term $\hat{d}_k$ is a fixed scale factor, and N_h denotes the number of heads. Finally, the aggregated embedding is calculated using Eq. 5.14.

$$\mathbf{f}^* = \frac{\sum_{\Omega_s \subseteq \Omega_r} \widetilde{\mathbf{f}}_a^{\Omega_s}}{N_s}, \tag{5.14}$$

where $N_s = N_T - 1$ denotes the total number of the relation sers.

Cross Domain Information Alignment (CDIA). When a source domain anchor is given as $\mathbf{f}^{source}$ after the TRAN-RD, we wish it could be closer to its corresponding cluster centers $\mathbf{f}_c^{target}$ calculated on the target domain while being far away from the negative anchors $\widetilde{\mathbf{f}}^{source}$ from different categories. The CDIA loss can be therefore calculated via Eq. 5.15,

$$\mathscr{L}_{CDIA} = -\frac{1}{N}\sum_{i=1}^{N} \log\left[\frac{e^{(cos(\mathbf{f}_i^{source}, \mathbf{f}_{(i,c)}^{target}))}}{\sum_{j=1}^{N_n} e^{(cos(\mathbf{f}_i^{source}, \widetilde{\mathbf{f}}_j^{source}))}}\right], \tag{5.15}$$

where N denotes the number of samples from the source domain and N_n denotes the number of negative anchors for the i-th anchor from the source domain. The term cos indicates the cosine similarity. $\mathbf{f}_{(i,c)}^{\text{target}}$ represents the nearest target domain center (of the c-th category) for the i-th anchor.

The target domain centers can be calculated using Eq. 5.16.

$$\mathbf{f}_c^{target} = \frac{\sum_{i=1}^{N_c} \mathbf{f}_{(i,c)}^{target}}{N_c}, \tag{5.16}$$

where N_c indicates the number of samples for class c, and $\mathbf{f}_c^{\text{target}}$ represents the target domain center for class c. In addition to the CDIA loss, we use supervised cross-entropy losses for samples from the source domain training set, the target domain few-shot training set, and the target domain generated training set, denoted as $\mathscr{L}_{CES}$, $\mathscr{L}_{CET}$, and $\mathscr{L}_{CEA}$, respectively. To obtain representative features from the generated target domain training set, we use an additional contrastive learning loss, as shown in Eq. 5.17.

$$\mathscr{L}_{aux} = -\frac{1}{N_g} \sum_{i=1}^{N_g} \log \left[\frac{e^{(cos(\mathbf{f}_i^{target}, \mathbf{f}_j^{target}))}}{\sum_{k=1}^{N_n} e^{(cos(\mathbf{f}_i^{target}, \widetilde{\mathbf{f}}_k^{target}))}} \right], \tag{5.17}$$

where the positive anchors $\widetilde{\mathbf{f}}_k^{\text{target}}$ are generated through random permutation of the input snippet along the temporal axis and $\mathbf{f}_j^{target}$ denotes another randomly selected sample from the same category with $\mathbf{f}_i^{target}$. N_g indicates the number of generated features in the shared latent space, and *cos* denotes cosine similarity. The overall supervision is achieved by a weighted sum of the aforementioned loss functions, as shown in Eq. 5.18.

$$\begin{aligned} \mathscr{L}_{all} = \omega_1 * \mathscr{L}_{CDIA} + \omega_2 * \mathscr{L}_{CES} + \omega_3 * \mathscr{L}_{CET} + \\ \omega_4 * \mathscr{L}_{CEA} + \omega_5 * \mathscr{L}_{aux}. \end{aligned} \tag{5.18}$$

5.1.3 Experiments

5.1.3.1 Datasets

We utilize five popular human action recognition datasets, namely HMDB-51 [96], UCF-101 [177], EPIC-KITCHENS-55 [45], Toyota Smart Home (TSH) [47], and Sims4Action [156], to investigate FSDA-AR. These datasets encompass a wide range of action types and recording environments, enabling a comprehensive evaluation of domain adaptation techniques. Methods categorized under *DG* use only the source domain samples, whereas methods under *FSDA-AR* utilize samples from both the source and target domains. **HMDB-51** [96] contains 6,766 video clips from various sources, covering 51 action categories with a minimum of 101 clips per action. For the DA task, 12 action classes are selected.

UCF-101 [177] consists of 13,320 video clips across 101 action categories. These datasets are employed in the HMDB $\rightarrow$ UCF and UCF $\rightarrow$ HMDB adaptation tasks, where they share 12 classes.

Table 5.1: Experimental results on UCF [177] → HMDB [96], HMDB [96] → UCF [177], and EPIC-KITCHEN [45]. S-X indicates Shot-X.

	Method	❶ UCF → HMDB				❷ HMDB → UCF				❸ EPIC-KITCHEN mean				❹ Sims4Action → TSH			
		S-1	S-5	S-10	S-20	S-1	S-5	S-10	S-20	S-1	S-5	S-10	S-20	S-1	S-5	S-10	S-20
DG	Random	——— 8.6 ———				——— 8.1 ———				——— 12.5 ———				——— 11.1 ———			
	KNN	——— 81.1 ———				——— 88.3 ———				——— 28.1 ———				——— 3.3 ———			
	Nearest Center	——— 83.9 ———				——— 91.5 ———				——— 27.4 ———				——— 28.0 ———			
	Nearest Neighbor	——— 80.1 ———				——— 88.6 ———				——— 25.5 ———				——— 3.27 ———			
FSDA-AR	CoMix [158]	83.1	88.1	89.7	90.8	91.0	93.2	96.8	96.3	31.2	31.8	32.1	32.7	24.2	20.6	28.9	35.5
	CO^2A [41]	83.9	88.1	89.1	91.1	92.5	94.0	96.7	97.5	32.6	36.4	38.0	38.2	21.9	26.6	34.0	42.1
	TA^3N [28]	83.3	88.9	88.3	91.7	93.7	95.1	97.5	98.0	37.9	41.2	42.1	43.0	21.1	29.8	35.8	42.7
	TranSVAE [198]	82.3	82.8	83.2	84.8	89.7	89.0	94.4	95.1	37.6	41.1	40.8	43.3	22.6	22.7	18.9	22.7
	TRX [151]	77.2	80.3	78.6	81.9	82.2	83.1	81.1	84.4	26.7	27.4	28.7	30.2	14.0	13.8	19.0	18.9
	HyRSM [194]	79.7	81.1	82.2	83.6	88.1	90.1	91.0	90.8	35.8	36.7	37.1	37.8	18.9	22.4	27.4	28.0
	FS-ADA [108]	82.7	87.2	88.6	87.2	91.9	94.4	93.7	96.5	37.0	39.7	39.3	40.4	17.1	22.6	28.3	28.0
	PASTN [64]	83.4	86.2	88.3	89.8	91.2	94.2	95.8	96.5	36.1	40.5	40.3	42.5	22.6	22.6	22.6	28.0
	SSA^2lign [204]	80.6	85.0	88.3	87.8	87.0	94.4	94.6	94.4	31.5	40.1	40.9	42.0	22.6	23.7	35.0	41.3
	RelaMiX (ours)	**85.6**	**91.1**	**91.1**	**92.2**	**94.1**	**97.2**	**97.9**	**98.4**	**40.7**	**43.9**	**44.4**	**45.2**	**27.0**	**31.0**	**38.9**	**49.2**

Figure 5.2: An overview of domain differences.

EPIC-KITCHENS-55[45] comprises 55 hours of egocentric videos, capturing kitchen activities from 32 participants. We leverage the domain adaptation benchmark defined by [141] on 8 overlapping activities.

Toyota Smart Home (TSH) [47] includes 16,115 video clips of 31 daily living activities, with 10 selected for our domain adaptation benchmark.

Sims4Action [156] is a synthetic dataset designed for cross-domain evaluation on TSH. It consists of 13,232 video clips depicting 10 daily living activities performed by avatars in the Sims4game, used in the Sims4Action → TSH adaptation task. These datasets provide a robust foundation for exploring DA techniques in various cross-domain scenarios, presenting challenges inherent in both real-world and synthetic video data.

Table 5.2: Experimental results on the EPIC-KITCHEN dataset considering six different adaptation settings.

	Method	❶ D1 → D2				❷ D2 → D1				❸ D1 → D3				❹ D3 → D1				❺ D2 → D3				❻ D3 → D2			
		S-1	S-5	S-10	S-20	S-1	S-5	S-10	S-20	S-1	S-5	S-10	S-20	S-1	S-5	S-10	S-20	S-1	S-5	S-10	S-20	S-1	S-5	S-10	S-20
DG	Random	———	12.5	———		———	12.3	———		———	12.7	———		———	12.6	———		———	12.3	———		———	12.4	———	
DG	KNN	———	25.5	———		———	26.9	———		———	26.2	———		———	27.4	———		———	28.5	———		———	33.9	———	
DG	Nearest Center	———	24.0	———		———	29.2	———		———	29.2	———		———	25.1	———		———	34.0	———		———	23.1	———	
DG	Nearest Neighbor	———	22.1	———		———	24.9	———		———	26.6	———		———	26.4	———		———	24.1	———		———	28.7	———	
FSDA-AR	CoMix [158]	31.5	32.0	34.3	35.7	25.2	27.9	31.1	29.9	30.0	30.4	28.5	30.8	30.4	30.2	28.8	30.4	34.0	34.5	35.3	34.4	36.0	35.7	34.5	35.1
FSDA-AR	CO^2A [41]	31.7	33.5	33.6	33.3	32.6	38.4	36.1	39.8	30.2	36.4	38.3	38.0	34.0	36.6	40.5	40.0	34.7	36.6	40.5	40.0	32.6	36.8	39.0	38.6
FSDA-AR	TA^3N [28]	36.8	39.0	40.2	43.5	36.8	38.9	40.4	40.5	36.7	40.2	41.0	40.3	33.1	40.0	40.9	41.8	41.1	43.4	43.5	45.6	42.8	45.8	46.5	46.5
FSDA-AR	TranSVAE [198]	32.9	39.5	39.5	42.8	35.3	40.4	37.5	41.7	37.0	39.1	40.3	42.3	36.1	38.2	37.5	41.4	42.8	44.9	44.5	45.9	41.2	44.4	45.6	45.6
FSDA-AR	TRX [151]	24.8	25.0	25.2	25.9	26.1	27.7	30.7	31.6	25.3	25.9	28.1	28.8	26.6	28.9	29.3	30.0	28.4	28.0	30.6	31.9	28.8	29.1	28.4	33.1
FSDA-AR	HyRSM [194]	31.1	33.5	34.0	37.2	33.4	32.7	33.9	34.8	33.2	37.2	36.5	36.7	35.0	34.8	35.7	35.0	40.4	40.3	41.2	41.4	41.6	41.8	41.5	41.5
FSDA-AR	FS-ADA [108]	36.4	38.1	38.4	37.7	34.7	36.8	39.1	39.3	36.1	37.4	38.2	40.5	32.2	39.8	35.9	38.6	42.4	42.5	42.0	44.3	40.4	43.5	42.1	42.2
FSDA-AR	PASTN [64]	33.3	38.2	37.7	41.3	34.0	38.9	36.8	40.9	35.3	39.4	39.0	41.0	33.6	37.9	38.2	41.1	39.2	43.1	44.4	44.6	43.0	45.5	45.9	45.8
FSDA-AR	SSA^2lign [204]	32.0	40.4	37.6	41.5	31.3	40.1	40.5	41.6	30.1	39.3	42.0	42.6	34.5	38.9	41.1	39.1	28.7	42.9	42.1	44.5	32.3	38.7	41.9	42.7
	RelaMiX (ours)	**39.1**	**43.9**	**43.7**	**47.9**	**38.4**	**41.6**	**42.1**	**42.8**	**38.4**	**42.1**	**42.5**	**43.1**	**37.9**	**41.6**	**42.3**	**42.5**	**45.1**	**46.2**	**47.4**	**46.5**	**45.5**	**48.0**	**48.1**	**48.1**

5.1.3.2 Implementation Details

We randomly select N_{shot} samples per class from the target domain training set to construct our benchmarks, with $N_{shot} \in \{1, 5, 10, 20\}$. The few-shot samples are fixed to enhance knowledge guidance through co-training using information from both the source and target domains. To ensure fair comparison, all feature extraction backbones are unified as I3D [19], initialized with pre-trained weights from Kinetics400 [89].

Our model is trained on an NVIDIA A100 GPU using PyTorch 1.12, with a batch size of 32. The learning rate decays in steps at epoch 60 and 80, and we use the Adam optimizer [92] with an initial learning rate of 0.0001 for 100 epochs. The sliding window size for feature extraction is set to $N_w = 16$, with temporal zero padding of 8. For SDFM, $K = 2$ is chosen. The weights of the losses are set as $\omega_1 = 0.0001$, $\omega_2 = 1$, $\omega_3 = 1$, $\omega_4 = 0.01$, and $\omega_5 = 0.0001$. The β parameter in TRAN-RD is set to 0.5. Both RD-MHSA and Scale-wise MHSA use 8 heads. In SDFM, α is set to 0.21, and 200 samples are generated for each action category. The computational complexity of our model is 108.92 GFLOPS.

5.1.3.3 Analysis of the Benchmark

The performance results for the transfer tasks involving UCF [177] to HMDB [96], HMDB [96] to UCF [177], EPIC-KITCHEN [45], and Sims4Action [156] to Toyota Smart Home [47] (TSH) settings are presented in Table 5.1 (❶, ❷, ❸, and ❹), respectively. We also provide per-split performances on the EPIC-KITCHEN dataset in Table 5.2, where S-1, S-5, S-10, and S-20 indicate shot 1, shot 5, shot 10, and shot 20, respectively. Regarding the experiments on EPIC-KITCHEN [45], the Top-1 accuracy under different shot settings for various methods are as follows: CoMix [158], CO^2A [41], TA^3N [28], TranSVAE [198], FS-ADA [108], and PASTN [64] achieve 31.2%, 32.6%,

Table 5.3: Task comparison between FSDA-AR and UDA.

❶ **UDA approaches on UDA task on EPIC-KITCHEN (2645 shots).**				
TranSVAE [198]	——————— 52.6 ———————			
CoMix [158]	——————— 43.2 ———————			
TA^{3}N [28]	——————— 39.9 ———————			
DANN [63]	——————— 39.2 ———————			
ADDA [189]	——————— 39.2 ———————			
❷ **TA^{3}N and RelaMiX on FSDA-AR task on EPIC-KITCHEN.**				
Method	**Shot-1**	**Shot-5**	**Shot-10**	**Shot-20**
TA^{3}N [28]	38.9	41.8	42.1	43.0
RelaMiX (ours)	41.0	44.4	44.5	45.1
❸ **UDA approaches for UDA tasks on UCF → HMDB (840 shots).**				
TranSVAE [198]	——————— 87.8 ———————			
CoMix [158]	——————— 86.7 ———————			
TA^{3}N [28]	——————— 81.4 ———————			
DANN [63]	——————— 80.1 ———————			
ADDA [189]	——————— 79.2 ———————			
❹ **TA^{3}N and RelaMiX for FSDA-AR task on UCF → HMDB.**				
Method	**Shot-1**	**Shot-5**	**Shot-10**	**Shot-20**
TA^{3}N [28]	83.3	88.9	88.3	91.7
RelaMiX (ours)	84.4	89.7	90.3	92.8
❺ **UDA approaches for UDA tasks on HMDB → UCF (1438 shots).**				
TranSVAE [198]	——————— 99.0 ———————			
CoMix [158]	——————— 93.9 ———————			
TA^{3}N [28]	——————— 90.5 ———————			
DANN [63]	——————— 88.1 ———————			
ADDA [189]	——————— 88.4 ———————			
❻ **TA^{3}N and RelaMiX for FSDA-AR task on HMDB → UCF.**				
Method	**Shot-1**	**Shot-5**	**Shot-10**	**Shot-20**
TA^{3}N [28]	93.7	95.1	97.5	98.0
RelaMiX (ours)	95.6	96.5	97.7	98.2
❼ **UDA approaches for UDA task on Sims4Action → TSH (8552 shots).**				
Schneider *et al.* [160]	——————— 36.3 ———————			
TA^{3}N [28] ([160])	——————— 8.0 ———————			
❽ **TA^{3}N and RelaMiX for FSDA-AR task on Sims4Action → TSH.**				
Method	**Shot-1**	**Shot-5**	**Shot-10**	**Shot-20**
TA^{3}N [28]	21.1	29.8	35.8	42.7
RelaMiX (ours)	24.6	31.4	36.7	45.7

38.9%, 37.6%, 37.0%, 36.1% under 1-shot, 31.8%, 36.4%, 41.8%, 42.1%, 39.7%, 40.5% under 5-shot, 32.1%, 38.0%, 42.1%, 40.8%, 39.3%, 40.3% under 10-shot, and 32.7%, 38.2%, 43.0%, 43.3%, 40.4%, 42.5% under 20-shot, respectively.

Both the UDA methods implemented within the FSDA-AR context and the previously published FSDA-AR techniques show notable performance, consistently outperforming random baseline. This observation underscores the effective reduction of domain gaps when utilizing a limited number of labeled shots. Notably, the required number of target domain samples is substantially reduced in FSDA-AR compared to traditional UDA tasks. For instance, with only 5 labeled shots, the total required sample count is approximately 1.6% of that used in the conventional UDA setting on EPIC-KITCHEN D1→D2.

However, it is noteworthy that the TranSVAE method [198] exhibits inferior performance in the FSDA-AR task, even though it outperforms TA^{3}N [28] in the UDA context, as reported in Table 5.3. This discrepancy suggests that the disentanglement method used in domain adaptation heavily relies on the availability of large-scale data from the target domain to effectively capture adaptation cues. Similar trends can be observed with the CoMix approach [158], which depends on the diversity of backgrounds in the target domain for adaptation, particularly in scenarios involving datasets with substantial domain gaps.

In all our experiments, we employ the I3D backbone [19] to ensure equitable comparisons. To facilitate a comparison with the SSA^{2}Lign approach [204], we replace the TimesFormer [12] backbone in SSA^{2}Lign with the I3D backbone. The reduction in performance observed in the adapted SSA^{2}Lign model on the leveraged datasets can be attributed to two primary factors. Firstly, there are disparities in domain gaps in our experimental configurations. Secondly, SSA^{2}Lign relies on transformer features from the TimesFormer [12] backbone.

However, to ensure a fair comparison with methods originally tailored for UDA, it is imperative to standardize the backbone to I3D [19]. This standardization is necessary because I3D is commonly employed in UDA works [198], allowing us to effectively demonstrate that the observed performance enhancements are a consequence of our proposed method, rather than a consequence of backbone substitution.

Comparing against the best baseline for each shot-setting, our proposed RelaMiX achieves 2.8%, 2.7%, 2.3%, and 1.9% performance improvements for the 1∼20-shot settings on EPIC-KITCHEN [45] in terms of FSDA-AR, as shown in Table 5.1 ❸. The per-split performances are showcased in Table 5.2, where RelaMiX consistently demonstrates superior results on the FSDA-AR task for each split of EPIC-KITCHEN [45]. Some UDA approaches adapted for FSDA-AR do not outperform the statistical baselines, indicating that fewer target domain samples can cause overfitting in approaches that require a large number of samples, especially on datasets with substantial domain differences,

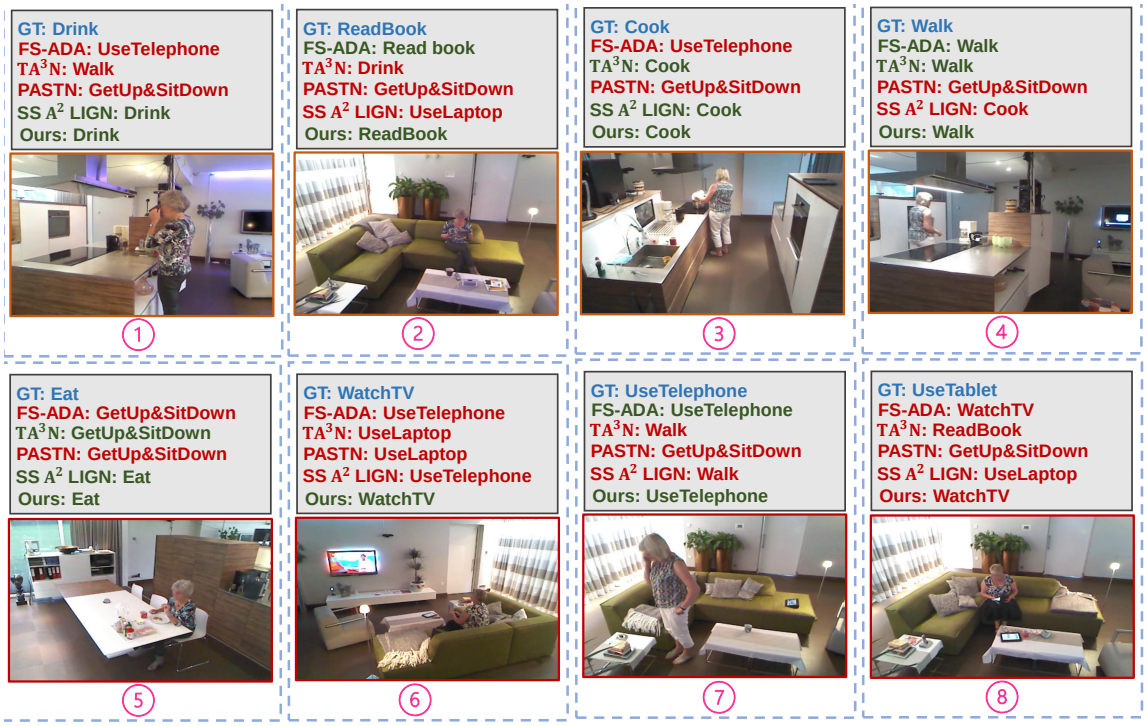


Figure 5.3: Qualitative results for FSDA-AR on Shot-20 Sims4Action [156] → TSH [47].

such as CoMix [158] on EPIC-KITCHEN [45].

Regarding the FSDA-AR task on UCF [177] → HMDB [96] and HMDB [96] → UCF [177], as introduced in Table 5.1 ❶ and ❷, all approaches show promising performance even under the 1-shot setting. When $N_{shot} \geq 5$ on UCF [177] → HMDB [96], CoMix [158], CO^{2}A [41], and TA^{3}N [28] under FSDA-AR outperform the state-of-the-art performance of 87.8% achieved by TranSVAE for the UDA task, as shown in Table 5.3 ❶. This demonstrates that FSDA-AR is more efficient than UDA when dealing with small domain gaps.

Compared with the approach that has the best performance among all the baselines for the 1∼20 shot settings, RelaMiX achieves performance improvements of 1.7%, 2.2%, 1.4%, and 0.5% for FSDA-AR on UCF [177] → HMDB [96] and 0.4%, 2.1%, 0.4%, and 0.4% on HMDB [96] → UCF [177], respectively.

Similarly, consistent performance improvements are observed on the Sims4Action [156] → TSH [47] task. Compared with the best-performing baseline for the 1∼20 shot settings, RelaMiX achieves performance improvements of 2.8%, 1.2%, 3.1%, and 6.5%, as introduced in Table 5.1 ❹.

The consistent performance enhancements produced by RelaMiX across various datasets indicate that the proposed method effectively utilizes the guidance provided by the few-shot labeled samples from the target domain. Furthermore, RelaMiX can achieve a generalizable temporal aggregation that accounts for diverse domain differences.

In terms of video-based domain adaptation, the FSDA-AR task exhibits comparable performance to the UDA task for human action recognition, as reported in Table 5.3. Consequently, our experiments confirm the feasibility of FSDA-AR, and we believe it to be an essential future research direction in domain adaptation for human action recognition.

5.1.3.4 Ablation of Each proposed Module

Table 5.4: Module ablation on EPIC-KITCHEN [45] D1 → D2.

Method	Shot-1	Shot-5	Shot-10	Shot-20
w/o TARD-RD	36.3	40.3	40.9	44.1
w/o CDIA	37.9	41.3	40.7	47.2
w/o SDFM	34.7	42.5	42.3	45.3
w/ All	**39.1**	**43.9**	**43.7**	**47.9**

To assess the efficacy of each component of our RelaMiX method, we conduct ablation experiments on the EPIC-KITCHEN [45] dataset, specifically the D1 → D2 split, as detailed in Table 5.4.

First, we compare RelaMiX against RelaMiX without TARD-RD, employing Temporal Relation Networks (TRN) as an alternative for temporal aggregation. The results indicate that RelaMiX outperforms RelaMiX without TARD-RD, achieving performance improvements of 2.8%, 3.6%, 2.8%, and 3.8% across the 1∼20 shot settings. These findings highlight the superiority of our TRAN-RD method for temporal aggregation in the FSDA-AR task. The integration of relational attention with relation dropout and scale-wise self-attention is shown to be effective in facilitating generalizable temporal aggregation and feature learning.

Next, we compare RelaMiX with RelaMiX without the SDFM component. In this scenario, RelaMiX demonstrates superiority over its SDFM-lacking counterpart, exhibiting performance improvements of 4.4%, 1.4%, 1.4%, and 2.6% for the 1∼20 shot settings. These results indicate that SDFM effectively enhances the learned embeddings of the target domain.

Finally, we compare RelaMiX with RelaMiX without CDIA. RelaMiX outperforms the CDIA-lacking variant by 1.2%, 2.6%, 3.0%, and 0.7% for the 1∼20 shot settings, revealing that CDIA plays a significant role in bridging the domain gap by extracting pertinent information from a few target-domain samples. More ablations can be found in the supplementary material. Notably, the contributions of each component in the proposed solution vary across the different shot settings.

5.1.3.5 Analysis of Qualitative Results

Apart from the quantitative analysis, we also assess the qualitative results of the proposed FSDA-AR task. As shown in ①–⑧ in Fig. 5.3, we visualize ten action recognition results from the 20-shot Sims4Action → TSH setting, encompassing all ten classes used for FSDA-AR. For each sample, we compare the action predictions from FS-ADA [108], TA^{3}N [28], SSA^{2}Lign [204], PASTN [64], and our RelaMiX.

Given the significant domain gap between the synthesized and real datasets, most of the baselines do not guarantee superior performance. However, RelaMiX, which considers temporal generalizability, latent space diversity, and cross-domain alignment, demonstrates impressive performance in challenging settings. This showcases the effectiveness of our novel techniques in enhancing temporal aggregation generalizability.

Thanks to the proposed method, which accounts for temporal generalizability, latent space diversity, and cross-domain alignment, our RelaMiX achieves much better generalization ability and yields state-of-the-art results in few-shot domain adaptation for video data. Our RelaMiX approach correctly classifies eight out of ten samples. The qualitative results support the assumption that the model benefits from the proposed techniques and achieves generalizable temporal aggregation as well.

5.1.3.6 Ablation of the TRAN-RD

We present ablation experiments for the TRAN-RD in Table 5.5a. Here, *w/o RD-MHSA* means replacing RD-MHSA with a multi-layer perceptron (MLP), *w/o Scale-wise MHSA* means using mean average for multi-scale aggregation, and *w/o RD* indicates discarding the relation dropout.

First, comparing TRAN-RD with *w/o RD-MHSA*, we find that using RD-MHSA improves performance by 7.5%, 6.8%, 4.5%, and 8.7% for the 1∼20 shot settings, showcasing the importance of RD-MHSA in snippet-wise temporal information aggregation. Next, comparing TRAN-RD with *w/o Scale-wise MHSA*, we observe performance gains of 11.0%, 6.0%, 4.4%, and 8.6%, indicating the superiority of scale-wise information reasoning. Finally, comparing TRAN-RD with *w/o RD*, the relation dropout enhances performance by 6.8%, 4.4%, 6.4%, and 8.2% for the 1∼20 shot settings. This study demonstrates that each component of the module design collaborates to achieve a generalizable temporal aggregator for the FSDA-AR setting.

5.1.3.7 Ablation of the CDIA

We conduct ablation experiments in Table 5.5b for CIDA loss. Two ablations are performed for CDIA: *w/o prototypes*, where prototype-based positive anchors are replaced with randomly tempo-

Table 5.5: Ablation studies for TRAN-RD and CDIA on EPIC-KITCHEN D1 → D2.

a Module ablation for TRAN-RD.

Method	S-1	S-5	S-10	S-20
w/o RD-MHSA	31.6	37.1	39.2	39.2
w/o Scale-wise MHSA	28.1	37.9	39.3	39.3
w/o RD	32.3	39.5	37.3	39.7
w/ All	**39.1**	**43.9**	**43.7**	**47.9**

b Module ablation for CDIA.

Method	S-1	S-5	S-10	S-20
w/o prototypes	31.5	35.7	35.7	42.5
w/o mixed domain negatives	34.4	38.3	37.2	39.3
w/ All	**39.1**	**43.9**	**43.7**	**47.9**

Table 5.6: Ablation study of the SDFM and temporal aggregation comparison.

a Module ablation for SDFM for the K nearest clusters.

Method	S-1	S-5	S-10	S-20
K=1	33.6	38.4	39.1	44.5
K=3	32.7	38.0	38.5	43.6
K=4	34.9	39.3	38.8	43.9
K=2	**39.1**	**43.9**	**43.7**	**47.9**

b Comparison between TRAN-RD with other temporal aggregation methods.

Method	GFLOPS	S-1	S-5	S-10	S-20
LSTM	0.09	31.3	40.9	35.2	37.7
GRU	0.10	32.5	32.9	36.5	39.1
TRN	0.04	33.2	43.1	40.1	41.5
TRAN-RD	0.92	**39.1**	**43.9**	**43.7**	**47.9**

rally permuted anchor embeddings, and *w/o mixed domain negatives*, where mixed domain negatives are replaced with source-domain negative anchor embeddings. Using target domain prototypes as positive anchors improves performance by 7.6%, 8.2%, 8.0%, and 5.4% compared to *w/o prototypes*. Using mixed domain negatives improves performance by 4.7%, 5.6%, 6.5%, and 8.6% compared to *w/o mixed domain negatives*. These observations indicate that mixed domain negatives and target domain prototypes together provide more effective FSDA-AR supervision with few-shot target domain samples.

5.1.3.8 Ablation of the SDFM

We perform the ablation study for SDFM in Table 5.6a to investigate the influence of different numbers of cluster centers, with experiments conducted for $K \in 1, 2, 3, 4$. The setting $K = 2$ generally performs well for feature mixture. Using more cluster centers may results in less discriminative generated target domain embedding while using less cluster center is harmful to the distribution diversity, thereby $K = 2$ is selected in SDFM.

5.1.3.9 Comparison with Other Temporal Aggregators

We conduct an ablation study between TRAN-RD and other existing temporal aggregators, such as LSTM, GRU, and TRN [227], as shown in Table 5.6b. TRAN-RD outperforms all others by a large margin for the 1∼20 shot settings. Although TRN performs well in the 5∼20 shot settings, it

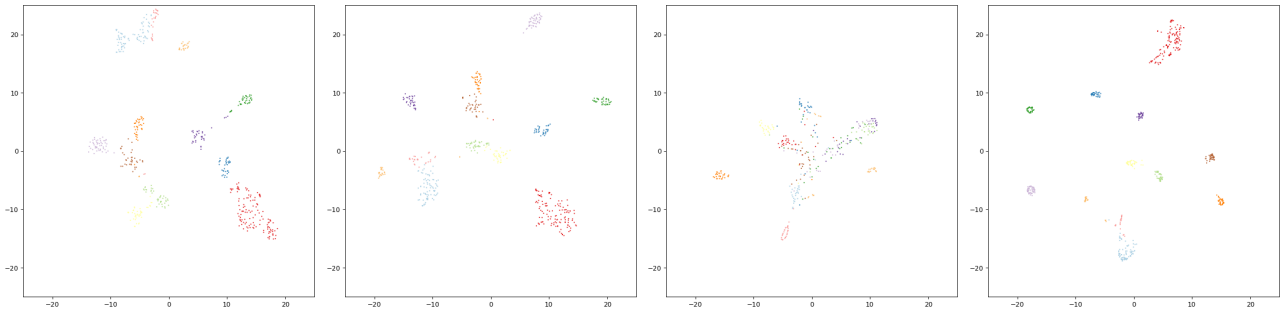
struggles with generalizable temporal aggregation with extremely small shot numbers, such as in the 1-shot setting. LSTM and GRU also face similar issues.

Our proposed TRAN-RD overcomes this difficulty using superior relation-based techniques for temporal aggregation in different scale settings. Alongside recognition performance, we also provide GFLOPS of different temporal aggregators to illustrate the computational complexity of our approach during inference. Since all baselines in our benchmark use I3D [19] as the feature extractor, which accounts for most of the computational complexity (108 GFLOPS), directly comparing the GFLOPS of the temporal aggregator is more revealing. Our TRAN-RD increases computational complexity by only 0.92 GFLOPS due to the RD-MHSA and Scale-wise MHSA mechanisms. This increase is minimal compared to the 108 GFLOPS of the I3D backbone. Moreover, CDIA and SDFM only participate in the training phase and do not contribute to computational complexity during inference.

5.1.3.10 Analysis of the Target Domain Sample Number of FSDA-AR and UDA

We present the required sample number to construct the training set on the target domain separately for FSDA-AR and UDA tasks across all leveraged DA settings in Table 5.7. Compared to UDA, FSDA-AR requires significantly less data from the target domain for training. Since labeling for action recognition does not require pixel-wise annotation and each sample only needs one label, data collection may take more time than labeling.

For example, on Sims4Action $\rightarrow$ TSH, FSDA-AR discards 97.7% of the target domain samples used in UDA, while delivering better performance. The comparable performance of FSDA-AR to UDA indicates that FSDA-AR is a more efficient setting, especially when data collection in the target domain is challenging. We emphasize that FSDA-AR is an important research direction, and our work serves as a crucial test bed for future studies in this area.



(a) FS-ADA **(b)** PASTN **(c)** SSA^{2}Lign **(d)** Ours

Figure 5.4: The t-SNE feature visualization [130] on the UCF test set [177] for FSDA-AR on 20−Shot HMDB [96] $\rightarrow$ UCF [177].

Table 5.7: Analysis of the sample number that is required for FSDA-AR and UDA under each setting.

Setting	FSDA				UDA
	S-1	S-5	S-10	S-20	
UCF [177]→HMDB [96]	12	60	120	240	840
HMDB [96]→UCF [177]	12	60	120	240	1438
Sims4Action [156]→TSH [47]	10	50	100	200	8552
D1→D2 [45]	8	40	80	160	2495
D2→D1 [45]	8	40	80	160	1543
D2→D3 [45]	8	40	80	160	3897
D3→D2 [45]	8	40	80	160	2495
D1→D3 [45]	8	40	80	160	3897
D3→D1 [45]	8	40	80	160	1543

5.1.3.11 Analysis of the t-SNE Visualization

To investigate the performance of few-shot domain adaptation in the latent space, we present the t-SNE distribution [130] in Fig. 5.4. Figures (a) to (d) show the t-SNE distributions for FS-ADA, PASTN, TA^{3}N, and our RelaMiX, respectively.

Under the setting of HMDB [96] → UCF [177], the samples shown in Fig. 5.4 are selected from the UCF101 test set. Compared to the other methods, the features from our RelaMiX are more distinguishable across different classes in the latent space, demonstrating a better generalization ability of RelaMiX, which is crucial in FSDA-AR.

5.1.4 Comparison with Other Domain Adaptation Settings

We begin by distinguishing the few-shot domain adaptation task from other commonly used domain adaptation settings. The task we address is Few-Shot Domain Adaptation (FSDA), where only a small number of labeled examples are available for each category in the target domain. The comparison among Semi-Supervised Domain Adaptation (SSDA) [159, 210, 216], Unsupervised Domain Adaptation (UDA) [28, 37, 87, 158, 198].

Unlike FSDA, both UDA and SSDA require a large amount of unlabeled data in the target domain to construct the training set, which may not always be feasible due to high-quality data collection constraints. FSDA, however, requires only a small number of labeled examples from the target domain, balancing data collection expenses and labeling efforts. In our benchmark, the few-shot setting results in a 70% to 98% reduction in the target domain data needed to construct the training set compared to SSDA and UDA. Since action recognition does not require pixel-level dense annotations, the trade-off between annotation and data collection is significant.

It is important to note that FSDA is distinct from domain adaptation for few-shot learning [39,

221], which focuses on adapting to new classes with limited examples. In FSDA, the goal is to adapt to new domains without introducing new action classes.

5.1.5 Discussion

This work makes significant exploration in FSDA-AR by addressing its distinct challenges and introducing a novel framework, RelaMiX, to enhance adaptation across diverse domains by effectively using few labelled target domain samples. Initially, we establish a new FSDA-AR benchmark using five well-known datasets: Sims4Action, ToyotaSmartHome, EPIC-KITCHENS, HMDB, and UCF. This benchmark is meticulously designed to evaluate the performance of FSDA-AR methods across a variety of domain adaptation settings, encompassing transitions from synthetic to real-world data, cinematic to real-life third-person views, and scenario changes in egocentric perspectives. By covering such a broad spectrum, the benchmark provides a rigorous testbed for assessing the robustness and adaptability of FSDA-AR methods.

This study reveals the limitations of existing domain adaptation techniques when applied to FSDA-AR tasks, particularly highlighting their struggles with generalization due to the scarcity of labeled data in the target domain. To overcome these challenges, we propose RelaMiX, a novel approach that integrates several advanced mechanisms to maximize the utility of limited labeled target samples and ensure more reliable domain adaptation.

RelaMiX incorporates a TRAN-RD to enhance the generalization of temporal feature aggregation. This new proposed temporal aggregation method improves the model's ability to capture and transfer temporal patterns across different domains by utilizing relation dropout and multi-scale self-attention mechanisms on both of the relation set and relation scale perspectives. This approach ensures that the temporal relationships within the data are effectively learned and generalized, which is critical for accurate action recognition.

Additionally, the framework includes SDFM, which enhances the latent space by blending features from both source and target domains. This mixture is achieved by calculating the statistical properties of the source domain features and generating additional target domain embeddings. This method enriches the latent space, promoting better feature diversity and improving the model's capacity for generalizable and discriminative feature learning.

To further bridge the domain gaps, RelaMiX employs CDIA loss, which uses few-shot samples from the target domain to align the feature distributions of the source and target domains. This alignment is achieved through a contrastive learning approach that minimizes the distance between similar samples from different domains while maximizing the distance between dissimilar ones. This mechanism ensures that the model can effectively transfer knowledge from the source domain to the target domain, even with minimal labeled data.

The experimental results demonstrate the effectiveness of RelaMiX, showing significant performance improvements across all datasets in the FSDA-AR benchmark. RelaMiX consistently outperforms existing FSDA-AR and UDA methods, achieving state-of-the-art results in various challenging settings. For instance, on the EPIC-KITCHENS dataset, RelaMiX achieves performance improvements of up to 2.8% in the 1-shot setting and maintains superior results across higher shot settings. These findings highlight the framework's ability to leverage few-shot target domain samples effectively, enhancing both open-set and close-set recognition capabilities.

By addressing the critical issues of temporal generalization, feature diversity, and domain alignment, this study provides a generalizable and versatile method for FSDA-AR. The proposed methods significantly enhance the reliability and accuracy of action recognition methods in real-world scenarios where labeled data is limited. This work lays a strong foundation for future research in FSDA-AR, suggesting practical applications in various fields where data labeling is labor-intensive or costly.

5.2 Towards Privacy Support RGB2Depth Domain Adaptation for Fall Detection

5.2.1 Introduction and Motivation

According to United Nations' predictions, 13% of the global population was aged 60 or elderly people [113], highlighting the importance of developing responsible technologies to support and assist the elderly. Falls pose a major danger, not only causing physical harm to elderly adults but also to young individuals who live alone. As reported by the World Health Organization[1], falls result in approximately 684,000 deaths annually, with 37.3 million falls severe enough to require medical attention.

Various approaches exist for fall detection, such as wearable equipment [1, 7, 16, 26, 111, 206, 215], using Wi-Fi signals [23, 48, 78, 191, 197], or video monitoring systems [6, 14, 57, 93, 183, 207, 226]. Video-based approaches are physically practical as they do not impose burdens on users or require complex operations compared to wearable devices [197]. These methods typically build on established action recognition models to achieve accurate results [6].

Most current datasets and methodologies for fall detection rely on RGB data [57]. However, privacy concerns have grown, and using RGB data has been criticized for potentially revealing detailed personal information. Therefore, there is increasing interest in privacy-supporting frameworks.

[1] https://www.who.int/news-room/fact-sheets/detail/falls

Depth data, or 3D data, represents the distance of objects from a camera or sensor without preserving detailed texture information, enhancing privacy while providing valuable information for fall detection. A fall detection method leveraging depth data at test time would be preferable if it achieves accurate results. However, existing depth-based datasets for fall detection are relatively small, providing limited data for training action recognition deep learning networks. Given the different privacy-preserving abilities of different modalities, users might choose different modalities at test time according to their needs. Cross-modal adaptation thus becomes an important research direction for fall detection, allowing the use of well-established models pretrained on large-scale RGB datasets to achieve depth-based fall detection at test time. In this work, we focus on using labeled RGB data and unlabeled depth data for training and transferring knowledge from the RGB domain to the depth domain (RGB2Depth), an area overlooked in fall detection research.

Since most depth-based fall detection datasets are small-scaled and insufficient for training and testing while video-based approaches, such as X3D [61], require large-scale pretraining for convergence, we reformulate and adopt the Kinetics dataset [89] for unsupervised domain adaptation in the RGB2Depth fall detection task. A subset of the data is converted to depth data through P^2Net [218] to provide sufficient test samples, and well-established RGB-based pretrained weights are used to initialize the models.

To bridge the gap between the RGB and depth domains for fall detection, we establish the cross-modal unsupervised domain adaptation pipeline UMA-FD using the X3D [61] model, a promising backbone for accurate action recognition. We utilize an intermediate domain module [44] to bridge RGB and depth representations, and employ multiple losses to constrain the latent space, such as an adversarial modality discrimination loss, triplet margin losses on the two domains, and classification loss on the source RGB data and pseudo-labeled depth data.

Since different losses contribute in various ways, a fixed scheme for weighting the losses may restrict the learning process during different stages. Therefore, we propose an adaptive weighting approach for the loss functions, using an additional multi-layer perceptron-based head to predict weighting parameters. Our contributions are summarized as follows:

- We propose the RGB-to-Depth (RGB2Depth) unsupervised domain adaptation task in fall detection and develop a new multi-source dataset and benchmark protocol.
- We introduce a new pipeline to address this task, employing 3D-CNN+LSTM [128], C3D [186], I3D [19], and X3D [61] as feature extraction backbones. We utilize the intermediate domain module, modality adversarial alignment, and triplet margin loss to minimize the cross-modal domain gap, and propose an adaptive weighting method for balancing the loss functions.
- Our model, UMA-FD, achieves state-of-the-art performance on the proposed RGB2Depth UDA

Table 5.8: Comparison with other fall detection dataset. For NTU-60 with multiple non-falling actions, we take the same number of other samples with falling action samples.

Datasets	Fall Samples	Other Samples	Total Number
UR Fall Detection [98]	30	40	70
NTU-60 [161]	948	948	1,896
Our Dataset	1,490	1,489	2,979

Table 5.9: The number of samples in our dataset.

	Training Set	Test Set	Total Number
Positive Sample	falling off bike /678 falling off chair /612	falling off bike /100 falling off chair /100	1,490
Negative Sample	washing hands /644 sweeping floor /645	washing hands /100 sweeping floor /100	1,489
Total Number	2,579	400	2,979

task compared to existing fall detection methods. Ablation studies demonstrate the efficiency of each proposed component within our framework.

5.2.2 Dataset

To study cross-modal unsupervised domain adaptation from RGB2Depth for fall detection, we require multi-modal data that includes both RGB and depth information in fall scenarios. The Kinetics-700 video dataset [18] contains 650,000 video clips spanning 700 human action classes, with each clip annotated with an action class and lasting approximately 10 seconds. From the Kinetics-700 dataset, we select two categories related to falling actions: the *falling off bike* class and the *falling off chair* class. These videos serve as positive samples for fall detection. Additionally, we randomly select two categories, *washing hands* and *sweeping floor*, as negative samples for fall detection.

The four categories of videos in our dataset initially contain only RGB data. To generate the corresponding depth data, we use advanced depth estimation algorithms, specifically P^2Net [218], to produce 288x384 depth data for each frame. To standardize the format, each RGB video frame is resized to 256x256, resulting in a labeled fall detection dataset containing both RGB and depth modalities.

Given the imbalance in sample numbers across the four categories and our goal for an equal number of positive and negative samples, we randomly sample the data from each category. Our dataset is compared with two existing datasets that include RGB videos and corresponding depth data for fall detection, as shown in Table 5.8. The UR Fall Detection dataset [98] has only 30 falls and 40 activities of daily living sequences, insufficient for training a deep model. The NTU-60 dataset [161], comprising 60 action classes including 948 fall videos, also includes indoor scenes, limiting its va-

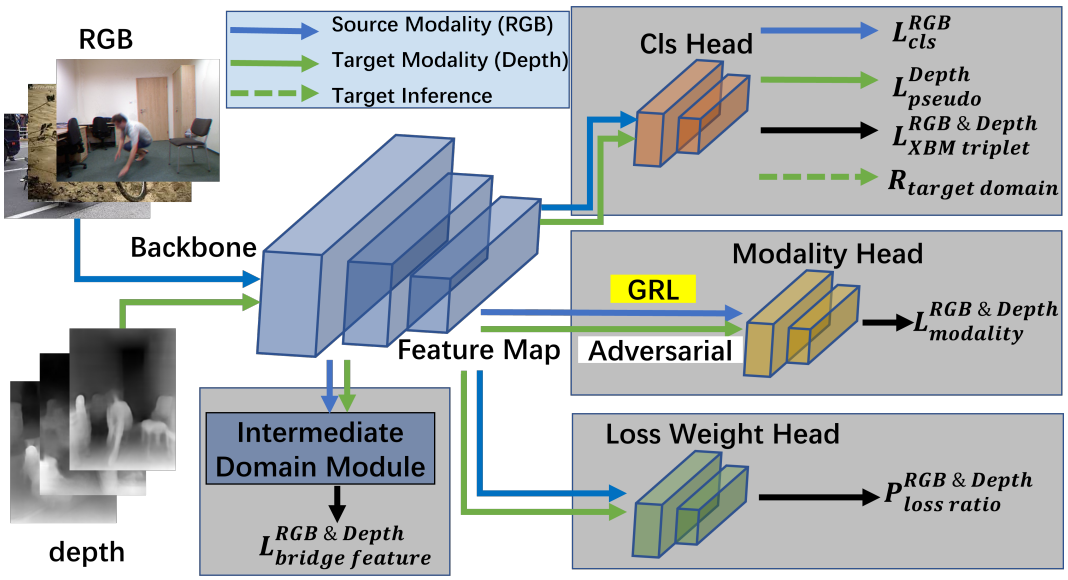


Figure 5.5: An overview of the proposed architecture.

riety. In contrast, our dataset contains 1,490 fall samples and 1,489 other samples, offering a larger and more diverse collection of scenes.

We then divide our dataset into training and test sets. The distribution of samples for each category after sampling is provided in Table 5.9. Our dataset includes 2,979 samples, with 1,490 positive and 1,489 negative samples. We randomly select 100 samples from each category for the test set, resulting in 2,579 training samples and 400 test samples, each containing both RGB and corresponding depth data. Additionally, we use the NTU-60 dataset [161] in later experiments to enhance the validity of our results.

5.2.3 Proposed Method

This section outlines our proposed method for RGB2Depth fall detection, which we refer to as *Unsupervised Cross-Modal Adaptation for Fall Detection (UMA-FD)*. Fig. 5.5 presents an overview of the UMA-FD method, using cross-modal unsupervised domain adaptation to transfer knowledge from the RGB source modality to the depth target modality. In UMA-FD, we preprocess the RGB and depth data to generate a compatible input format and use a unified backbone network to generate feature maps for both data streams.

In order to achieve effective cross-modal feature learning, we incorporate an Intermediate Domain Module (IDM) [44] to generate an intermediate modality feature map, and then compute the bridge feature loss [44]. The classification layers of our network consist of three heads: the label classification head, the modality agnostic head, and the loss weight adaptive head. Next, we define the unsupervised modality adaptation problem and describe each building block of our proposed

method in detail.

5.2.3.1 IDM and Bridge Feature Loss

During the training process, $\mathbf{v}^R$ and $\mathbf{v}^D$ are randomly paired as inputs to the backbone network. The IDM accepts data from both modalities as input and generates an intermediate modality feature map by weighted summation of the feature maps from the two modalities. The weighting coefficients are obtained through adaptive learning within the network.

IDM can be added between any two hidden layers of the backbone, generating the intermediate modality to bridge the distinct two modalities. The data from this intermediate modality and the other modalities is then fed into the subsequent backbone network layers. The intermediate modality embeddings generated by IDM can be represented as the following two equations.

$$\mathbf{f}_a = SoftMax(\mathbf{M}_\alpha(\mathbf{M}_{\beta_1}(Concat(\mathbf{f}^R_{h_{avg}}, \mathbf{f}^R_{h_{max}})) + \mathbf{M}_{\beta_2}(Concat(\mathbf{f}^D_{h_{avg}}; \mathbf{f}^D_{h_{max}})))), \tag{5.19}$$

$$\mathbf{f}^{\text{inter}} = \mathbf{f}^R_a \cdot \mathbf{f}^R_h + \mathbf{f}^D_a \cdot \mathbf{f}^D_h. \tag{5.20}$$

where $\mathbf{f}_h$ represents the feature map of the hidden layer, subscript *avg* and *max* denote average pooling and max pooling, $\mathbf{M}_\beta$ denotes fully connect layer, $\mathbf{M}_\alpha$ stands for multi-layer perceptrons. In our case, different backbones are employed, and we always add the IDM module after the first convolution block.

The backbone network generates the final feature maps for the RGB, depth, and intermediate modalities. We employ the bridge feature loss [44] to constrain the weighted sum of distances between the intermediate modality feature map and those of the RGB and depth modalities. The weighted sum employs the weighting coefficients derived from the IDM module. This ensures that when the RGB modality significantly influences the intermediate modality, the bridge feature loss emphasizes the distance between the RGB and intermediate modalities. Similarly, this procedure applies to the depth modality.

The bridge feature loss is computed as Eq. 5.21.

$$\mathscr{L}^{R\&D}_{bridge} = \frac{1}{N_B}\sum_{i=1}^{N_B}\sum_{k\in\{R,D\}}\left[a^k_i \cdot ||\mathbf{f}^k_i - \mathbf{f}^{inter}_i||_2\right], \tag{5.21}$$

where $\mathbf{f}$ is the feature map of the final output of the backbone, a is the weighting coefficient generated by the IDM module, and $||\cdot||$ represents the L2 norm to calculate the spatial Euclidean distance between two feature maps.

The bridge feature loss ensures that the feature map of the intermediate modality lies between

the RGB and depth modalities in the spatial distribution, thus constraining the backbone to learn appropriate feature maps for both modalities.

5.2.3.2 Unsupervised Modality Adaptation

Even though the source and target modalities are different, the underlying clues used for fall detection have strong potential correlations. In this context, supervised training on the source modality can help uncover informative cues in the target modality. Motivated by this, our method minimizes both the classification loss of the source modality and the distribution discrepancy between the source and target modalities.

Following the backbone, a supervised classification head is constructed using a fully connected network. For the RGB modality data, which has associated sample labels, we compute the cross-entropy loss according to Eq. 5.22,

$$\mathscr{L}_{cls}^{R} = \frac{1}{N_B} \sum_{i=1}^{N_B} -y \log p(\mathbf{v}^R), \tag{5.22}$$

where $p(\mathbf{v}^R)$ denotes the prediction from the classification head when $\mathbf{v}^R$ is ultilized as input to the neural network. Since there are no labels available for the depth data, a threshold-based pseudo-labeling technique is employed for supervision, allowing us to obtain the classification loss on the depth data. With this approach, we estimate pseudo-labels for the depth samples that meet the threshold condition according to Eq. 5.23.

$$Y_{pseudo}^{D} = \begin{cases} 0, & Sigmoid(p(\mathbf{v}^D)) \leq \tau \\ \\ 1, & Sigmoid(p(\mathbf{v}^D)) > \tau \end{cases} \tag{5.23}$$

where τ is the pseudo label threshold. We then calculate the corresponding pseudo-label cross-entropy loss as follows Eq. 5.24.

$$\mathscr{L}_{pseudo}^{D} = \sum_{\mathbf{v}^D \in T^{part}} -y^{pseudo} log p(\mathbf{v}^D). \tag{5.24}$$

where T^{part} represents the depth samples set that meets the threshold condition.

Triplet loss [75] is a commonly used loss for metric learning, and can be also used to enhance the discriminative capability of the learnt embeddings. Due to the use of a 3D convolutional network model for video data and the limited memory of a single GPU, the batch size of training data is quite small. This makes it challenging to compute the triplet loss within a single batch. The cross-batch

memory mechanism (XBM) [196] addresses this issue by serving as a module that stores previously processed training data during the training process. During the training process, the XBM component retains the feature maps from previous batches. Triplet loss is then calculated based on the current batch's embeddings and those stored in the XBM. As a result, we obtain the XBM_triplet loss ($\mathscr{L}^{R\&D}_{XBM\_triplet}$), which can be represented as Eq. 5.25.

$$\mathscr{L}^{R\&D}_{XBM\_triplet} = max(d(\mathbf{f}_{cur}, \mathbf{f}^{p}_{pre}) - d(\mathbf{f}_{cur}, \mathbf{f}^{n}_{pre}) + margin, 0). \tag{5.25}$$

where $\mathbf{f}_{cur}$ represents the embeddings of the current sample, and $\mathbf{f}^{p}_{pre}, \mathbf{f}^{n}_{pre}$ respectively represent the previous sample embeddings with the same label and the current sample with different label, respectively. The $\mathscr{L}^{R\&D}_{XBM\_triplet}$ constraints the maximum distance of embeddings between samples with same label smaller than the minimum distance between samples with different label, which improves the discriminative ability of the learned latent representations on both of the depth and RGB modalities.

5.2.3.3 Modalities Adversarial Alignment

In unsupervised domain adaptation, both generative and discriminative adversarial approaches have been proposed for bridging the distribution discrepancy between source and target domains. For high-dimensional data streams, such as video, discriminative approaches are more suitable [141]. Discriminative methods train a discriminator, $\mathbf{M}_D(\cdot)$, to predict the modality of an input from the learnt features from the backbone $\mathbf{M}_H(\cdot)$. By maximising the discriminator loss, the network learns a feature representation that is invariant to both modalities.

In our scenario, to align the RGB and depth data, we propose a modality discriminator that penalizes feature variability between the modalities. The modality discriminator, $\mathbf{M}_D(\cdot)$, contains a Gradient Reversal Layer (GRL) [105] and a fully connected network to learn the modality representation. Given a binary modality label, y_m, indicating if a sample $\mathbf{v}$ belongs to the RGB or depth domain, we propose the following modality agnostic loss as described in Eq. 5.26.

$$\mathscr{L}^{R\&D}_{modality} = \sum_{k \in \{R,D\}} -y_m log(\mathbf{M}_D(\mathbf{M}_H(\mathbf{v}^k))) - (1 - y_m) log(\mathbf{M}_D(\mathbf{M}_H(\mathbf{v}^k))). \tag{5.26}$$

The $\mathscr{L}^{R\&D}_{modality}$ loss reduces the variance between different modalities in the backbone's feature maps. This ensures that the features trained on the RGB data become more applicable to the depth data.

5.2.3.4 Total Loss and Loss Weight Adaptation

To summarize the aforementioned components, the final loss can be expressed as Eq. 5.27,

$$\mathscr{L} = \lambda_a \mathscr{L}^{R}_{cls} + \lambda_b \mathscr{L}^{D}_{pseudo} + \lambda_c \mathscr{L}^{R\&D}_{modality} + \lambda_d \mathscr{L}^{R\&D}_{bridge} + \lambda_e \mathscr{L}^{R\&D}_{XBM_{triplet}}. \tag{5.27}$$

where $\lambda_a, \lambda_b, \lambda_c, \lambda_d, \lambda_e$ are the proportional coefficients. During the training process, the overall network consists of five losses, and the impact of each loss on the final depth data's classification accuracy is unknown. Therefore, we need to adjust the values of $\lambda_a, \lambda_b, \lambda_c, \lambda_d, \lambda_e$ accordingly. Manual adjustment can be time-consuming and may not yield an optimal combination. To address this problem, we consider using an adaptive network to automatically learn the loss weights, aiming to obtain the optimal solution. The loss weight adaptive network $\mathbf{M}_W(\cdot)$ consists of a three-layer fully connected network with corresponding activation functions. The network output is a five-dimensional weight coefficient as follows,

$$[\lambda_a, \lambda_b, \lambda_c, \lambda_d, \lambda_e] = SoftMax(\mathbf{M}_W(\mathbf{M}_H(\mathbf{v}))). \tag{5.28}$$

5.2.4 Experiments and Results

In this section, we first discuss the implementation details and evaluation metrics. Then, we evaluate our proposed method UMA-FD and compare the results with the baseline and the supervised target method for fall detection on the NTU-60 dataset [161] and our dataset. To enhance the validity of the results, we compare the performance of a fall detection backbone 3D-CNN+LSTM [128] and three other CNN-based backbones: C3D [186], I3D [19], and X3D [61]. Next, we discuss the results of various ablation experiments. Finally, we present qualitative results of UMA-FD and analyze the classification results of some samples.

5.2.4.1 Implementation Details and Evaluation Metrics

During training, we use the dataset described in Section 5.2.2. The NTU-60 dataset includes 948 videos of falls (positive samples) and 948 videos of other actions (negative samples). We use 180 positive and 180 negative samples for testing, with the rest for training. Additionally, we create a dataset from the Kinetics-700 database, generating RGB and corresponding depth data for each frame using a depth estimation model. The training set includes 5,158 samples (2,579 labeled RGB and 2,579 unlabeled depth clips). The test set has 400 depth samples, equally split between positive and negative examples. Experiments are conducted on two NVIDIA GeForce RTX 3090 GPUs with 24GB memory. We use mmaction2 [40] and pre-trained models C3D [186], I3D [19], and X3D [61]. Model

parameters for ablation experiments are based on previous best configurations. We use SGD [3] with a momentum of 0.9. Initial learning rates are 0.0001 for I3D and X3D, and 0.001 for C3D. Models are trained for 120 epochs, with the learning rate decaying by one-tenth after 60 epochs. Pseudo-label thresholds are 0.8 for 3D-CNN+LSTM and C3D, I3D, and 0.7 for X3D. Evaluation metrics include accuracy, F1 score, and AUC. Metrics are calculated based on predicted labels.

5.2.4.2 Comparison with Baseline and Supervised Target Method

Since fall detection through cross-modal unsupervised adaptive learning is a novel problem, we need to define a baseline for this new task to justify the soundness of our proposed method. The most straightforward approach, utilizing the concept of transfer learning, involves training a fall detection model on the labeled RGB data and then directly predicting using the unlabeled depth dataset. This method is used as the baseline for comparison in this study. Additionally, we can obtain results from a supervised target method, in which we assume the depth data labels are known. In this supervised target method, we use both labeled RGB and depth data training sets for training. We then validate the performance of the model on the test set of depth data. This target supervised method is regarded as the upper bound of this task. To verify the cross-backbone generalizability of our proposed method, comparative experiments are conducted on 3D-CNN+LSTM [128], C3D [186], I3D [19], and X3D [61] by implementing our proposed mechanisms on those backbones.

The results comparison on the NTU-60 dataset and our generated Kinetics dataset are provided in Table 5.10a and Table 5.10b, respectively. On the NTU-60 dataset, compared with the baselines using 3D-CNN+LSTM, C3D, I3D, and X3D backbones, the accuracy of UMA-FD increases by 10.83%, 5.95%, 6.66%, and 10.83%, the F1 scores increases by 45.05%, 9.06%, 9.89%, and 10.61%, and the AUC increases by 5.96%, 20.16%, 20.77%, and 6.63%, respectively. Note that 3D-CNN+LSTM is a well-established fall detection backbone proposed by [128], while the others are backbones designed for general action recognition. On our generated Kinetics dataset, compared with the baseline using the fall detection backbone, *i.e.*, 3D-CNN+LSTM, and conventional action recognition backbones, *i.e.*, C3D, I3D, and X3D, the accuracy of UMA-FD increases by 3.25%, 4.75%, 4.25%, and 5.5%, the F1 scores increases by 10.56%, 5.14%, 9.19%, and 8.61%, and the AUC increases by 4.19%, 3.92%, 3.12%, and 8.61%, and the AUC increases by 4.19%, 3.92%, 3.12% and 4.26%, respectively. The results consistently demonstrate that our cross-modal unsupervised adaptation method achieve superior performance on various datasets and various feature extraction backbones for the RGB2Depth unsupervised domain adaptation task. This improvement is independent of the specific backbone used, highlighting the versatility of our method.

The I3D backbone, which has fewer parameters, achieves results comparable to the C3D backbone. In contrast, the X3D backbone, being one of the most promising backbones for action recog-

Table 5.10: Experimental results on NTU-60 and Kinetics datasets.

a Experimental results on NTU-60 dataset.

methods	**3D-CNN+LSTM**		
	Accuracy	F1 Score	AUC
Baseline	57.78	28.97	74.49
UMA-FD	**68.61**	**74.02**	**80.45**
Supervised Target	94.44	94.38	98.55
methods	**C3D**		
	Accuracy	F1 Score	AUC
Baseline	60.32	65.03	62.90
UMA-FD	**66.27**	**74.09**	**83.06**
Supervised Target	95.96	96.33	98.37
methods	**I3D**		
	Accuracy	F1 Score	AUC
Baseline	59.17	63.88	61.20
UMA-FD	**65.83**	**73.77**	**81.97**
Supervised Target	95.83	95.73	98.16
methods	**X3D**		
	Accuracy	F1 Score	AUC
Baseline	83.89	84.16	91.63
UMA-FD	**94.72**	**94.77**	**98.26**
Supervised Target	98.33	98.33	99.50

b Experimental results on Kinetics dataset.

Methods	**3D-CNN+LSTM**		
	Accuracy	F1 Score	AUC
Baseline	61.25	57.88	65.74
UMA-FD	**64.50**	**68.44**	**69.93**
Supervised Target	73.25	72.21	81.18
Methods	**C3D**		
	Accuracy	F1 Score	AUC
Baseline	67.00	64.89	72.67
UMA-FD	**71.75**	**70.03**	**76.59**
Supervised Target	79.75	80.20	85.43
Methods	**I3D**		
	Accuracy	F1 Score	AUC
Baseline	68.00	62.13	74.70
UMA-FD	**72.25**	**71.32**	**77.82**
Supervised Target	79.25	78.66	86.33
Methods	**X3D**		
	Accuracy	F1 Score	AUC
Baseline	73.00	70.00	81.79
UMA-FD	**78.50**	**78.61**	**86.05**
Supervised Target	92.50	92.43	97.90

nition, yields significantly better results than the other two backbones. However, when compared to the supervised target method, the accuracy of UMA-FD is still lower due to the absence of label information for depth data.

There remains room for improvement in the accuracy of our method, which will be the focus of future research. Given that the X3D backbone produces the best results, all subsequent experiments in the abkation will be conducted using the X3D backbone.

5.2.4.3 Ablation Study

Next, we analyze the individual contributions of different components of UMA-FD. Using the X3D backbone, we add various module parts and loss functions of our proposed method based on the baseline method to conduct ablation experiments. The results, detailed in Table 5.11, illustrating the contribution of different building blocks and corresponding loss functions. For convenience, each experiment is numbered; for instance, the baseline is denoted as V-01.

First, we add the modality agnostic head and modality adversarial loss to the baseline model, referred to as V-02. The accuracy increases from 73.00% to 75.25% compared to V-01, an absolute improvement of 2.25%, with significant improvements in F1 score and AUC. These results indicate

Table 5.11: Ablation of our proposed method UMA-FD, showing the contribution of the various module and corresponding loss functions.

Method	Accuracy	F1 Score	AUC
Baseline (V-01)	73.00	70.00	81.79
+Modality Loss (V-02)	75.25	74.02	83.04
+Pseudo Loss (V-03)	76.25	78.16	83.05
+Bridge Feature Loss (V-04)	77.25	75.34	84.67
+XBM_riplet Loss (V-05)	77.75	77.00	84.10
UMA-FD (V-06)	**78.50**	**78.61**	**86.05**

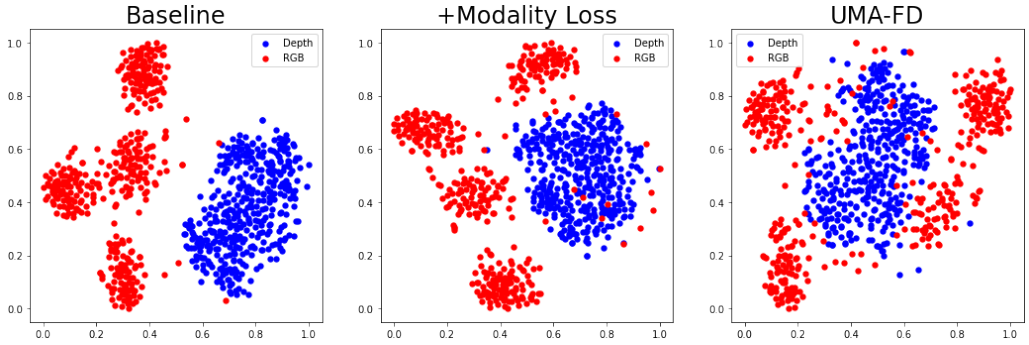


Figure 5.6: T-SNE plots of RGB data (red points) and Depth data (blue point) feature spaces produced by baseline, the method of adding modality loss and our proposed method UMA-FD.

that the modality agnostic head and modality adversarial loss effectively mitigate the differences between the features of different modalities, allowing more information learned from RGB data to be applied to depth data. This conclusion will be further confirmed by the qualitative results in the next section.

Building upon the optimal model from V-02, we set a threshold to distinguish positive and negative sample pseudo-labels and incorporate a pseudo loss, denoted by the V-03 model. The V-03's accuracy further improves, reaching 76.25%, a 1% increase compared to V-02. The F1 score also shows a significant improvement, and the AUC is slightly better than the previous version, demonstrating an overall noticeable enhancement. V-03 capitalizes on V-02's ability to classify depth data and employs partial pseudo-label information with higher confidence, enabling the model to learn more depth data information.

Next, we verify the effect of the bridge feature loss based on the optimal model of V-03. We add the IDM and bridge feature loss (V-04), leading to an accuracy gain from 76.25% to 77.25%. Adjusting the threshold improves the outcome significantly: the AUC increases to 84.67% compared to 83.05% in V-03. Overall, the performance has been improved, indicating that the IDM module and bridge feature loss indeed reduce the difference in representations between different modalities.

Next, we add $XBM_{triplet}$ loss to V-04, referred to as V-05. The $XBM_{triplet}$ loss also requires the labels of samples. For the RGB sample data, the real label is used directly. For the unlabeled depth data, the pseudo label that meets the threshold is used to calculate the triplet loss. Comparing the optimal results of V-05 with V-04, the accuracy increases from 77.25% to 77.75%. The $XBM_{triplet}$ loss benefits the model by learning discriminative features for both modalities. A comprehensive comparison of accuracy with F1 score and AUC shows that V-05 outperforms V-04, demonstrating that the $XBM_{triplet}$ loss is also effective for cross-modal fall detection. Finally, we verify the effectiveness of the loss weight adaptation method. Based on the optimal model from V-05, we incorporate the loss weight adaptive network to automatically learn the weight of each loss. This model, numbered V-06, represents our proposed method, UMA-FD. The accuracy increases from 77.75% to 78.5%, an absolute improvement of 0.75%. The F1 score and AUC are the best among all experimental results, showcasing a clear enhancement. This demonstrates that the superiority of the proposed loss weight adaptive network for the training assistance. The ablation experiments provide a comprehensive verification of each individual module. By employing cross-modal unsupervised adaptive learning, the classification accuracy of the unlabeled depth data is increased from 73% (baseline) to 78.5%, a significant gain of 5.5%. The above analyses indicate that the proposed UMA-FD can achieve superior RGB2Depth domain adaptation performance and show large performance improvement compared withe the baseline. The proposed method can serve as a good solution in real-world application for cross-modal fall detection task.

5.2.4.4 Analyses of the T-SNE Results

We present the t-SNE visualization of the RGB data and depth data feature spaces $M(\cdot)$ generated by the baseline method, the method incorporating modality loss, and our proposed method UMA-FD in Fig. 5.6. It is evident that our proposed method mitigates the differences between the source and target modalities to some extent.

In the baseline method, where the RGB model is directly used to predict depth data, the resulting feature distributions differ significantly. Adding modality loss to mitigate these differences improves the situation considerably. In our final proposed method UMA-FD, the feature distributions are essentially mixed, which is the desired outcome. Therefore, we can more effectively utilize the information learned from the RGB data when applied to the depth data, thereby improving the classification accuracy of depth data. However, the feature space distributions of the two modalities' data still differ significantly, and even with our proposed method UMA-FD, these differences are not completely eliminated. Identifying ways to further reduce these differences will be an important direction for future work in this task.

5.2.5 Discussion

In this work, we addressed the RGB to depth unsupervised domain adaptation for fall detection task, extending unsupervised domain adaptation to meet specific application needs. We generated a dual-modality fall detection dataset with RGB and depth data, comprising $2,979$ samples, surpassing most existing fall detection datasets in the scale of the dataset. This dataset was created using the public Kinetics-700 and NTU-60 datasets and an off-the-shelf depth estimation algorithm. To enhance classification accuracy for unlabeled depth data, we applied various UDA methods to the RGB to Depth unsupervised domain adaptation scenario, achieving scene adaptation.

In single-task multi-loss scenarios, manually adjusting loss weights is inefficient and suboptimal. To solve this, we designed a loss weight adaptive network that automatically learns each loss's weight. Integrating these optimization methods significantly improved our model's performance, raising classification accuracy from 73% (baseline) to 78.5% on our generated Kinetics-700 dataset, demonstrating the feasibility of cross-modal unsupervised adaptive learning.

This study shows the superiority of the modality agnostic loss, modality agnostic head, cross-batch triplet margin loss, and the pseudo label based target domain supervision method on the RGB2Depth unsupervised fall detection task, which is an essential application of human action recognition. However, due to the large performance difference between our proposed method and the upper boundary by using supervised target domain for training, future work on this task is expected, *e.g.*, using foundation model.

6 Conclusions and Remarks

6.1 Impact to the Community

The contributions made in this thesis have a significant impact in the field of human action recognition, which involve the observation of limitations of the existing works and the proposal of solutions from multiple perspectives to explore how to achieve more generalizable deep learning in the field of human action recognition. This thesis first highlights the significant challenge posed by diverse occlusions, which are common real-world perturbations, on skeleton data in tasks such as one-shot and self-supervised action recognition. They point out that most current methods in these fields struggle with occlusion perturbations, demonstrating limited performance. This underscores the difficulty of addressing occlusions in tasks that focus on generalizability. Besides, the introduction of the first open-set recognition skeleton-based human action recognition benchmark opens the vistas for open-set model confidence calibration on skeleton-based human action recognition and highlights the limitations of existing methods developed for video/image data when employed on skeleton data. This study spurs the development of more effective solutions like the CrossMax approach.

Moreover, exploration on the generalizable challenges on video-based human action recognition are explored in this thesis, which focuses specifically on different types of domain adaptation. The construction of a large-scale few-shot domain adaptation benchmark and the proposed RelaMiX method demonstrate the efficiency and potential of few-shot domain adaptation over unsupervised domain adaptation, promoting its application in diverse real-world scenarios, especially when the target domain sample is hard to collect. Additionally, the cross-modal fall detection approach using unlabelled depth domain data offers valuable advancements for assistive technology, particularly in elderly care, where privacy supporting is likely to be considered by the users.

Overall, this thesis provides novel solutions and benchmarks on various unexplored generalizable challenges in the field of the human action recognition field, which will significantly benefit the research community and practical applications and pave the way to accomplish more reliable and generalizable deep learning.

6.2 New Generalizable Benchmarks

In this thesis, several benchmarks are constructed to address various generalization challenges in human action recognition field. First, we benchmark one-shot skeleton-based human action recognition task and the self-supervised skeleton-based human action recognition task under diverse occlusions, where state-of-the art one-shot skeleton-based action recognition approaches and self-supervised skeleton-based human action recognition approaches are involved. This benchmark is delivered to demonstrate the existing approaches designed for each specific generalizable challenge under the perturbation of different kinds of occlusion.

Moreover, an open-set skeleton-based human action recognition is constructed, which includes three datasets, different open-set splits, different validation settings, and incorporates three skeleton-based human action recognition backbones. This benchmark is designed to evaluate the performance of existing open-set recognition methods, which often struggle with the sparsity of skeleton data and lack of visual background information.

Additionally, the thesis introduces a benchmark for few-shot domain adaptation in video-based human action recognition. This benchmark compares the performance of few-shot domain adaptation with domain adaptation settings across various domains, highlighting the efficiency of few-shot domain adaptation in scenarios where collecting large-scale unlabeled target domain data is challenging.

Finally, this thesis includes a small study about cross-modal unsupervised RGB2Depth domain adaptation for fall detection. These comprehensive benchmarks provide foundations for future research on generalizable deep learning, enabling the evaluation and development of more reliable action recognition methods across various challenges.

6.3 Novel Methods for the Generalizable Challenges

This thesis introduces several innovative methods to address the generalization challenges in human action recognition. For one-shot skeleton-based action recognition under occlusions, the Trans4SOAR method is proposed. It utilizes a transformer architecture to achieve multi-modal feature fusion at the patch embedding level, combining human body joints, bones, and velocities derived from 3D motion data. By incorporating a prototype-based latent space consistency loss, Trans4SOAR enhances the generalizability and robustness of learned embeddings and demonstrates superior performance in both occluded and non-occluded scenarios. To address self-supervised skeleton-based action recognition under occlusions, the OPSTL method is developed. This two-stage imputation method can be integrated into various self-supervised learning pipelines. It employs three-stream

contrastive learning with adaptive spatial masking for data augmentation during training. After pre-training, KMeans clustering finds cluster centers, and the K nearest neighbors within each cluster are used for imputation. This approach is validated across different camera settings, datasets, and occlusion scenarios. For open-set skeleton-based human action recognition, the CrossMax method is introduced. CrossMax relies on cross-modal mean max discrepancy training across three branches for different modalities: joints, bones, and velocities. During the test phase, a channel normalized distance-based logits calibration approach combines the advantages of both SoftMax open-set probability scores and channel normalized distance-based open-set probability scores. This method delivers significant performance improvements across various datasets and GCN backbones.

In the field of video-based human action recognition, the RelaMiX method is introduced for few-shot domain adaptation. RelaMiX includes temporal relational dropout, snippet-wise and scale-wise attentional fusion for temporal aggregation, and source domain statistics-based feature mixture. It uses cross-domain information alignment loss to enhance the representation of the source and target domains for the same category. RelaMiX shows promising performances on leveraged datasets and serves as a significant baseline in this field. Additionally, a cross-modal fall detection approach is proposed to achieve test-time depth-based fall detection using unlabelled depth domain data during training. This method relies on domain agnostic adversarial learning and cross-batch triplet margin loss to learn discriminative embeddings. An intermediate domain module bridges the latent spaces from different modalities, proving effective on various datasets and backbones. These methods represent significant advancements in addressing generalization challenges in human action recognition, providing more reliable solutions that enhance the generalizability and effectiveness of action recognition deep learning methods in diverse conditions and scenarios.

6.4 Open Questions to Future Works

Building on the advancements in this thesis, several future directions can enhance human action recognition models. One promising area is developing more sophisticated methods for handling occlusions in real-time applications. This could involve exploring advanced imputation methodologies using deep cluster approaches instead of KMeans and integrating them with robust machine learning models to improve performance in dynamic and cluttered environments. Expanding the scope of open-set recognition in skeleton-based action recognition is another potential direction. Future research could focus on open-set human action localization, which is more challenging as it requires predicting the spatio-temporal localization of the person alongside the actions. In video-based human action recognition, future work could explore few-shot semi-supervised learning techniques across various domains and tasks. Developing methods to efficiently transfer knowledge from syn-

thetic to real-world data would significantly enhance system generalizability. Additionally, integrating unsupervised domain adaptation techniques with few-shot learning approaches could yield more flexible and adaptive models. Improving the interpretability and transparency of human action recognition models is also important. As these systems are deployed in critical applications like healthcare and surveillance, understanding their decision-making processes becomes essential. This could involve developing techniques for visualizing learned features and representations or creating frameworks that provide intuitive explanations of model outputs. Overall, these future directions have the potential to significantly advance human action recognition, making systems more robust, generalizable, and ethically sound while expanding their applicability to a broader range of real-world scenarios.

7 | Publication List

This doctoral research contains the following publications or submissions under review, where * indicates corresponding author and + indicates shared first author (supervised students as first author).

1. Kunyu Peng, Cheng Yin, Junwei Zheng, Ruiping Liu, David Schneider, Jiaming Zhang, Kailun Yang, M. Saquib Sarfraz, Rainer Stiefelhagen, Alina Roitberg: Navigating Open Set Scenarios for Skeleton-Based Action Recognition. AAAI Conference on Artificial Intelligence 2024: 4487-4496

2. Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, Rainer Stiefelhagen: Delving Deep Into One-Shot Skeleton-Based Action Recognition With Diverse Occlusions. IEEE Transactions on Multimedia, 25: 1489-1504 (2023)

3. Kunyu Peng, Di Wen, David Schneider, Jiaming Zhang, Kailun Yang, M. Saquib Sarfraz, Rainer Stiefelhagen, Alina Roitberg: Exploring Few-shot Domain Adaptation for Video-based Activity Recognition. AAAI Conference on Artificial Intelligence 2025 (under review)

4. Yifei Chen, Kunyu Peng*, Alina Roitberg, David Schneider, Jiaming Zhang, Junwei Zheng, Ruiping Liu, Yufan Chen, Kailun Yang, Rainer Stiefelhagen: Unveiling the Hidden Realm: Self-supervised Skeleton-based Action Recognition in Occluded Environments. IEEE International Conference on Acoustics, Speech, and Signal Processing 2025 (under review soon)

5. Hejun Xiao, Kunyu Peng+, Xiangsheng Huang, Alina Roitberg, Hao Li, Zhaohui Wang, Rainer Stiefelhagen: Toward Privacy-Supporting Fall Detection via Deep Unsupervised RGB2Depth Adaptation. IEEE Sensors Journal, 23: 29143-29155 (2023)

The publications and submissions under review list which are unrelated to this thesis is as follows,

1. Kunyu Peng, Fu Jia, Kailun Yang, Di Wen, Yufan Chen, Ruiping Liu, Junwei Zheng, Jiaming Zhang, M. Saquib Sarfraz, Rainer Stiefelhagen and Alina Roitberg. Referring Atomic Video Action Recognition. European Conference on Computer Vision 2024.

2. Kunyu Peng, David Schneider, Alina Roitberg, Kailun Yang, Jiaming Zhang, M. Saquib Sarfraz, Rainer Stiefelhagen: Towards Video-based Activated Muscle Group Estimation. ACM Multimedia 2024

3. Yi Xu, Kunyu Peng*, Di Wen, Ruiping Liu, Junwei Zheng, Yufan Chen, Jiaming Zhang, Alina Roitberg, Kailun Yang, Rainer Stiefelhagen: Skeleton-Based Human Action Recognition with Noisy Labels. IEEE/RSJ International Conference on Intelligent Robots and Systems 2024

4. Yiping Wei, Kunyu Peng*, Alina Roitberg, Jiaming Zhang, Junwei Zheng, Ruiping Liu, Yufan Chen, Kailun Yang, Rainer Stiefelhagen: Elevating Skeleton-Based Action Recognition with Efficient Multi-Modality Self-Supervision. IEEE International Conference on Acoustics, Speech, and Signal Processing 2024

5. Ping-Cheng Wei, Kunyu Peng+, Alina Roitberg, Kailun Yang, Jiaming Zhang, Rainer Stiefelhagen: Multi-modal Depression Estimation Based on Sub-attentional Fusion. European Conference on Computer VisionWorkshops (6) 2022: 623-639

6. Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, Rainer Stiefelhagen: TransDARC: Transformer-based Driver Activity Recognition with Latent Space Feature Calibration. IEEE/RSJ International Conference on Intelligent Robots and Systems 2022: 278-285

7. Kunyu Peng, Alina Roitberg, David Schneider, Marios Koulakis, Kailun Yang, Rainer Stiefelhagen: Affect-DML: Context-Aware One-Shot Recognition of Human Affect using Deep Metric Learning.IEEE International Conference on Automatic Face and Gesture Recognition 2021: 1-8

8. Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, Rainer Stiefelhagen: Should I take a walk? Estimating Energy Expenditure from Video Data. The IEEE / CVF Computer Vision and Pattern Recognition Conference Workshops 2022: 2074-2084

9. Alina Roitberg, Kunyu Peng, David Schneider, Kailun Yang, Marios Koulakis, Manuel Martínez, Rainer Stiefelhagen: Is My Driver Observation Model Overconfident? Input-Guided Calibration Networks for Reliable and Interpretable Confidence Estimates. IEEE Trans. Intell. Transp. Syst. 23(12): 25271-25286 (2022)

10. Alina Roitberg, Kunyu Peng, Zdravko Marinov, Constantin Seibold, David Schneider, Rainer Stiefelhagen: A Comparative Analysis of Decision-Level Fusion for Multimodal Driver Behaviour Understanding. IEEE Intelligent Vehicles Symposium 2022: 1438-1444

11. Calvin Tanama, Kunyu Peng, Zdravko Marinov, Rainer Stiefelhagen, Alina Roitberg: Quantized Distillation: Optimizing Driver Activity Recognition Models for Resource-Constrained Environments. IEEE/RSJ International Conference on Intelligent Robots and Systems 2023

12. Kunyu Peng, Juncong Fei, Kailun Yang, Alina Roitberg, Jiaming Zhang, Frank Bieder, Philipp Heidenreich, Christoph Stiller, Rainer Stiefelhagen: MASS: Multi-Attentional Semantic Segmentation of LiDAR Data for Dense Top-View Understanding. IEEE Trans. Intell. Transp. Syst. 23(9): 15824-15840 (2022)

13. M. Saquib Sarfraz, Mei-Yen Chen, Lukas Layer, Kunyu Peng, Marios Koulakis: Position: Quo Vadis, Unsupervised Time Series Anomaly Detection? International Conference on Machine Learning 2024

14. Yufan Chen, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ruiping Liu, Philip Torr, Rainer Stiefelhagen: RoDLA: Benchmarking the Robustness of Document Layout Analysis Models. The IEEE / CVF Computer Vision and Pattern Recognition Conference (2024)

15. Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, Rainer Stiefelhagen: Delivering Arbitrary-Modal Semantic Segmentation. The IEEE / CVF Computer Vision and Pattern Recognition Conference 2023: 1136-1147

16. Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, Rainer Stiefelhagen: Bending Reality: Distortion-aware Transformers for Adapting to Panoramic Semantic Segmentation. The IEEE / CVF Computer Vision and Pattern Recognition Conference 2022: 16896-16906

17. Ruiping Liu, Jiaming Zhang, Kunyu Peng, Yufan Chen, Ke Cao, Junwei Zheng, M. Saquib Sarfraz, Kailun Yang, Rainer Stiefelhagen: Fourier Prompt Tuning for Modality-Incomplete Scene Segmentation. IEEE Intelligent Vehicles Symposium 2024

18. Hao Shi, Yu Li, Kailun Yang, Jiaming Zhang, Kunyu Peng, Alina Roitberg, Yaozu Ye, Huajian Ni, Kaiwei Wang, Rainer Stiefelhagen: FishDreamer: Towards Fisheye Semantic Completion via Unified Image Outpainting and Segmentation. The IEEE / CVF Computer Vision and Pattern Recognition Conference Workshops 2023: 6434-6444

19. Ruiping Liu, Jiaming Zhang, Kunyu Peng, Junwei Zheng, Ke Cao, Yufan Chen, Kailun Yang, Rainer Stiefelhagen: Open Scene Understanding: Grounded Situation Recognition Meets Segment Anything for Helping People with Visual Impairments. IEEE/CVF International Conference on Computer Vision (Workshops) 2023: 1849-1859

20. Chang Chen, Jiaming Zhang, Kailun Yang, Kunyu Peng, Rainer Stiefelhagen: Trans4Map: Revisiting Holistic Bird's-Eye-View Mapping from Egocentric Images to Allocentric Semantics with Vision Transformers. IEEE/CVF Winter Conference on Applications of Computer Vision 2023: 4002-4011

21. Junwei Zheng, Jiaming Zhang, Kailun Yang, Kunyu Peng, Rainer Stiefelhagen: MateRobot: Material Recognition in Wearable Robotics for People with Visual Impairments. IEEE International Conference on Robotics and Automation 2024

22. Zhifeng Teng, Jiaming Zhang, Kailun Yang, Kunyu Peng, Hao Shi, Simon Reiß, Ke Cao, Rainer Stiefelhagen: 360BEV: Panoramic Semantic Mapping for Indoor Bird's-Eye View. IEEE/CVF Winter Conference on Applications of Computer Vision 2024: 372-381

23. Wenyan Ou, Jiaming Zhang, Kunyu Peng, Kailun Yang, Gerhard Jaworek, Karin Müller, Rainer Stiefelhagen: Indoor Navigation Assistance for Visually Impaired People via Dynamic SLAM and Panoptic Segmentation with an RGB-D Sensor. ICCHP-AAATE (1) 2022: 160-168

24. Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, Rainer Stiefelhagen: MatchFormer: Interleaving Attention in Transformers for Feature Matching. IEEE Asian Conference on Computer Vision. 2022: 2746-2762.

25. Jiaming Zhang, Chaoxiang Ma, Kailun Yang, Alina Roitberg, Kunyu Peng, Rainer Stiefelhagen: Transfer Beyond the Field of View: Dense Panoramic Semantic Segmentation via Unsupervised Domain Adaptation. IEEE Trans. Intell. Transp. Syst. 23(7): 9478-9491 (2022)

26. Xinyu Luo, Jiaming Zhang, Kailun Yang, Alina Roitberg, Kunyu Peng, Rainer Stiefelhagen: Towards Robust Semantic Segmentation of Accident Scenes via Multi-Source Mixed Sampling and Meta-Learning. The IEEE / CVF Computer Vision and Pattern Recognition Conference Workshops 2022: 4428-4438

Bibliography

[1] Bruno Aguiar, Tiago Rocha, Joana Silva, and Ines Sousa. "Accelerometer-based fall detection for smartphones". In: *IEEE International Symposium on Medical Measurements and Applications*. 2014, pp. 1–6.

[2] Dasom Ahn, Sangwon Kim, Hyunsu Hong, and Byoung Chul Ko. "STAR-Transformer: A spatio-temporal cross attention transformer for human action recognition". In: *Proc. WACV*. 2023, pp. 3330–3339.

[3] Shun-ichi Amari. "Backpropagation and stochastic gradient descent method". In: *Neurocomputing* 5.4-5 (1993), pp. 185–196.

[4] Federico Angelini, Zeyu Fu, Yang Long, Ling Shao, and Syed Mohsen Naqvi. "2D Pose-Based Real-Time Human Action Recognition With Occlusion-Handling". In: *IEEE Transactions on Multimedia* (2020).

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. "Vivit: A video vision transformer". In: *IEEE International Conference on Computer Vision*. 2021, pp. 6836–6846.

[6] Umar Asif, Benjamin Mashford, Stefan Von Cavallar, Shivanthan Yohanandan, Subhrajit Roy, Jianbin Tang, and Stefan Harrer. "Privacy preserving human fall detection using video data". In: *Machine Learning for Health Workshop*. PMLR. 2020, pp. 39–51.

[7] Woon-Sung Baek, Dong-Min Kim, Faisal Bashir, and Jae-Young Pyun. "Real life applicable fall detection system based on wireless body area network". In: *IEEE Consumer Communications and Networking Conference*. 2013, pp. 62–67.

[8] Ruwen Bai. "Hierarchical Graph Convolutional Skeleton Transformer for Action Recognition". In: *arXiv preprint arXiv:2109.02860* (2021).

[9] Chaitanya Bandi and Ulrike Thomas. "Skeleton-based action recognition for human-robot interaction using self-attention mechanism". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. 2021, pp. 1–8.

[10] Wentao Bao, Qi Yu, and Yu Kong. "Evidential deep learning for open set action recognition". In: *IEEE International Conference on Computer Vision*. 2021, pp. 13349–13358.

[11] Yassir Bendou et al. "EASY: Ensemble Augmented-Shot Y-shaped Learning: State-Of-The-Art Few-Shot Classification with Simple Ingredients". In: *arXiv preprint arXiv:2201.09699* (2022).

[12] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. "Is space-time attention all you need for video understanding?" In: *International Conference on Machine Learning*. Vol. 2. 3. 2021, p. 4.

[13] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. "Memory matching networks for one-shot image recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4080–4088.

[14] Xi Cai, Suyuan Li, Xinyue Liu, and Guang Han. "Vision-based fall detection with multi-task hourglass convolutional auto-encoder". In: *IEEE Access* 8 (2020), pp. 44493–44502.

[15] Congqi Cao, Yajuan Li, Qinyi Lv, Peng Wang, and Yanning Zhang. "Few-shot action recognition with implicit temporal alignment and pair similarity optimization". In: *Computer Vision and Image Understanding* (2021).

[16] Huiqiang Cao, Shuicai Wu, Zhuhuang Zhou, Chung-Chih Lin, Chih-Yu Yang, Shih-Tseng Lee, and Chieh-Tsai Wu. "A fall detection method based on acceleration data and hidden Markov model". In: *IEEE International Conference on Signal and Image Processing*. 2016, pp. 684–689.

[17] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1302–1310.

[18] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. "A short note on the kinetics-700 human action dataset". In: *arXiv preprint arXiv:1907.06987* (2019).

[19] Joao Carreira and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6299–6308.

[20] Eduardo Casilari, Jose A Santoyo-Ramón, and Jose M Cano-García. "Umafall: A multisensor dataset for the research on automatic fall detection". In: *Procedia Computer Science* 110 (2017), pp. 32–39.

[21] Jun Cen, Di Luan, Shiwei Zhang, Yixuan Pei, Yingya Zhang, Deli Zhao, Shaojie Shen, and Qifeng Chen. "The devil is in the wrongly-classified samples: Towards unified open-set recognition". In: *arXiv preprint arXiv:2302.04002* (2023).

[22] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. "Domain-specific batch normalization for unsupervised domain adaptation". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7354–7362.

[23] Diansheng Chen, Wei Feng, Yu Zhang, Xiyu Li, and Tianmiao Wang. "A wearable wireless fall detection system with accelerators". In: *IEEE International Conference on Robotics and Biomimetics*. IEEE. 2011, pp. 2259–2263.

[24] Guangyao Chen, Peixi Peng, Xiangqian Wang, and Yonghong Tian. "Adversarial reciprocal points learning for open set recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11 (2021), pp. 8065–8081.

[25] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. "Learning open set network with discriminative reciprocal points". In: *European Conference on Computer Vision*. Springer. 2020.

[26] J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajcsy. "Wearable Sensors for Reliable Fall Detection". In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. 2005, pp. 3551–3554. DOI: 10.1109/IEMBS.2005.1617246.

[27] Lin Chen, Rong Li, Hang Zhang, Lili Tian, and Ning Chen. "Intelligent fall detection method based on accelerometer data from a wrist-worn smart watch". In: *Measurement* 140 (2019), pp. 215–226.

[28] 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*. 2019, pp. 6321–6330.

[29] Xiong-Hui Chen, Shengyi Jiang, Feng Xu, Zongzhang Zhang, and Yang Yu. "Cross-modal domain adaptation for cost-efficient visual reinforcement learning". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12520–12532.

[30] Yatong Chen, Hongwei Ge, Yuxuan Liu, Xinye Cai, and Liang Sun. "Agpn: Action granularity pyramid network for video action recognition". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[31] Yifei Chen, Kunyu Peng, Alina Roitberg, David Schneider, Jiaming Zhang, Junwei Zheng, Ruiping Liu, Yufan Chen, Kailun Yang, and Rainer Stiefelhagen. "Unveiling the Hidden Realm: Self-supervised Skeleton-based Action Recognition in Occluded Environments". In: *IEEE International Conference on Intelligent Robots and Systems (under review)*. 2024.

[32] Yong Chen, Weitong Li, Lu Wang, Jiajia Hu, and Mingbin Ye. “Vision-based fall event detection in complex background using attention guided bi-directional LSTM”. In: *IEEE Access* 8 (2020), pp. 161337–161348.

[33] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. “Channel-wise topology refinement graph convolution for skeleton-based action recognition”. In: *IEEE International Conference on Computer Vision*. 2021, pp. 13359–13368.

[34] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. “Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition”. In: *AAAI Conference on Artificial Intelligence*. Vol. 35. 2. 2021, pp. 1113–1122.

[35] Zhenjie Chen, Hongsong Wang, and Jie Gui. “Occluded Skeleton-Based Human Action Recognition with Dual Inhibition Training”. In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 2625–2634.

[36] Yi-Bin Cheng, Xipeng Chen, Dongyu Zhang, and Liang Lin. “Motion-transformer: Self-supervised pre-training for skeleton-based action recognition”. In: *ACM International Conference on Multimedia in Asia*. 2021, pp. 1–6.

[37] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. “Shuffle and attend: Video domain adaptation”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 678–695.

[38] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. “Twins: Revisiting the design of spatial attention in vision transformers”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 9355–9366.

[39] Xin Cong, Bowen Yu, Tingwen Liu, Shiyao Cui, Hengzhu Tang, and Bin Wang. “Inductive unsupervised domain adaptation for few-shot classification via clustering”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*. Springer. 2021, pp. 624–639.

[40] MMAction2 Contributors. *OpenMMLab’s Next Generation Video Understanding Toolbox and Benchmark*. https://github.com/open-mmlab/mmaction2. 2020.

[41] 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: *IEEE Winter Conference on Applications of Computer Vision*. 2022, pp. 1181–1190.

[42] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. “Mixformer: End-to-end tracking with iterative mixed attention”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13608–13618.

[43] 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 (2022), pp. 2533–2550.

[44] Yongxing Dai, Jun Liu, Yifan Sun, Zekun Tong, Chi Zhang, and Ling-Yu Duan. "Idm: An intermediate domain module for domain adaptive person re-id". In: *IEEE International Conference on Computer Vision*. 2021, pp. 11864–11874.

[45] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. "Scaling egocentric vision: The epic-kitchens dataset". In: *European Conference on Computer Vision*. Springer. 2018, pp. 720–736.

[46] Somayeh Danafar and Niloofar Gheissari. "Action recognition for surveillance applications using optic flow and SVM". In: *Computer Vision–ACCV 2007: 8th Asian Conference on Computer Vision, Tokyo, Japan, November 18-22, 2007, Proceedings, Part II 8*. Springer. 2007, pp. 457–466.

[47] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. "Toyota smarthome: Real-world activities of daily living". In: *IEEE International Conference on Computer Vision*. 2019, pp. 833–842.

[48] Jianyang Ding and Yong Wang. "A WiFi-based smart home fall detection system using recurrent neural network". In: *IEEE Transactions on Consumer Electronics* 66.4 (2020), pp. 308–317.

[49] Kaize Ding, Albert Jiongqian Liang, Bryan Perozzi, Ting Chen, Ruoxi Wang, Lichan Hong, Ed H Chi, Huan Liu, and Derek Zhiyuan Cheng. "HyperFormer: Learning expressive sparse feature representations via hypergraph transformer". In: *SIGIR*. 2023.

[50] Xiaolu Ding, Shuqiong Zhu, Wei Qu, and Wai Chen. "Generalized graph convolutional networks for action recognition with occluded skeletons". In: *International Conference on Computing and Pattern Recognition*. 2020, pp. 43–49.

[51] John K Dixon. "Pattern recognition with partly missing data". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.10 (1979), pp. 617–621.

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[53] Dawei Du, Ameya Shringi, Anthony Hoogs, and Christopher Funk. “Reconstructing humpty dumpty: Multi-feature graph autoencoder for open set action recognition”. In: *IEEE Winter Conference on Applications of Computer Vision*. 2023, pp. 3371–3380.

[54] Yong Du, Wei Wang, and Liang Wang. “Hierarchical recurrent neural network for skeleton based action recognition”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1110–1118.

[55] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. “Revisiting skeleton-based action recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2022, pp. 2969–2978.

[56] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. “How well do self-supervised models transfer?” In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5414–5423.

[57] Ricardo Espinosa, Hiram Ponce, Sebastián Gutiérrez, Lourdes Martínez-Villaseñor, Jorge Brieva, and Ernesto Moya-Albor. “A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-Fall detection dataset”. In: *Computers in biology and medicine* 115 (2019), p. 103520.

[58] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. “Multiscale vision transformers”. In: *IEEE International Conference on Computer Vision*. 2021, pp. 6824–6835.

[59] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. “Cian: Cross-image affinity net for weakly supervised semantic segmentation”. In: *AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 10762–10769.

[60] 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*. 2017, pp. 2334–2343.

[61] Christoph Feichtenhofer. “X3d: Expanding architectures for efficient video recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 203–213.

[62] Dario Fontanel, Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. “Boosting deep open world recognition by clustering”. In: *IEEE Robotics and Automation Letters* 5.4 (2020), pp. 5985–5992.

[63] Yaroslav Ganin and Victor Lempitsky. “Unsupervised domain adaptation by backpropagation”. In: *International conference on machine learning*. PMLR. 2015, pp. 1180–1189.

[64] Zan Gao, Leming Guo, Weili Guan, An-An Liu, Tongwei Ren, and Shengyong Chen. "A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-R2". In: *IEEE Transactions on Image Processing* 30 (2020), pp. 767–782.

[65] Chuanxing Geng and Songcan Chen. "Collective decision for open set recognition". In: *IEEE Transactions on Knowledge and Data Engineering* 34.1 (2020), pp. 192–204.

[66] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. "Recent Advances in Open Set Recognition: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.10 (2021), pp. 3614–3631.

[67] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. "Recent advances in open set recognition: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 43.10 (2020), pp. 3614–3631.

[68] Mehwish Ghafoor and Arif Mahmood. "Quantification of Occlusion Handling Capability of a 3D Human Pose Estimation Framework". In: *IEEE Transactions on Multimedia* 25 (2023), pp. 3311–3318.

[69] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. "Levit: a vision transformer in convnet's clothing for faster inference". In: *IEEE International Conference on Computer Vision*. 2021, pp. 12259–12269.

[70] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. "A kernel two-sample test". In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[71] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition". In: *AAAI Conference on Artificial Intelligence*. Vol. 36. 1. 2022, pp. 762–770.

[72] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. "Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition". In: *AAAI Conference on Artificial Intelligence*. Vol. 36. 1. 2022, pp. 762–770.

[73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[74] Dan Hendrycks and Kevin Gimpel. "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". In: *International Conference on Learning Representations*. 2017.

[75] Alexander Hermans, Lucas Beyer, and Bastian Leibe. "In defense of the triplet loss for person re-identification". In: *arXiv preprint arXiv:1703.07737* (2017).

[76] James Hong, Matthew Fisher, Michaël Gharbi, and Kayvon Fatahalian. "Video pose distillation for few-shot, fine-grained sports action recognition". In: *IEEE International Conference on Computer Vision*. 2021, pp. 9254–9263.

[77] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. "Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9068–9077.

[78] Yuqian Hu, Feng Zhang, Chenshu Wu, Beibei Wang, and KJ Ray Liu. "A WiFi-based passive fall detection system". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2020, pp. 1723–1727.

[79] Yuqing Hu, Stéphane Pateux, and Vincent Gripon. "Squeezing backbone feature distributions to the max for efficient few-shot learning". In: *Algorithms* 15.5 (2022), p. 147.

[80] Guoliang Hua, Hong Liu, Wenhao Li, Qian Zhang, Runwei Ding, and Xin Xu. "Weakly-Supervised 3D Human Pose Estimation With Cross-View U-shaped Graph Convolutional Network". In: *IEEE Transactions on Multimedia* 25 (2023), pp. 1832–1843.

[81] Shengqi Huang, Wanqi Yang, Lei Wang, Luping Zhou, and Ming Yang. "Few-shot unsupervised domain adaptation with image-to-class sparse similarity encoding". In: *ACM International Conference on Multimedia*. 2021, pp. 677–685.

[82] Forrest N. Iandola et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size". In: *arXiv preprint arXiv:1602.07360* (2016).

[83] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. "A survey on contrastive self-supervised learning". In: *Technologies* 9.1 (2020), p. 2.

[84] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. "xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 12605–12614.

[85] Taotao Jing, Haifeng Xia, Jihun Hamm, and Zhengming Ding. "Marginalized augmented few-shot domain adaptation". In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[86] Gregory Kahn, Adam Villaflor, Bosen Ding, Pieter Abbeel, and Sergey Levine. "Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation". In: *International Conference on Robotics and Automation*. 2018, pp. 5129–5136.

[87] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. "Contrastive adaptation network for unsupervised domain adaptation". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4893–4902.

[88] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k-means clustering algorithm: analysis and implementation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.7 (2002), pp. 881–892.

[89] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[90] Chadia Khraief, Faouzi Benzarti, and Hamid Amiri. "Elderly fall detection based on multistream deep convolutional networks". In: *Multimedia Tools and Applications* 79 (2020), pp. 19537–19560.

[91] Miran Kim, Xiaoqian Jiang, Kristin Lauter, Elkhan Ismayilzada, and Shayan Shams. "Secure human action recognition by encrypted neural network inference". In: *Nature communications* 13.1 (2022), p. 4799.

[92] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations*. Ed. by Yoshua Bengio and Yann LeCun. 2015.

[93] Konstantina N Kottari, Konstantinos K Delibasis, and Ilias G Maglogiannis. "Real-time fall detection using uncalibrated fisheye cameras". In: *IEEE Transactions on Cognitive and Developmental Systems* 12.3 (2019), pp. 588–600.

[94] Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. "BAR: Bayesian activity recognition using variational inference". In: *arXiv preprint arXiv:1811.03305* (2018).

[95] Alex Krizhevsky. "One weird trick for parallelizing convolutional neural networks". In: *arXiv preprint arXiv:1404.5997* (2014).

[96] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. "HMDB: a large video database for human motion recognition". In: *IEEE International Conference on Computer Vision*. IEEE. 2011, pp. 2556–2563.

[97] Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. "Featmatch: Feature-based augmentation for semi-supervised learning". In: *European Conference on Computer Vision*. Springer. 2020, pp. 479–495.

[98] 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 (2014), pp. 489–501.

[99] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. "Sliced wasserstein discrepancy for unsupervised domain adaptation". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10285–10295.

[100] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoon Lee. "Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition". In: *arXiv preprint arXiv:2208.10741* (2022).

[101] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition". In: *IEEE International Conference on Computer Vision*. 2023, pp. 10410–10419.

[102] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. "Skeleton-based action recognition with convolutional neural networks". In: *ICMEW*. 2017.

[103] Dan Li and Wuzhen Shi. "Partially occluded skeleton action recognition based on multi-stream fusion graph convolutional networks". In: *Computer Graphics International Conference*. Springer. 2021, pp. 178–189.

[104] Jinfeng Li, Weifeng Liu, Yicong Zhou, Jun Yu, Dapeng Tao, and Changsheng Xu. "Domain-invariant graph for adaptive semi-supervised domain adaptation". In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 18.3 (2022), pp. 1–18.

[105] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. "Unsupervised learning of view-invariant action representations". In: *Advances in Neural Information Processing Systems* 31 (2018).

[106] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. "3D Human Action Representation Learning via Cross-View Consistency Pursuit". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 4741–4750.

[107] Mengxue Li, Yi-Ming Zhai, You-Wei Luo, Peng-Fei Ge, and Chuan-Xian Ren. "Enhanced transport distance for unsupervised domain adaptation". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 13936–13944.

[108] Xinyu Li, Yuan He, J. Andrew Zhang, and Xiaojun Jing. "Supervised domain adaptation for few-shot radar-based human activity recognition". In: *IEEE Sensors Journal* (2021).

[109] Xuanfeng Li, Jian Lu, Xiaogai Chen, and Xiaodan Zhang. “Spatial-Temporal Adaptive Metric Learning Network for One-Shot Skeleton-Based Action Recognition”. In: *IEEE Signal Processing Letters* 31 (2024), pp. 321–325.

[110] Zhidong Liang, Ming Yang, Liuyuan Deng, Chunxiang Wang, and Bing Wang. “Hierarchical depthwise graph convolutional neural network for 3D semantic segmentation of point clouds”. In: *International Conference on Robotics and Automation*. IEEE. 2019, pp. 8152–8158.

[111] Dongha Lim, Chulho Park, Nam Ho Kim, Sang-Hoon Kim, and Yun Seop Yu. “Fall-detection algorithm using 3-axis acceleration: combination with simple threshold and hidden Markov model”. In: *Journal of Applied Mathematics* 2014 (2014).

[112] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. “MS2L: Multi-task Self-Supervised Learning for Skeleton Based Action Recognition”. In: *ACM International Conference on Multimedia (MM)*. 2020, pp. 2490–2498.

[113] Hong Liu and Jianping Xie. “Comparative genomics of Mycobacterium tuberculosis drug efflux pumps and their transcriptional regulators”. In: *Critical Reviews™ in Eukaryotic Gene Expression* 24.2 (2014).

[114] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. “NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.10 (2020), pp. 2684–2701.

[115] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. “NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

[116] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. “NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.10 (2020), pp. 2684–2701.

[117] Jun Liu, Amir Shahroudy, Dong Xu, Alex C Kot, and Gang Wang. “Skeleton-based action recognition using spatio-temporal LSTM network with trust gates”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.12 (2017), pp. 3007–3021.

[118] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. “Global context-aware attention lstm networks for 3d action recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 1647–1656.

[119] Mengyuan Liu, Hong Liu, and Chen Chen. “Robust 3D action recognition through sampling local appearances and global distributions”. In: *IEEE Transactions on Multimedia* 20.8 (2017), pp. 1932–1947.

[120] Xiao Liu et al. “Self-supervised learning: Generative or contrastive”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.1 (2023), pp. 857–876.

[121] Yanan Liu, Hao Zhang, Yanqiu Li, Kangjian He, and Dan Xu. “Skeleton-based Human Action Recognition via Large-kernel Attention Graph Convolutional Network”. In: *IEEE Transactions on Visualization and Computer Graphics* 29 (2023), pp. 2575–2585.

[122] 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: *IEEE International Conference on Computer Vision*. 2021, pp. 10012–10022.

[123] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. “Video swin transformer”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3202–3211.

[124] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. “Learning transferable features with deep adaptation networks”. In: *International conference on machine learning*. PMLR. 2015, pp. 97–105.

[125] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. “Deep transfer learning with joint adaptation networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 2208–2217.

[126] Ilya Loshchilov and Frank Hutter. “Decoupled Weight Decay Regularization”. In: *7th International Conference on Learning Representations*. 2019.

[127] Jing Lu, Yunlu Xu, Hao Li, Zhanzhan Cheng, and Yi Niu. “Pmal: Open set recognition via robust prototype mining”. In: *AAAI Conference on Artificial Intelligence*. Vol. 36. 2. 2022, pp. 1872–1880.

[128] Na Lu, Yidan Wu, Li Feng, and Jinbo Song. “Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data”. In: *IEEE Journal of Biomedical and Health Informatics* 23.1 (2018), pp. 314–323.

[129] Lei Ma, Yuhui Zheng, Zhao Zhang, Yazhou Yao, Xijian Fan, and Qiaolin Ye. “Motion stimulation for compositional action recognition”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[130] Laurens van der Maaten and Geoffrey E. Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605.

[131] 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 (2019), p. 1988.

[132] Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. "Action transformer: A self-attention model for short-time pose-based human action recognition". In: *Pattern Recognition* 124 (2022), p. 108487.

[133] Amir Mehmood, Adnan Nadeem, Muhammad Ashraf, Turki Alghamdi, and Muhammad Shoaib Siddiqui. "A novel fall detection algorithm for elderly using SHIMMER wearable sensors". In: *Health and Technology* 9 (2019), pp. 631–646.

[134] Raphael Memmesheimer, Simon Häring, Nick Theisen, and Dietrich Paulus. "Skeleton-dml: Deep metric learning for skeleton-based one-shot action recognition". In: *IEEE Winter Conference on Applications of Computer Vision*. 2022, pp. 3702–3710.

[135] Raphael Memmesheimer, Nick Theisen, and Dietrich Paulus. "SL-DML: Signal Level Deep Metric Learning for Multimodal One-Shot Action Recognition". In: *2020 25th International Conference on Pattern Recognition (ICPR)* (2020), pp. 4573–4580.

[136] Benjamin J Meyer and Tom Drummond. "The importance of metric learning for robotic vision: Open set recognition and active learning". In: *International Conference on Robotics and Automation*. IEEE. 2019, pp. 2924–2931.

[137] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. "Dropout sampling for robust object detection in open-set conditions". In: *International Conference on Robotics and Automation*. 2018, pp. 3243–3249.

[138] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6707–6717.

[139] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. "Few-shot adversarial domain adaptation". In: *Advances in Neural Information Processing Systems* 30 (2017).

[140] Debadyuti Mukherjee, Riktim Mondal, Pawan Kumar Singh, Ram Sarkar, and Debotosh Bhattacharjee. "EnsemConvNet: a deep learning approach for human activity recognition using smartphone sensors for healthcare applications". In: *Multimedia Tools and Applications* 79 (2020), pp. 31663–31690.

[141] Jonathan Munro and Dima Damen. "Multi-modal domain adaptation for fine-grained action recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 122–132.

[142] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. "Attention bottlenecks for multimodal fusion". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14200–14213.

[143] Poojan Oza and Vishal M Patel. "C2ae: Class conditioned auto-encoder for open-set recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2307–2316.

[144] Jay Patravali, Gaurav Mittal, Ye Yu, Fuxin Li, and Mei Chen. "Unsupervised few-shot action recognition via action-appearance aligned meta-adaptation". In: *IEEE International Conference on Computer Vision*. 2021, pp. 8484–8494.

[145] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[146] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. "Delving deep into one-shot skeleton-based action recognition with diverse occlusions". In: *IEEE Transactions on Multimedia* (2023).

[147] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. "TransDARC: Transformer-based driver activity recognition with latent space feature calibration". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2022, pp. 278–285.

[148] Kunyu Peng, Di Wen, David Schneider, Jiaming Zhang, Kailun Yang, M. Saquib Sarfraz, Rainer Stiefelhagen, and Alina Roitberg. "Exploring Few-Shot Adaptation for Activity Recognition on Diverse Domains". In: *ACM Multimedia (under review)*. 2024.

[149] Kunyu Peng, Cheng Yin, Junwei Zheng, Ruiping Liu, David Schneider, Jiaming Zhang, Kailun Yang, M Saquib Sarfraz, Rainer Stiefelhagen, and Alina Roitberg. "Navigating open set scenarios for skeleton-based action recognition". In: *AAAI Conference on Artificial Intelligence*. Vol. 38. 5. 2024, pp. 4487–4496.

[150] Wei Peng, Xiaopeng Hong, Haoyu Chen, and Guoying Zhao. "Learning graph convolutional network for skeleton-based human action recognition by neural searching". In: *AAAI Conference on Artificial Intelligence*. Vol. 34. 03. 2020, pp. 2669–2676.

[151] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. "Temporal-relational crosstransformers for few-shot action recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 475–484.

[152] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. "Skeleton-based action recognition via spatial and temporal transformer networks". In: *Computer Vision and Image Understanding* 208 (2021), p. 103219.

[153] Chiara Plizzari, Marco Cannici, and Matteo Matteucci. "Spatial temporal transformer network for skeleton-based action recognition". In: *Pattern recognition. ICPR international workshops and challenges: virtual event, January 10–15, 2021, Proceedings, Part III*. Springer. 2021, pp. 694–701.

[154] Can Qin, Lichen Wang, Qianqian Ma, Yu Yin, Huan Wang, and Yun Fu. "Contradictory structure learning for semi-supervised domain adaptation". In: *SIAM International Conference on Data Mining*. SIAM. 2021, pp. 576–584.

[155] Alina Roitberg, Chaoxiang Ma, Monica Haurilet, and Rainer Stiefelhagen. "Open set driver activity recognition". In: *IEEE Intelligent Vehicles Symposium*. 2020, pp. 1048–1053.

[156] 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: *IEEE International Conference on Intelligent Robots and Systems*. 2021, pp. 8563–8569.

[157] Alberto Sabater, Laura Santos, Jose Santos-Victor, Alexandre Bernardino, Luis Montesano, and Ana C Murillo. "One-shot action recognition in challenging therapy scenarios". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 2777–2785.

[158] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. "Contrast and mix: Temporal contrastive video domain adaptation with background mixing". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 23386–23400.

[159] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. "Semi-supervised domain adaptation via minimax entropy". In: *IEEE International Conference on Computer Vision*. 2019, pp. 8050–8058.

[160] David Schneider, Saquib Sarfraz, Alina Roitberg, and Rainer Stiefelhagen. "Pose-based contrastive learning for domain agnostic activity representations". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2022, pp. 3433–3443.

[161] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. "NTU RGB+D: A large scale dataset for 3D human activity analysis". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 1010–1019.

[162] Muhammad Bilal Shaikh and Douglas Chai. "RGB-D data-based action recognition: A review". In: *Sensors* 21.12 (2021), p. 4246.

[163] Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. "Automatic operating room surgical activity recognition for robot-assisted surgery". In: *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*. 2020, pp. 385–395.

[164] Zhongwei Shen, Xiao-Jun Wu, and Tianyang Xu. "FEXNet: Foreground extraction network for human action recognition". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.5 (2021), pp. 3141–3151.

[165] Wuzhen Shi, Dan Li, Yang Wen, and Wu Yang. "Occlusion-aware graph neural networks for skeleton action recognition". In: *IEEE Transactions on Industrial Informatics* 19.10 (2023), pp. 10288–10298.

[166] Yujie Shi. "Open set action recognition based on skeleton". In: *IEEE International Conference on Computer and Communication Systems*. 2023, pp. 1062–1066.

[167] Xiangbo Shu, Jiawen Yang, Rui Yan, and Yan Song. "Expansion-squeeze-excitation fusion network for elderly activity recognition". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.8 (2022), pp. 5281–5292.

[168] Yu Shu, Yemin Shi, Yaowei Wang, Yixiong Zou, Qingsheng Yuan, and Yonghong Tian. "Odn: Opening the deep network for open-set action recognition". In: *International conference on multimedia and expo*. 2018, pp. 1–6.

[169] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan. "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 1227–1236.

[170] Karen Simonyan and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos". In: *Advances in Neural Information Processing Systems* 27 (2014).

[171] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. "Constructing stronger and faster baselines for skeleton-based action recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 45.2 (2022), pp. 1474–1488.

[172] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. "Richly activated graph convolutional network for robust skeleton-based action recognition". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.5 (2020), pp. 1915–1925.

[173] Yi-Fan Song, Zhang Zhang, and Liang Wang. "Richly activated graph convolutional network for action recognition with incomplete skeletons". In: *IEEE International Conference on Image Processing*. IEEE. 2019, pp. 1–5.

[174] Liangchen Song, Gang Yu, Junsong Yuan, and Zicheng Liu. "Human pose estimation and its application to action recognition: A survey". In: *Journal of Visual Communication and Image Representation* 76 (2021), p. 103055.

[175] Ziyang Song et al. "Attention-Oriented Action Recognition for Real-Time Human-Robot Interaction". In: *International Conference on Pattern Recognition*. 2020, pp. 7087–7094.

[176] Tae Soo Kim and Austin Reiter. "Interpretable 3d human action analysis with temporal convolutional networks". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 20–28.

[177] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. "UCF101: A dataset of 101 human actions classes from videos in the wild". In: *arXiv preprint arXiv:1212.0402* (2012).

[178] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. "Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference". In: *IEEE International Conference on Computer Vision*. 2019, pp. 6301–6310.

[179] Xin Sun, Zhenning Yang, Chi Zhang, Keck-Voon Ling, and Guohao Peng. "Conditional gaussian distribution learning for open set recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 13480–13489.

[180] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. "Pix3d: Dataset and methods for single-image 3d shape modeling". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 2974–2983.

[181] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9.

[182] Takeshi Teshima, Issei Sato, and Masashi Sugiyama. "Few-shot domain adaptation by causal mechanism transfer". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9458–9469.

[183] Jantima Thummala and Suree Pumrin. "Fall detection using motion history image and shape deformation". In: *IEEE International Electrical Engineering Congress*. 2020, pp. 1–4.

[184] Anyang Tong, Chao Tang, and Wenjian Wang. "Semi-supervised action recognition from temporal augmentation using curriculum learning". In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.3 (2022), pp. 1305–1319.

[185] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. "Going deeper with image transformers". In: *IEEE International Conference on Computer Vision*. 2021, pp. 32–42.

[186] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. "Learning spatiotemporal features with 3d convolutional networks". In: *IEEE International Conference on Computer Vision*. 2015, pp. 4489–4497.

[187] Satoshi Tsutsui, Yanwei Fu, and David Crandall. "Meta-reinforced synthetic data for one-shot fine-grained visual recognition". In: *Advances in Neural Information Processing Systems* 32 (2019).

[188] Juanhui Tu, Mengyuan Liu, and Hong Liu. “Skeleton-based human action recognition using spatial temporal 3D convolutional neural networks”. In: *IEEE International conference on multimedia and expo*. 2018, pp. 1–6.

[189] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. “Adversarial discriminative domain adaptation”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7167–7176.

[190] Viacheslav Voronin, Marina Zhdanova, Evgenii Semenishchev, Aleksander Zelenskii, Yigang Cen, and Sos Agaian. “Action recognition for the robotics and manufacturing automation using 3-D binary micro-block difference”. In: *The International Journal of Advanced Manufacturing Technology* 117 (2021), pp. 2319–2330.

[191] Hao Wang, Daqing Zhang, Yasha Wang, Junyi Ma, Yuxiang Wang, and Shengjie Li. “RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices”. In: *IEEE Transactions on Mobile Computing* 16.2 (2016), pp. 511–526.

[192] 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: *European Conference on Computer Vision*. Springer. 2016, pp. 20–36.

[193] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. “Action recognition based on joint trajectory maps using convolutional neural networks”. In: *ACM International Conference on Multimedia*. 2016, pp. 102–106.

[194] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. “Hybrid relation guided set matching for few-shot action recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2022, pp. 19948–19957.

[195] Xiao Wang, Weirong Ye, Zhongang Qi, Xun Zhao, Guangge Wang, Ying Shan, and Hanzi Wang. “Semantic-guided relation propagation network for few-shot action recognition”. In: *ACM International Conference on Multimedia*. 2021, pp. 816–825.

[196] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R Scott. “Cross-batch memory for embedding learning”. In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6388–6397.

[197] Yuxi Wang, Kaishun Wu, and Lionel M Ni. “Wifall: Device-free fall detection by wireless networks”. In: *IEEE Transactions on Mobile Computing* 16.2 (2016), pp. 581–594.

[198] Pengfei Wei, Lingdong Kong, Xinghua Qu, Yi Ren, Zhiqiang Xu, Jing Jiang, and Xiang Yin. “Unsupervised Video Domain Adaptation for Action Recognition: A Disentanglement Perspective”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[199] Hanbo Wu, Xin Ma, and Yibin Li. "Spatiotemporal multimodal learning with 3D CNNs for video action recognition". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.3 (2021), pp. 1250–1261.

[200] Hejun Xiao, Kunyu Peng, Xiangsheng Huang, Alina Roitberg, Hao Li, Zhaohui Wang, and Rainer Stiefelhagen. "Toward Privacy-Supporting Fall Detection via Deep Unsupervised RGB2Depth Adaptation". In: *IEEE Sensors Journal* 23.23 (2023), pp. 29143–29155.

[201] Ni Xiao and Lei Zhang. "Dynamic weighted learning for unsupervised domain adaptation". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15242–15251.

[202] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification". In: *European Conference on Computer Vision*. Springer. 2018, pp. 305–321.

[203] Yuliang Xiu, Jiefeng Li, Haoyu Wang, Yinghong Fang, and Cewu Lu. "Pose Flow: Efficient Online Pose Tracking". In: *British Machine Vision Conference*. 2018.

[204] Yuecong Xu, Jianfei Yang, Yunjiao Zhou, Zhenghua Chen, Min Wu, and Xiaoli Li. "Augmenting and Aligning Snippets for Few-Shot Video Domain Adaptation". In: *IEEE International Conference on Computer Vision*. 2023, pp. 13445–13456.

[205] Wanqi Xue and Wei Wang. "One-shot image classification by learning to restore prototypes". In: *AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 6558–6565.

[206] Diana Yacchirema, Jara Suárez de Puga, Carlos Palau, and Manuel Esteve. "Fall detection system for elderly people using IoT and ensemble machine learning algorithm". In: *Personal and Ubiquitous Computing* 23 (2019), pp. 801–817.

[207] Apichet Yajai, Annupan Rodtook, Krisana Chinnasarn, and Suwanna Rasmequan. "Fall detection using directional bounding box". In: *International Joint Conference on Computer Science and Software Engineering*. 2015, pp. 52–57.

[208] Hang Yan, Beichen Hu, Gang Chen, and E Zhengyuan. "Real-time continuous human rehabilitation action recognition using OpenPose and FCN". In: *2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*. 2020, pp. 239–242.

[209] Sijie Yan, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition". In: *AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[210] Luyu Yang, Yan Wang, Mingfei Gao, Abhinav Shrivastava, Kilian Q Weinberger, Wei-Lun Chao, and Ser-Nam Lim. "Deep co-training with task decomposition for semi-supervised domain adaptation". In: *IEEE International Conference on Computer Vision*. 2021, pp. 8906–8916.

[211] Shuo Yang, Lu Liu, and Min Xu. "Free Lunch for Few-shot Learning: Distribution Calibration". In: *International Conference on Learning Representations*. 2021.

[212] Siyuan Yang, Jun Liu, Shijian Lu, Er Meng Hwa, and Alex C. Kot. "One-Shot Action Recognition via Multi-Scale Spatial-Temporal Skeleton Matching". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.7 (2024), pp. 5149–5156.

[213] Yang Yang, Chunping Hou, Yue Lang, Dai Guan, Danyang Huang, and Jinchen Xu. "Open-set human activity recognition based on micro-Doppler signatures". In: *Pattern Recognition* 85 (2019), pp. 60–69.

[214] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. "Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition". In: *ACM International Conference on Multimedia*. 2020, pp. 55–63.

[215] Haben Yhdego, Jiang Li, Steven Morrison, Michel Audette, Christopher Paolini, Mahasweta Sarkar, and Hamid Okhravi. "Towards musculoskeletal simulation-aware fall injury mitigation: transfer learning with deep CNN for fall detection". In: *Spring Simulation Conference*. IEEE. 2019, pp. 1–12.

[216] Jeongbeen Yoon, Dahyun Kang, and Minsu Cho. "Semi-supervised Domain Adaptation via Sample-to-Sample Self-Distillation". In: *IEEE Winter Conference on Applications of Computer Vision* (2021), pp. 1686–1695.

[217] Ryota Yoshihashi, Wen Shao, Rei Kawakami, Shaodi You, Makoto Iida, and Takeshi Naemura. "Classification-reconstruction learning for open-set recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4016–4025.

[218] Zehao Yu, Lei Jin, and Shenghua Gao. "P^2Net: Patch-Match and Plane-Regularization for Unsupervised Indoor Depth Estimation". In: *European Conference on Computer Vision*. Springer. 2020, pp. 206–222.

[219] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In: *International Conference on Machine Learning (ICML)*. Vol. 139. 2021, pp. 12310–12320.

[220] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. "Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world". In: *IEEE International Conference on Computer Vision*. 2021, pp. 1760–1770.

[221] Jiawen Zhang, Jiaqi Zhu, Yi Yang, Wandong Shi, Congcong Zhang, and Hongan Wang. "Knowledge-enhanced domain adaptation in few-shot relation classification". In: *ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 2183–2191.

[222] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. "View adaptive neural networks for high performance skeleton-based human action recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.8 (2019), pp. 1963–1978.

[223] Qinglong Zhang and Yu-Bin Yang. "Rest: An efficient transformer for visual recognition". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15475–15485.

[224] Yuhan Zhang, Bo Wu, Wen Li, Lixin Duan, and Chuang Gan. "STST: Spatial-temporal specialized transformer for skeleton-based action recognition". In: *ACM International Conference on Multimedia*. 2021, pp. 3229–3237.

[225] Zhengyou Zhang. "Microsoft Kinect Sensor and Its Effect". In: *IEEE MultiMedia* 19.2 (2012), pp. 4–10. DOI: 10.1109/MMUL.2012.24.

[226] Cankun Zhong, Wing WY Ng, Shuai Zhang, Chris D Nugent, Colin Shewell, and Javier Medina-Quero. "Multi-occupancy fall detection using non-invasive thermal vision sensor". In: *IEEE Sensors Journal* 21.4 (2020), pp. 5377–5388.

[227] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. "Temporal relational reasoning in videos". In: *European Conference on Computer Vision*. Springer. 2018, pp. 803–818.

[228] Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. "Learning placeholders for open-set recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2021, pp. 4401–4410.

[229] Yujie Zhou, Haodong Duan, Anyi Rao, Bing Su, and Jiaqi Wang. "Self-supervised Action Representation Learning from Partial Spatio-Temporal Skeleton Sequences". In: *AAAI Conference on Artificial Intelligence*. 2023, pp. 3825–3833.

[230] Yuxuan Zhou, Chao Li, Zhi-Qi Cheng, Yifeng Geng, Xuansong Xie, and Margret Keuper. "Hypergraph transformer for skeleton-based action recognition". In: *arXiv preprint arXiv:2211.09590* (2022).

[231] Anqi Zhu, Qiuhong Ke, Mingming Gong, and James Bailey. "Adaptive local-component-aware graph convolutional network for one-shot skeleton-based action recognition". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 6038–6047.

[232] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. "Crossclr: Cross-modal contrastive learning for multi-modal video representations". In: *IEEE International Conference on Computer Vision*. 2021, pp. 1450–1459.

[233] Yixiong Zou, Yemin Shi, Daochen Shi, Yaowei Wang, Yongsheng Liang, and Yonghong Tian. "Adaptation-oriented feature projection for one-shot action recognition". In: *IEEE Transactions on Multimedia* 22.12 (2020), pp. 3166–3179.

[234] Yixiong Zou, Yemin Shi, Yaowei Wang, Yu Shu, Qingsheng Yuan, and Yonghong Tian. "Hierarchical temporal memory enhanced one-shot distance learning for action recognition". In: *IEEE International Conference on Multimedia and Expo*. 2018, pp. 1–6.