Navigating Open Set Scenarios for Skeleton-Based Action Recognition

Kunyu Peng[1], Cheng Yin[1], Junwei Zheng[1], Ruiping Liu[1], David Schneider[1], Jiaming Zhang[1], Kailun Yang[2,*], M. Saquib Sarfraz[1,3], Rainer Stiefelhagen[1], Alina Roitberg[4]

[1]Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany
[2]School of Robotics, Hunan University, China
[3]Mercedes-Benz Tech Innovation, Germany
[4]Institute for Artificial Intelligence, University of Stuttgart, Germany
kunyu.peng@kit.edu, ujfib@student.kit.edu, junwei.zheng@kit.edu, ruiping.liu@kit.edu, david.schneider@kit.edu, jiaming.zhang@kit.edu, kailun.yang@hnu.edu.cn, saquib.safraz@kit.edu, rainer.stiefelhagen@kit.edu, alina.roitberg@ki.uni-stuttgart.de

Abstract

In real-world scenarios, human actions often fall outside the distribution of training data, making it crucial for models to recognize known actions and reject unknown ones. However, using pure skeleton data in such open-set conditions poses challenges due to the lack of visual background cues and the distinct sparse structure of body pose sequences. In this paper, we tackle the unexplored **O**pen-**S**et **S**keleton-based **A**ction **R**ecognition (OS-SAR) task and formalize the benchmark on three skeleton-based datasets. We assess the performance of seven established open-set approaches on our task and identify their limits and critical generalization issues when dealing with skeleton information. To address these challenges, we propose a distance-based cross-modality ensemble method that leverages the cross-modal alignment of skeleton joints, bones, and velocities to achieve superior open-set recognition performance. We refer to the key idea as CrossMax - an approach that utilizes a novel cross-modality mean max discrepancy suppression mechanism to align latent spaces during training and a cross-modality distance-based logits refinement method during testing. CrossMax outperforms existing approaches and consistently yields state-of-the-art results across all datasets and backbones. We will release the benchmark, code, and models to the community.

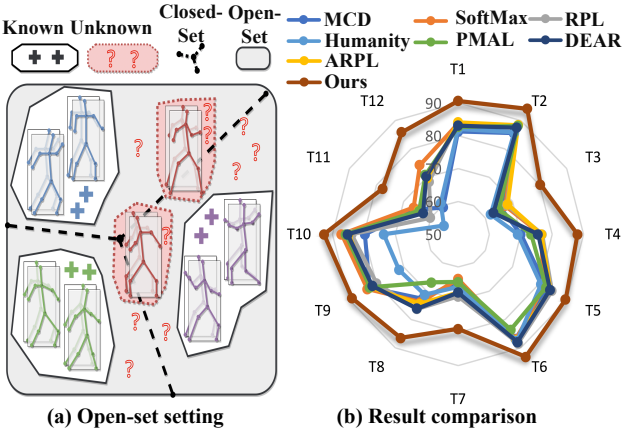

Figure 1: (a) The open-set skeleton-based action recognition setting. (b) Compared to previous methods, our method consistently achieves state-of-the-art performance. The tasks T1-T4 are based on the CTR-GCN backbone and use Cross-Subject and Cross-View splits of NTU60 to evaluate with O-AUROC and O-AUPR metrics, T5-T8 are with HD-GCN, and T9-T12 are with Hyperformer, respectively.

Introduction

Leveraging body pose sequences for human action recognition offers several benefits, such as enhanced privacy, reduced data volume, and better generalization to novel human appearances. Modern skeleton-based approaches (Zhou et al. 2022), once trained, remain static in their set of possible predictions. A more realistic scenario is the model's exposure to *open sets*, where both, known and novel action categories may occur at any time (Meyer and Drummond 2019). Out-of-distribution actions – those that fall outside the model's known repertoire – typically result in misclassifications as one of the known categories, eventually leading to significant disruptions, particularly when these recognition outputs directly steer decision-making, *e.g.*, in assistive robots. As pointed out by researchers in the past (Miller et al. 2018; Fontanel et al. 2020) there is a pressing need for further exploration in open-set, skeleton-based human action recognition, which is the main motivation of our work.

Several methods target open-set action recognition *in videos* (Bao, Yu, and Kong 2021), but the problem of detecting novel behaviors from *skeleton* streams has been overlooked so far. These tasks pursue similar goals, yet differ substantially: the absence of visual background as an additional cue context and the sparse characteristic structure of body pose sequences introduce unique challenges in managing out-of-distribution actions. To address the lack of a suitable evaluation testbed, we first build an expansive benchmark for **O**pen-**S**et **S**keleton-based **A**ction **R**ecognition (OS-SAR), comprising three prominent skeleton-based action recognition backbones: CTRGCN (Chen et al. 2021), HDGCN (Lee et al. 2022), and Hyperformer (Ding et al. 2023). This benchmark is derived from three public datasets for action recognition from

*Corresponding author.

body pose sequences – NTU60 (Shahroudy et al. 2016), NTU120 (Liu et al. 2020), and ToyotaSmartHome (Dai et al. 2023) – for which we formalize the open-set splits and an evaluation protocol. Effective and generalizable open-set recognition techniques should maintain stable performance for diverse combinations of datasets and backbones. Following the open-set recognition practices in image classification (Lu et al. 2022), we randomly sample sets of unseen classes and compute the averaged performance over five random splits. However, presumably due to inherent differences between image/video and skeletal data, common open-set recognition strategies struggle to deliver consistent OS-SAR results and the recognition quality considerably fluctuates when considering different backbones and datasets. This inconsistency underscores that the current methodologies struggle when deployed for OS-SAR challenges. A deeper examination reveals that the predicted open-set probability estimates of the existing methods are not realistic when exposed to a mix of in- and out-of-distribution skeletal sequences, which detrimentally affects open-set performance, steering models towards unwarranted overconfidence.

To tackle this problem, we introduce a new approach for OS-SAR. Our method is multimodal and builds on three streams: joints, velocities, and bones, which we enable distribution-wise information exchange in their latent space via a novel Cross-Modality Mean Max Discrepancy (CrossMMD) suppression mechanism. We also need to address the overconfidence of the SoftMax-normalized probability estimation when mixing in- and out-of-distribution samples (Liu et al. 2020). To this intent, we introduce a distance-based confidence measure based on the **C**hannel **N**ormalized **E**uclidean distance (CNE-distance) to the nearest latent space embeddings from the training set. This distance-based approach significantly improves the open-set recognition performance but falls short when it comes to the conventional close-set results compared to the vanilla SoftMax. To have the best of both worlds, we propose a cross-modality distance-based logits refinement technique, which combines logits averaged across the modalities and the proposed CNE-distances. We refer to our complete method as CrossMax, as it considers both, CrossMMD during training and cross-modality distance-based refinement during testing. CrossMax achieves state-of-the-art performances across datasets, backbones, and evaluation settings, shown in Fig. 1. The benchmark and code can be found in https://github.com/KPeng9510/OS-SAR.

Our main contributions are as follows:

- A large-scale benchmark for Open-Set Skeleton-based Action Recognition (OS-SAR), featuring three datasets for classification from body pose sequences, seven open-set recognition baselines, and three well-established backbones for skeleton data streams.
- A multimodal approach for OS-SAR leveraging three streams: joints, velocities, and bones, and enabling the distribution-wise information exchange among them using the novel Cross-Modality Mean Max Discrepancy (CrossMMD) suppression mechanism.
- A distance-based confidence measure, the Channel Normalized Euclidean distance (CNE-distance), to address overconfidence in SoftMax-normalized probability estimates and enhance open-set recognition.
- The complete CrossMax methodology combines the aforementioned CrossMMD and the distance-based logits refinement technique, achieving state-of-the-art performance across various evaluations.

Related Work

Skeleton-based action recognition aims at recognizing action categories using the skeletal geometric information (Ke et al. 2017; Liu, Liu, and Chen 2017; Duan et al. 2022). Most well-established methods are graph convolutional neural networks (GCN)-based (Kipf and Welling 2016; Yan, Xiong, and Lin 2018; Shi et al. 2019; Cheng et al. 2020; Ye et al. 2020; Chen et al. 2021), more recent approaches leverage transformer architectures (Shi et al. 2020; Plizzari, Cannici, and Matteucci 2021; Lee et al. 2022; Zhou et al. 2022; Ding et al. 2023; Xin et al. 2023). CTRGCN (Chen et al. 2021), HDGCN (Liang et al. 2019), and Hyperformer (Ding et al. 2023) serve as backbones in our OS-SAR experiments due to their superior performances and large architecture discrepancy which allows for an evaluation regarding cross-backbone generalizability.

Open-set recognition, aiming at distinguishing classes, unseen during training (Scheirer et al. 2013), 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. (Berti et al. 2022) presented an approach for one-shot OS-SAR, but do not present methods for the general OS-SAR task. Due to the large discrepancy regarding this task, we resort to several well-established open-set image classification and open-set video-based action recognition approaches which can be adapted for OS-SAR by replacing backbone and input data. Shi *et al* (Shi 2023) proposed 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 (Bao, Yu, and Kong 2021), 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, multiple works (Hendrycks and Gimpel 2017; Yoshihashi et al. 2019; Sun et al. 2020; Chen et al. 2020, 2022; Lu et al. 2022; Geng and Chen 2020; Oza and Patel 2019) were presented. (Hendrycks and Gimpel 2017) first used the highest SoftMax score as the open-set probability, followed by reconstruction-based approaches (Yoshihashi et al. 2019; Oza and Patel 2019; Sun et al. 2020; Cen et al. 2023). Recently, the most promising works are prototype-based methods (Chen et al. 2020, 2022; Sun et al. 2020). Reciprocal points distance served as open-set probability in (Chen et al. 2020) and (Chen et al. 2022) while PMAL (Lu et al. 2022) is the state-of-the-art approach. (Cen et al. 2023) proposed a new task for uni-

fied few-shot open-set recognition. We choose to use SoftMax, RPL, ARPL, and PMAL as OS-SAR baselines. SoftMax could serve as a lower bound for OS-SAR while the rest have large potential to deliver superior performances in OS-SAR due to the success of these methods in the open-set image classification. In open-set video-based action recognition task, at the early stage, (Shu et al. 2018) proposed Open Deep Network (ODN) by adding novel classes incrementally to the recognition head to achieve awareness of new classes. (Krishnan, Subedar, and Tickoo 2018) and (Subedar et al. 2019) leveraged bayesian neural networks to achieve reliable uncertainty estimation. DEAR (Bao, Yu, and Kong 2021) constructed a large-scale benchmark for open-set video-based human action recognition. They also proposed an architecture that uses deep evidential learning and delivers state-of-the-art performance. Humpty Dumpty (Du et al. 2023) (renamed as Humpty in our benchmark) uses clip-wise relational graphical reconstruction error as the open-set probability. Monte Carlo Dropout with Voting (MCD-V) is proposed by (Roitberg et al. 2020) for open-set video-based driver action recognition. (Yang et al. 2019) leveraged micro-doppler radar data, we do not adapt this model due to its specific architecture for such a modality. DEAR, Humpty, and MCD-V serve as OS-SAR baselines. These baselines do not show consistent performances across datasets and backbones, displaying the need for a generalizable OS-SAR method. Thereby, we propose CrossMax which uses cross-modality mean max suppression in the training to enable cross-modality information exchange, and cross-modality distance-based logits refinement in the testing to refine the salient and non-salient logits position separately, introduced in the following section in detail.

Method

Benchmark

We introduce OS-SAR, a large-scale benchmark for **O**pen-**S**et **S**keleton-based **A**ction **R**ecognition, leveraging CTR-GCN (Chen et al. 2021), HDGCN (Liang et al. 2019), and Hyperformer (Ding et al. 2023) as the backbones to validate the cross-backbone generalizability. We build on the NTU60 (Shahroudy et al. 2016), NTU120 (Liu et al. 2020), and ToyotaSmartHome (Dai et al. 2023) datasets for human action recognition from body pose sequences and adapt their splits to suit open set conditions. Backbones and baselines are presented in the following, while the dataset introduction will be covered in the experiments section.

Skeleton Representation Backbones. *CTRGCN* (Chen et al. 2021) used Channel-wise Topology Refinement Graph Convolution (CTRGC) to dynamically learn distinct topologies and efficiently aggregate features in different channels. *HDGCN* (Liang et al. 2019) is based on Hierarchically Decomposed (HD) GCN by leveraging an HD-Graph that decomposes nodes into multiple sets to capture both structurally adjacent and distant edges with semantic relevance. *Hyperformer* (Ding et al. 2023) is a transformer-based approach that incorporates bone connectivity via graph distance embedding. We selected these architectures to validate the cross-backbone generalizability of open-set methods due to their strong performance in conventional skeleton-based human action recognition benchmarks and the different building blocks of their underlying architectures.

Existing Open-Set Recognition Baselines. Our baselines are open-set recognition methods from the image/video classification task, as there are no existing method for open-set recognition from skeleton yet, which can be adapted to diverse skeleton-based backbones. *Open-set baselines from image classification:* We selected principal-point distance-based approaches, *i.e.*, RPL (Chen et al. 2020) and ARPL (Chen et al. 2022), prototype learning-based approach, *i.e.*, PMAL (Lu et al. 2022), which is current state-of-the-art in open-set image classification, and the vanilla SoftMax score (Hendrycks and Gimpel 2017) as baselines. *Open-set baselines from video-based action recognition:* We choose DEAR (Bao, Yu, and Kong 2021), which uses deep evidential learning for open-set probability estimation, Monte Carlo Dropout + Voting (MCD-V) (Roitberg et al. 2020), and Humpty (Du et al. 2023), which uses temporal graph reconstruction as the open-set probability. We use the aforementioned skeleton backbones individually in all the selected open-set recognition baselines to achieve a fair comparison, where the image/video backbones are replaced by the selected skeleton-based backbones.

CrossMax

We propose CrossMax, a novel OS-SAR method leveraging three complementary skeleton modalities: joints, bones, and velocities. CrossMax first employs ensembled backbones for feature extraction while using Cross-modality Mean-Max Discrepancy suppression (CrossMMD) in training to enhance information exchange and reduce modality disparities. We further introduce a novel cross-modality distance-based logits refinement using Channel-Normalized Euclidean distance (CNE-distance). This refinement method significantly improves open-set probability estimation and close-set classification, as demonstrated in Fig. 2.

Skeleton Modalities. Given skeleton joints as $j = \{j_1^{\{1,...,N_j\}}, ..., j_T^{\{1,...,N_j\}}\}$, where T denotes the total frame number of the skeleton sequence and N_j indicates the joint number, bones b and velocities v can be calculated through $v = \{j_t^{\{1,...,N_j\}} - j_{t-1}^{\{1,...,N_j\}} | t \in [1, T]\}$, indicating the velocity of joints during motion at timestamp t, and $b = \{j^x_{\{1,...,T\}} - j^y_{\{1,...,T\}} | (x, y) \in \Omega_b\}$, indicating the bone vector, where Ω_b indicates the set of bones, as depicted in Fig. 2.

CrossMMD. To enable better cross-modal exchange, we present CrossMMD. The Mean Maximal Discrepancy (MMD) is used to quantify the dissimilarity between probability distributions (Gretton et al. 2012). In our approach, MMD serves as a loss function, encouraging greater similarity between distributions. Note, that there is a lack of MMD-related research tailored for cross-modality scenarios. Our primary goal is to diminish the significant discrepancy between latent spaces originating from diverse modalities. By doing so, we aim to leverage the intrinsic open-set discriminative cues to grasp the advantage of each branch, thereby

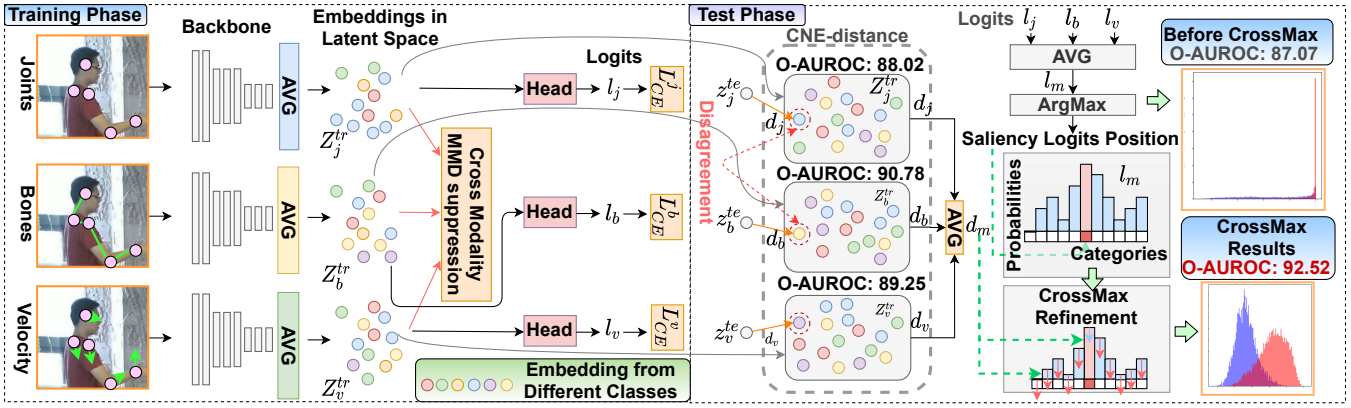

Figure 2: An overview of CrossMax. During training, we utilize the Cross-modality Mean Maximum Discrepancy (CrossMMD), to better align the latent spaces across different modalities. At test-time, for each modality, we calculate the Euclidean distance to the closest training set sample and combine this 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.

facilitating an exchange of information based on distribution. We introduce the Cross-modality Mean Max Discrepancy Suppression Mechanism (CrossMMD) to address this challenge. The Gaussian kernel is used to Reproduce Kernel Hilbert Space (RKHS). Let Ω_x and Ω_y denote two embedding batches from two different modalities, which can be interpreted as two distributions. First, we concatenate them to form $z = Concat(\Omega_x, \Omega_y)$. Then, we compute the pairwise L2 Norm distance between all samples of z denoted as d_z. The bandwidths are chosen according to Eq. 1 to determine the scales of the kernel function, which are influenced by both the sum of the distances and the sample number,

$$BW = \frac{\sum(d_z)}{(N_z)^2 - N_z}, \tag{1}$$

where N_z denotes the sample number. Let N_k indicate the kernel number, we obtain the bandwidth list L_{BW} as $\{BW * (\alpha)^i | \ i \in [0, N_k)\}$, where α is a scaling factor. Small bandwidths focus on capturing fine-grained dissimilarities among embeddings, which can be useful when the distributions have intricate local shapes. Large bandwidths capture broader cues and global discrepancies. The kernel matrix of the given embeddings is obtained by,

$$\mathbb{H}_k = \{exp(-\frac{d_z}{\beta}) | \beta \in L_{BW}\}. \tag{2}$$

For $\mathcal{K} \in \mathbb{H}_k$, the intra-source differences can be calculated via Eq. 3,

$$\begin{aligned} \text{Intra}(Z_j^{tr}, Z_b^{tr}, Z_v^{tr}) = \mathbb{E}\left[\sum_{\mathcal{K}\in\mathbb{H}_k} \mathcal{K}(Z_j^{tr}, Z_j^{tr})\right] + \\ \mathbb{E}\left[\sum_{\mathcal{K}\in\mathbb{H}_k} \mathcal{K}(Z_b^{tr}, Z_b^{tr})\right] + \mathbb{E}\left[\sum_{\mathcal{K}\in\mathbb{H}_k} \mathcal{K}(Z_v^{tr}, Z_v^{tr})\right], \end{aligned} \tag{3}$$

where $\mathbb{E}$ indicates empirical mean average, Z_j^{tr}, Z_b^{tr}, and Z_v^{tr} indicate the embeddings learned from three modalities during training as shown in Fig. 2, while the inter-source differences among different modalities is calculated by Eq. 4,

$$\begin{aligned} \text{Inter}(Z_j^{tr}, Z_b^{tr}, Z_v^{tr}) = \mathbb{E}\left[\sum_{\mathcal{K}\in\mathbb{H}_k} \mathcal{K}(Z_j^{tr}, Z_v^{tr})\right] + \\ \mathbb{E}\left[\sum_{\mathcal{K}\in\mathbb{H}_k} \mathcal{K}(Z_j^{tr}, Z_b^{tr})\right] + \mathbb{E}\left[\sum_{\mathcal{K}\in\mathbb{H}_k} \mathcal{K}(Z_b^{tr}, Z_v^{tr})\right]. \end{aligned} \tag{4}$$

The final CrossMMD is calculated as the discrepancy between the intra- and inter-source differences as Eq. 5,

$$\begin{aligned} \text{CrossMMD}(Z_j^{tr}, Z_b^{tr}, Z_v^{tr}) = \text{Intra}(Z_j^{tr}, Z_b^{tr}, Z_v^{tr}) - \\ \text{Inter}(Z_j^{tr}, Z_b^{tr}, Z_v^{tr}). \end{aligned} \tag{5}$$

CrossMMD is chosen as a loss function L_{MMD} to enable the multi-scale information exchange among different modalities on comparable scales provided by the Gaussian kernels. Apart from the L_{MMD}, we also use the cross-entropy loss on the training set, depicted as Eq. 6,

$$L_{\text{overall}} = L_{\text{CE}}^j + L_{\text{CE}}^b + L_{\text{CE}}^v + \lambda \cdot L_{\text{MMD}}, \tag{6}$$

where λ is chosen as a fixed value to keep the two losses having the same gradients scale. L_{CE}^j, L_{CE}^b, and L_{CE}^v denote cross entropy losses for three branches, respectively.

Cross-modality Distance-based Logits Refinement. By utilizing the averaged logits from three branches constrained by CrossMMD, the model yields a predicted open-set probability by using the highest score from SoftMax on the logits. However, though the performance of the open-set probability prediction increases, we find that SoftMax-based probability prediction suffers from a bad disentanglement between in- and out-of-distribution samples in terms of the open-set probability distribution, which limits the further improvement of OS-SAR.

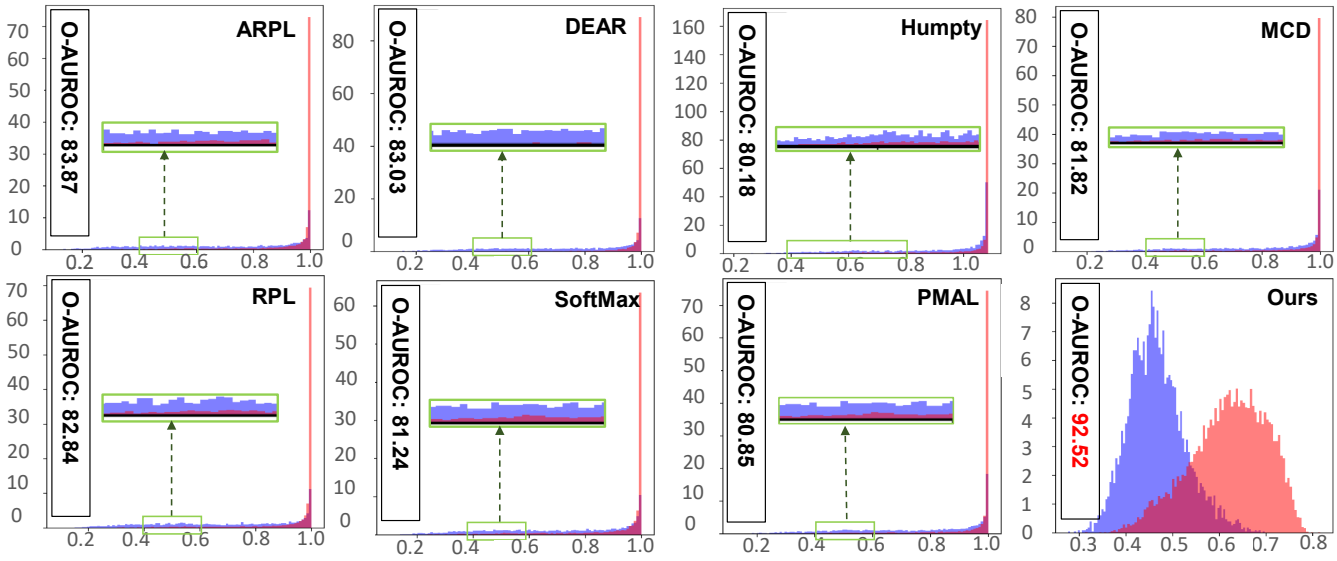

Figure 3: Comparison of the open-set probabilities. All the methods use HD-GCN on NTU60 (CS) on one random split.

To overcome this limitation, we propose a Channel-Normalized Euclidean distance (CNE-distance). This mechanism achieves Gaussian-wise probability distributions and ensures better disentanglement between the in- and out-of-distribution samples. We first extract the embedding for the training samples considering three modalities and obtain the embedding sets as Z_a^{tr}, where $a \in \{j, b, v\}$. Then, we follow the same procedure to extract the embedding of the test sample, *i.e.*, z_a^{te}. For each sample from the test set, we can obtain three distances according to the corresponding nearest embedding in the Z_a^{tr}, *i.e.*, d_j, d_b, and d_v. We first utilize L2 normalization along the channel dimension for each embedding to map the feature value between 0 and 1. Then the Euclidean distance to the nearest training set embedding is used as the open-set probability. In summary, the CNE-distance can be calculated as Eq. 7,

$$\begin{aligned} &d_j, d_b, d_v = D\left[\mathcal{N}_C(z_j^{te}), \mathcal{N}_C(Z_j^{tr})\right], \\ &D\left[\mathcal{N}_C(z_b^{te}), \mathcal{N}_C(Z_b^{tr})\right], D\left[\mathcal{N}_C(z_v^{te}), \mathcal{N}_C(Z_v^{tr})\right], \end{aligned} \tag{7}$$

where te and tr indicate the test and training set and $D[\cdot]$ is Euclidean distance. $\mathcal{N}_C(\cdot)$ indicates the channel normalization. The averaged distance can be obtained by Eq. 8,

$$d_m = Mean(d_j, d_b, d_v). \tag{8}$$

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. 2, the results 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 closed-set classification. We first acquire the position with the highest logit value of the averaged logits by Eq. 9,

$$M_P = ArgMax((l_j + l_b + l_v)/3), \tag{9}$$

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. 10 considering a given sample, where the salient logit position is indicated by a one-hot mask M_P,

$$l_m\left[M_P\right] := Log(exp(l_m\left[M_P\right] * d_m^2)(\frac{1}{d_m} - 1)). \tag{10}$$

While the not salient positions are indicated by mask M_{NP}, the not saliency logits are refined by Eq. 11,

$$l_m\left[M_{NP}\right] := l_m\left[M_{NP}\right] * d_m^2. \tag{11}$$

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 P_{prob}=$Max(SoftMax(l_m))$, 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.

Experiments

Datasets and Metrics

Datasets. NTU60 (Shahroudy et al. 2016) involves $56,880$ samples of 60 action classes. We randomly choose

B	Method	NTU60						NTU120					
		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**

Table 1: Experiments on NTU60 (Shahroudy et al. 2016) and NTU120 (Liu et al. 2020) 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 (Chen et al. 2020), ARPL (Chen et al. 2022), PMAL (Lu et al. 2022), the vanilla SoftMax score (SoftMax) (Hendrycks and Gimpel 2017), Monte Carlo Dropout + Voting (MCD-V) (Roitberg et al. 2020), DEAR (Bao, Yu, and Kong 2021), and Humpty (Du et al. 2023) are chosen as open-set baselines to construct the benchmark.

20 classes as out-of-distribution classes. NTU120 (Liu et al. 2020) involves 120 action classes. We randomly choose 90 classes as out-of-distribution classes. ToyotaSmartHome (Dai et al. 2023) contains $16,115$ samples with 31 classes, which is challenging since occlusion from real-world scenarios is involved. 18 action classes are selected as out-of-distribution classes.

Metrics. The area under the receiver operating characteristic (O-AUROC) and area under the precision-recall curve (O-AUPR) are the most important metrics to evaluate the open-set performance with different focuses regarding the category balancing. Alongside, close-set classification accuracy (C-ACC) is chosen to evaluate whether the open-set method can preserve good classification capability or not. O-AUROC and C-ACC metrics are selected following PMAL (Lu et al. 2022), while O-AUPR is additionally provided since ToyotaSmartHome is unbalanced.

Implementation Details

Our method relies on PyTorch1.8.0 and is trained with SGD optimizer with learning rate (lr) 0.1, step-wise lr scheduler with decay rate 0.1, steps for decay at $\{35, 55, 70\}$, weight decay 0.0004, and batch size 64 for 100 epochs on 4 Nvidia A100 GPUs with Intel Xeon Gold 6230 processor. λ, N_k, and α are chosen as 0.1, 5, and 2.0, respectively. In total, our method has 4.29 MB, 5.04 MB, and 7.8 MB number of parameters on CTRGCN, HDGCN, and Hyperformer.

Benchmark Analysis

We first give a comprehensive analysis the existing open-set recognition approaches on our benchmark in Tab. 1 and Tab.2. Taking cross-backbone generalizability into consideration, principal points distance-based approaches, *i.e.*, RPL (Chen et al. 2020) and ARPL (Chen et al. 2022), achieve 0.34% and 0.45% O-AUROC improvements on CTRGCN and 1.40% and 2.40% O-AUROC improvements on HDGCN compared with SoftMax (Hendrycks and Gimpel 2017) on NTU60 (CS). However, their performances are below SoftMax on the HyperFormer on NTU60, showing that these approaches can not well generalize to different GCN backbones.

Then we focus on cross-dataset generalizability. On NTU120, RPL and ARPL can achieve better performances compared with SoftMax on HDGCN and Hyperformer, while on ToyotaSmartHome (CS), RPL fails to work well

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

Table 2: Experiments on ToyotaSmartHome (Dai et al. 2023) dataset, where CS, CV, and B indicate **C**ross-**S**ubject/**V**iew evaluations and **B**ackbone. The results are averaged for five random splits.

on CTRGCN and both of them only work better compared with SoftMax on HDGCN. Considering the prototypical learning approach, *i.e.*, PMAL (Lu et al. 2022), it generally does not work well on the OS-SAR. Three open-set approaches from video-based action recognition task, the deep evidential learning approach, *i.e.*, DEAR (Bao, Yu, and Kong 2021), the Monte Carlo Dropout + Voting approach, *i.e.*, MCD-V (Roitberg et al. 2020), and the temporal relationship reconstruction approach, *i.e.*, Humpty (Du et al. 2023), unfortunately, deliver limited performances for OS-SAR. This consequence is caused by the large differences between RGB image/video data and the skeleton data, as skeleton data lacks a majority of the background cues and visual appearance cues while the data format is quite sparse. Another unignorable reason is that the networks utilized for feature extraction of skeleton data are mostly GCNs instead of convolutional neural networks (CNNs) or rely on graph architecture, where different manifolds on the latent space could be delivered due to the backbone discrepancy. This observation indicates the critical need to develop a generalizable OS-SAR approach that can work well across datasets and backbones.

To handle the underlying issue for existing open-set recognition approaches, we analyze the disentanglement between the in- and out-of-distribution samples considering the open-set probability in Fig. 3, 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 NTU60 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.

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

Table 3: Module ablation on NTU60 (CS) on CTRGCN, where the results are averaged among five random splits.

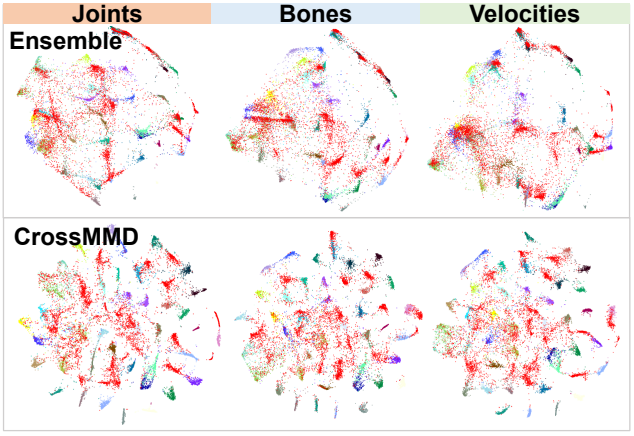

Figure 4: T-SNE (van der Maaten and Hinton 2008) visualizations on NTU60 (CS) using CTRGCN. Out- and in-of-distribution samples are marked by red and other colors.

Analysis of Observations and Ablations

Benefits by using CrossMMD. To analyze the benefits from CrossMMD, t-SNE visualizations are provided considering training w/o CrossMMD (marked as Ensemble), and w/ CrossMMD (marked as CrossMMD), in Fig. 4. We ob-

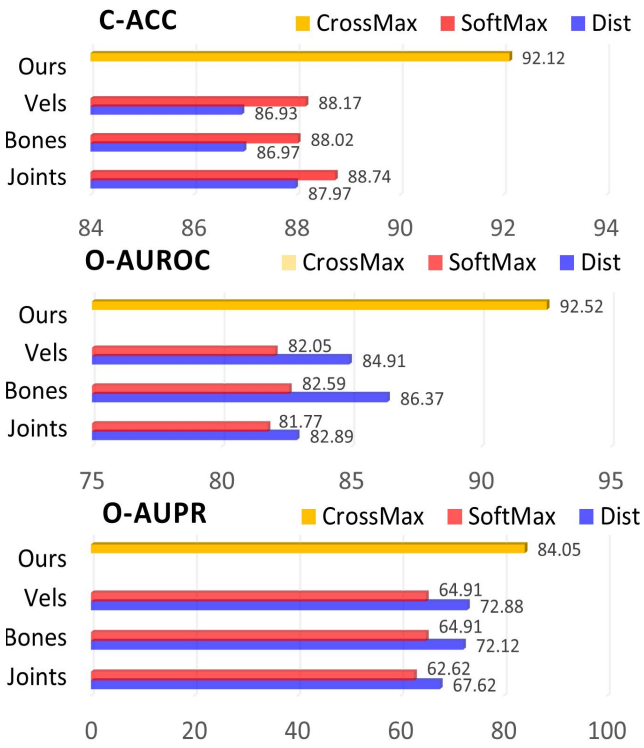

Figure 5: Comparison of SoftMax, CNE-distance, and CrossMax using HDGCN on NTU60 (CS) on one split.

serve that by using CrossMMD, the latent spaces are more discriminative and structured for in- and out-of-distribution samples, which matches the performance benefits introduced in Tab. 3, where Ensemble and CrossMMD both use SoftMax score for the open-set probability estimation.

Comparison between CNE-distance *vs*. vanilla SoftMax. We observe that the open-set recognition performances for O-AUROC and O-AUPR of CNE-distance on different modalities are much better compared with those of the vanilla SoftMax w/o logits refinement and deliver the proof in Fig. 5. Compared with vanilla SoftMax, CNE-distance can achieve 2.86%, 3.78%, and 1.12% O-AUROC improvements for joints, bones, and velocities, respectively. However, as mentioned in the introduction, CNE-distance has a shortcoming in achieving a satisfied decision on the close-set classification. The proposed logits refinement takes advantage of both vanilla SoftMax and the CNE-distance and it achieves superior open-set and close-set performances.

Comparison between logits refinement *vs*. CNE-distance. After the above-mentioned analysis, there would be a question regarding how well the proposed cross-modality distance-based logits refinement outperforms CNE-distance ablations in terms of the open-set probability prediction ability. To delve deeper into this question, we showcase a comparison in Fig. 6 where the results are from five random splits marked as R1 to R5. We choose CNE-distance ablations as joint-modality distance (Dist_joints), bone-modality distance (Dist_bones), velocity-modality distance (Dist_velocities), the min aggregation (Dist_min), and the max aggregation (Dist_max) over three branches. If the shape of the curve for one approach is a regular pentagon, the performance is stable and robust over different splits. We observe that the prediction by using our logits refinement achieves the best stable performance compared with the others, indicating that using the cross-modality logits refinement method can even harvest more stable open-set probability estimation performances while preserving the superior close-set classification ability.

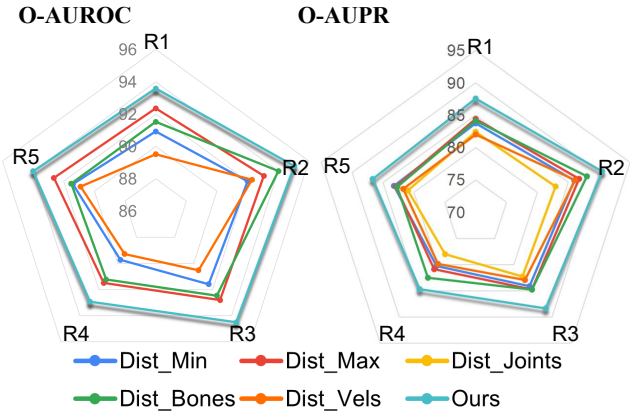

Figure 6: Comparison for different methods using CTRGCN on NTU60 (CV) for five random splits.

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 Tab. 3, where *Ensemble* indicates using ensembled modalities and vanilla SoftMax, *CrossMMD* indicates using CrossMMD and vanilla SoftMax, and *CrossMax* indicates using CrossMMD and cross-modality logits refinement. *CrossMMD* achieves 2.08%, 3.45%, and 0.37% improvements for O-AUROC, O-AUPR, and C-ACC, 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.

Conclusion

We propose the OS-SAR benchmark to contribute a large-scale test bed for open-set skeleton-based action recognition across backbones and datasets while selecting seven well-established open-set recognition methods serving as baselines. We identify that most existing open-set recognition methods do not work well on OS-SAR and thereby propose CrossMax using CrossMMD during the training phase and cross-modality distance-based logits refinement during the test phase. Our approach achieves state-of-the-art performances on OS-SAR while indicating great ability in disentangling the in- and out-of-distribution samples in terms of the predicted open-set probability.

Acknowledgments

This work was supported in part by the SmartAge project sponsored by the Carl Zeiss Stiftung (P2019-01-003; 2021-2026), the University of Excellence through the "KIT Future Fields" project, in part by the Helmholtz Association Initiative and Networking Fund on the HoreKA@KIT partition and the state of Baden-Württemberg through bwHPC and the German Research Foundation through grant INST 35/1597-1 FUGG. This work is also supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition. A. Roitberg was supported by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy - EXC 2075.

References

Bao, W.; Yu, Q.; and Kong, Y. 2021. Evidential deep learning for open set action recognition. In *ICCV*.

Berti, S.; Rosasco, A.; Colledanchise, M.; and Natale, L. 2022. One-shot open-set skeleton-based action recognition. In *Humanoids*.

Cen, J.; Luan, D.; Zhang, S.; Pei, Y.; Zhang, Y.; Zhao, D.; Shen, S.; and Chen, Q. 2023. The devil is in the wrongly-classified samples: Towards unified open-set recognition. *arXiv preprint arXiv:2302.04002*.

Chen, G.; Peng, P.; Wang, X.; and Tian, Y. 2022. Adversarial reciprocal points learning for open set recognition. *TPAMI*.

Chen, G.; Qiao, L.; Shi, Y.; Peng, P.; Li, J.; Huang, T.; Pu, S.; and Tian, Y. 2020. Learning open set network with discriminative reciprocal points. In *ECCV*.

Chen, Y.; Zhang, Z.; Yuan, C.; Li, B.; Deng, Y.; and Hu, W. 2021. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *ICCV*.

Cheng, K.; Zhang, Y.; Cao, C.; Shi, L.; Cheng, J.; and Lu, H. 2020. Decoupling gcn with dropgraph module for skeleton-based action recognition. In *ECCV*.

Dai, R.; Das, S.; Sharma, S.; Minciullo, L.; Garattoni, L.; Bremond, F.; and Francesca, G. 2023. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *TPAMI*.

Ding, K.; Liang, A. J.; Perozzi, B.; Chen, T.; Wang, R.; Hong, L.; Chi, E. H.; Liu, H.; and Cheng, D. Z. 2023. HyperFormer: Learning expressive sparse feature representations via hypergraph transformer. In *SIGIR*.

Du, D.; Shringi, A.; Hoogs, A.; and Funk, C. 2023. Reconstructing humpty dumpty: Multi-feature graph autoencoder for open set action recognition. In *WACV*.

Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; and Dai, B. 2022. Revisiting skeleton-based action recognition. In *CVPR*.

Fontanel, D.; Cermelli, F.; Mancini, M.; Bulo, S. R.; Ricci, E.; and Caputo, B. 2020. Boosting deep open world recognition by clustering. *RA-L*.

Geng, C.; and Chen, S. 2020. Collective decision for open set recognition. *TKDE*.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A kernel two-sample test. *JMLR*.

Hendrycks, D.; and Gimpel, K. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*.

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A new representation of skeleton sequences for 3D action recognition. In *ICCV*.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krishnan, R.; Subedar, M.; and Tickoo, O. 2018. BAR: Bayesian activity recognition using variational inference. *arXiv preprint arXiv:1811.03305*.

Lee, J.; Lee, M.; Lee, D.; and Lee, S. 2022. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. *arXiv preprint arXiv:2208.10741*.

Liang, Z.; Yang, M.; Deng, L.; Wang, C.; and Wang, B. 2019. Hierarchical depthwise graph convolutional neural network for 3D semantic segmentation of point clouds. In *ICRA*.

Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.; and Kot, A. C. 2020. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *TPAMI*.

Liu, M.; Liu, H.; and Chen, C. 2017. Enhanced skeleton visualization for view invariant human action recognition. *PR*.

Lu, J.; Xu, Y.; Li, H.; Cheng, Z.; and Niu, Y. 2022. PMAL: Open set recognition via robust prototype mining. In *AAAI*.

Meyer, B. J.; and Drummond, T. 2019. The importance of metric learning for robotic vision: Open set recognition and active learning. In *ICRA*.

Miller, D.; Nicholson, L.; Dayoub, F.; and Sünderhauf, N. 2018. Dropout sampling for robust object detection in open-set conditions. In *ICRA*.

Oza, P.; and Patel, V. M. 2019. C2AE: Class conditioned auto-encoder for open-set recognition. In *CVPR*.

Plizzari, C.; Cannici, M.; and Matteucci, M. 2021. Spatial temporal transformer network for skeleton-based action recognition. In *ICPRW*.

Roitberg, A.; Ma, C.; Haurilet, M.; and Stiefelhagen, R. 2020. Open set driver activity recognition. In *IV*.

Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; and Boult, T. E. 2013. Toward open set recognition. *TPAMI*.

Shahroudy, A.; Liu, J.; Ng, T.-T.; and Wang, G. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *CVPR*.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*.

Shi, L.; Zhang, Y.; Cheng, J.; and Lu, H. 2020. Decoupled spatial-temporal attention network for skeleton-based action-gesture recognition. In *ACCV*.

Shi, Y. 2023. Open set action recognition based on skeleton. In *ICCCS*.

Shu, Y.; Shi, Y.; Wang, Y.; Zou, Y.; Yuan, Q.; and Tian, Y. 2018. ODN: Opening the deep network for open-set action recognition. In *ICME*.

Subedar, M.; Krishnan, R.; Meyer, P. L.; Tickoo, O.; and Huang, J. 2019. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *ICCV*.

Sun, X.; Yang, Z.; Zhang, C.; Ling, K.-V.; and Peng, G. 2020. Conditional Gaussian distribution learning for open set recognition. In *CVPR*.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *JMLR*.

Xin, W.; Liu, R.; Liu, Y.; Chen, Y.; Yu, W.; and Miao, Q. 2023. Transformer for skeleton-based action recognition: A review of recent advances. *Neurocomputing*.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*.

Yang, Y.; Hou, C.; Lang, Y.; Guan, D.; Huang, D.; and Xu, J. 2019. Open-set human activity recognition based on micro-Doppler signatures. *PR*.

Ye, F.; Pu, S.; Zhong, Q.; Li, C.; Xie, D.; and Tang, H. 2020. Dynamic GCN: Context-enriched topology learning for skeleton-based action recognition. In *MM*.

Yoshihashi, R.; Shao, W.; Kawakami, R.; You, S.; Iida, M.; and Naemura, T. 2019. Classification-reconstruction learning for open-set recognition. In *CVPR*.

Zhou, Y.; Li, C.; Cheng, Z.-Q.; Geng, Y.; Xie, X.; and Keuper, M. 2022. Hypergraph transformer for skeleton-based action recognition. *arXiv preprint arXiv:2211.09590*.