Revisiting Skeleton-based Action Recognition

Haodong Duan[1,3] Yue Zhao[2] Kai Chen[3,5] Dahua Lin[1,3] Bo Dai[3,4] ✉
[1]The Chinese University of HongKong [2]The University of Texas at Austin
[3]Shanghai AI Laboratory [4]S-Lab, Nanyang Technological University [5]SenseTime Research

Abstract

Human skeleton, as a compact representation of human action, has received increasing attention in recent years. Many skeleton-based action recognition methods adopt GCNs to extract features on top of human skeletons. Despite the positive results shown in these attempts, GCN-based methods are subject to limitations in robustness, interoperability, and scalability. In this work, we propose PoseConv3D, a new approach to skeleton-based action recognition. PoseConv3D relies on a 3D heatmap volume instead of a graph sequence as the base representation of human skeletons. Compared to GCN-based methods, PoseConv3D is more effective in learning spatiotemporal features, more robust against pose estimation noises, and generalizes better in cross-dataset settings. Also, PoseConv3D can handle multiple-person scenarios without additional computation costs. The hierarchical features can be easily integrated with other modalities at early fusion stages, providing a great design space to boost the performance. PoseConv3D achieves the state-of-the-art on five of six standard skeleton-based action recognition benchmarks. Once fused with other modalities, it achieves the state-of-the-art on all eight multi-modality action recognition benchmarks. Code has been made available at: https://github.com/kennymckormick/pyskl.

1. Introduction

Action recognition is a central task in video understanding. Existing studies have explored various modalities for feature representation, such as RGB frames [6, 54, 59], optical flows [47], audio waves [62], and human skeletons [60, 64]. Among these modalities, skeleton-based action recognition has received increasing attention in recent years due to its action-focusing nature and compactness. In practice, human skeletons in a video are mainly represented as a sequence of joint coordinate lists, where the coordinates are extracted by pose estimators. Since only the pose information is included, skeleton sequences capture only action information while being immune to contextual nuisances, such as background variation and lighting changes.

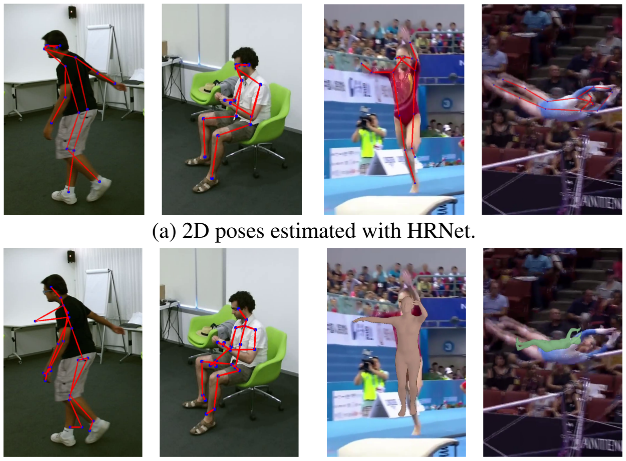
(a) 2D poses estimated with HRNet.
(b) 3D poses collected with Kinect. (c) 3D poses estimated with VIBE.

Figure 1. **PoseConv3D takes 2D poses as inputs.** In general, 2D poses are of better quality than 3D poses. We visualize 2D poses estimated with HRNet for videos in NTU-60 and FineGYM in (a). Apparently, their quality is much better than 3D poses collected by sensors (b) or estimated with state-of-the-art estimators (c).

Table 1. **Differences between PoseConv3D and GCN.**

	Previous Work	PoseConv3D
Input	2D / 3D Skeleton	2D Skeleton
Format	Coordinates	3D Heatmap Volumes
Architecture	GCN	3D-CNN

Among all the methods for skeleton-based action recognition [15, 57, 58], graph convolutional networks (GCN) [64] have been one of the most popular approaches. Specifically, GCNs regard every human joint at every timestep as a node. Neighboring nodes along the spatial and temporal dimensions are connected with edges. Graph convolution layers are then applied to the constructed graph to discover action patterns across space and time. Due to the good performance on standard benchmarks for skeleton-based action recognition, GCNs have been a standard approach when processing skeleton sequences.

While encouraging results have been observed, GCN-based methods are limited in the following aspects: (1) *Robustness:* While GCN directly handles coordinates of human joints, its recognition ability is significantly affected by the distribution shift of coordinates, which can often occur when applying a different pose estimator to acquire the coordinates. A small perturbation in coordinates often leads to completely different predictions [66]. (2) *Interoperability:* Previous works have shown that representations from different modalities, such as RGB, optical flows, and skeletons, are complementary. Hence, an effective combination of such modalities can often result in a performance boost in action recognition. However, GCN is operated on an irregular graph of skeletons, making it difficult to fuse with other modalities that are often represented on regular grids, especially in the early stages. (3) *Scalability:* In addition, since GCN regards every human joint as a node, the complexity of GCN scales linearly with the number of persons, limiting its applicability to scenarios that involve multiple persons, such as group activity recognition.

In this paper, we propose a novel framework **PoseConv3D** that serves as a competitive alternative to GCN-based approaches. In particular, PoseConv3D takes as input 2D poses obtained by modern pose estimators shown in Figure 1. The 2D poses are represented by stacks of heatmaps of skeleton joints rather than coordinates operated on a human skeleton graph. The heatmaps at different timesteps will be stacked along the temporal dimension to form a 3D heatmap volume. PoseConv3D then adopts a 3D convolutional neural network on top of the 3D heatmap volume to recognize actions. Main differences between PoseConv3D and GCN-based approaches are summarized in Table 1.

PoseConv3D can address the limitations of GCN-based approaches stated above. First, using 3D heatmap volumes is more robust to the up-stream pose estimation: we empirically find that PoseConv3D generalizes well across input skeletons obtained by different approaches. Also, PoseConv3D, which relies on heatmaps of the base representation, enjoys the recent advances in convolutional network architectures and is easier to integrate with other modalities into multi-stream convolutional networks. This characteristic opens up great design space to further improve the recognition performance. Finally, PoseConv3D can handle different numbers of persons without increasing computational overhead since the complexity over 3D heatmap volume is independent of the number of persons. To verify the efficiency and effectiveness of PoseConv3D, we conduct comprehensive studies across several datasets, including FineGYM [43], NTURGB-D [34], UCF101 [51], HMDB51 [26], Kinetics400 [6], and Volleyball [22], where PoseConv3D achieves state-of-the-art performance compared to GCN-based approaches.

2. Related Work

3D-CNN for RGB-based action recognition. 3D-CNN is a natural extension of 2D-CNN for spatial feature learning to spatiotemporal in videos. It has long been used in action recognition [23, 54]. Due to a large number of parameters, 3D-CNN requires huge amounts of videos to learn good representation. 3D-CNN has become the mainstream approach for action recognition since I3D [6]. From then on, many advanced 3D-CNN architectures [17, 18, 55, 56] have been proposed by the action recognition community, which outperform I3D both in precision and efficiency. In this work, we first propose to use 3D-CNN with 3D heatmap volumes as inputs and obtain the state-of-the-art in skeleton-based action recognition.

GCN for skeleton-based action recognition. Graph convolutional network is widely adopted in skeleton-based action recognition [3,7,19,49,50,64]. It models human skeleton sequences as spatiotemporal graphs. ST-GCN [64] is a well-known baseline for GCN-based approaches, which combines spatial graph convolutions and interleaving temporal convolutions for spatiotemporal modeling. Upon the baseline, adjacency powering is used for multiscale modeling [30, 36], while self-attention mechanisms improve the modeling capacity [28, 45]. Despite the great success of GCN in skeleton-based action recognition, it is also limited in robustness [66] and scalability. Besides, for GCN-based approaches, fusing features from skeletons and other modalities may need careful design [13].

CNN for skeleton-based action recognition. Another stream of work adopts convolutional neural networks for skeleton-based action recognition. 2D-CNN-based approaches first model the skeleton sequence as a pseudo image based on manually designed transformations. One line of works aggregates heatmaps along the temporal dimension into a 2D input with color encodings [10] or learned modules [1, 63]. Although carefully designed, information loss still occurs during the aggregation, which leads to inferior recognition performance. Other works [2,24,25,29,37] directly convert the coordinates in a skeleton sequence to a pseudo image with transformations, typically generate a 2D input of shape $K \times T$, where K is the number of joints, T is the temporal length. Such input cannot exploit the locality nature of convolution networks, which makes these methods not as competitive as GCN on popular benchmarks [2]. Only a few previous works have adopted 3D-CNNs for skeleton-based action recognition. To construct the 3D input, they either stack the pseudo images of distance matrices [21, 32] or directly sum up the 3D skeletons into a cuboid [33]. These approaches also severely suffer from information loss and obtain much inferior performance to the state-of-the-art. Our work stacks heatmaps along the temporal dimension to form 3D heatmap volumes, preserving all information during this process. Besides, we use 3D-CNN instead of 2D-CNN due to its good capability for spatiotemporal feature learning.

3. Framework

We propose **PoseConv3D**, a **3D-CNN**-based approach for skeleton-based action recognition, which can be a competitive alternative to GCN-based approaches, outperforming GCN under various settings in terms of accuracy with improved robustness, interoperability, and scalability. An overview of PoseConv3D is depicted in Figure 2, and details of PoseConv3D will be covered in the following sections. We begin with a review of skeleton extraction, which is the basis of skeleton-based action recognition but is often overlooked in previous literature. We point out several aspects that should be considered when choosing a skeleton extractor and motivate the use of 2D skeletons in PoseConv3D[1]. Subsequently, we introduce 3D Heatmap Volume that is the representation of a 2D skeleton sequence used in PoseConv3D, followed by the structural designs of PoseConv3D, including a variant that focuses on the modality of human skeletons as well as a variant that combines the modalities of human skeletons and RGB frames to demonstrate the interoperability of PoseConv3D.

3.1. Good Practices for Pose Extraction

Being a critical pre-processing step for skeleton-based action recognition, human skeleton or pose extraction largely affects the final recognition accuracy. However, its importance is often overlooked in previous literature, in which poses estimated by sensors [34, 42] or existing pose estimators [4, 64] are used without considering the potential effects. Here we conduct a review on key aspects of pose extraction to find a good practice.

In general, 2D poses are of better quality compared to 3D poses, as shown in Figure 1. We adopt 2D Top-Down pose estimators [39, 53, 61] for pose extraction. Compared to its 2D Bottom-Up counterparts [5, 8, 38], Top-Down methods obtain superior performance on standard benchmarks such as COCO-keypoints [31]. In most cases, we feed proposals predicted by a human detector to the Top-Down pose estimators, which is sufficient enough to generate 2D poses of good quality for action recognition. When only a few persons are of interest out of dozens of candidates [2], some priors are essential for skeleton-based action recognition to achieve good performance, *e.g.*, knowing the interested person locations at the first frame of the video. In terms of the storage of estimated heatmaps, they are often stored as coordinate-triplets (x, y, c) in previous literature, where c marks the maximum score of the heatmap and (x, y) is the corresponding coordinate of c. In experiments, we find that coordinate-triplets (x, y, c) help save the majority of storage space at the cost of little performance drop. The detailed ablation study is included in Appendix Sec 4.1.

3.2. From 2D Poses to 3D Heatmap Volumes

After 2D poses are extracted from video frames, to feed into PoseConv3D, we reformulate them into a 3D heatmap volume. Formally, we represent a 2D pose as a heatmap of size $K \times H \times W$, where K is the number of joints, H and W are the height and width of the frame. We can directly use the heatmap produced by the Top-Down pose estimator as the target heatmap, which should be zero-padded to match the original frame given the corresponding bounding box. In case we have only coordinate-triplets (x_k, y_k, c_k) of skeleton joints, we can obtain a joint heatmap $\boldsymbol{J}$ by composing K gaussian maps centered at every joint:

$$\boldsymbol{J}_{kij} = e^{-\frac{(i-x_k)^2+(j-y_k)^2}{2*\sigma^2}} * c_k, \tag{1}$$

σ controls the variance of gaussian maps, and (x_k, y_k) and c_k are respectively the location and confidence score of the k-th joint. We can also create a limb heatmap $\boldsymbol{L}$:

$$\boldsymbol{L}_{kij} = e^{-\frac{\mathcal{D}((i,j),seg[a_k,b_k])^2}{2*\sigma^2}} * \min(c_{a_k}, c_{b_k}). \tag{2}$$

The k_{th} limb is between two joints a_k and b_k. The function $\mathcal{D}$ calculates the distance from the point (i, j) to the segment $[(x_{a_k}, y_{a_k}), (x_{b_k}, y_{b_k})]$. It is worth noting that although the above process assumes a single person in every frame, we can easily extend it to the multi-person case, where we directly accumulate the k-th gaussian maps of all persons without enlarging the heatmap. Finally, a 3D heatmap volume is obtained by stacking all heatmaps ($\boldsymbol{J}$ or $\boldsymbol{L}$) along the temporal dimension, which thus has the size of $K \times T \times H \times W$.

In practice, we further apply two techniques to reduce the redundancy of 3D heatmap volumes. (1) **Subjects-Centered Cropping.** Making the heatmap as large as the frame is inefficient, especially when the persons of interest only act in a small region. In such cases, we first find the smallest bounding box that envelops *all* the 2D poses across frames. Then we crop all frames according to the found box and resize them to the target size. Consequently, the size of the 3D heatmap volume can be reduced spatially while all 2D poses and their motion are kept. (2) **Uniform Sampling.** The 3D heatmap volume can also be reduced along the temporal dimension by sampling a subset of frames. Unlike previous works on RGB-based action recognition, where researchers usually sample frames in a short temporal window, such as sampling frames in a 64-frame temporal window as in SlowFast [18], we propose to use a uniform sampling strategy [59] for 3D-CNNs instead. In particular, to sample n frames from a video, we divide the video into n segments of equal length and randomly select one frame

[1] PoseConv3D can also work with 3D skeletons. An example solution is to divide a 3D skeleton (x, y, z) into three 2D skeletons respectively using (x, y), (y, z) and (x, z).

[2] In FineGym, there exists dozens of audience, while only the pose of the athlete matters.

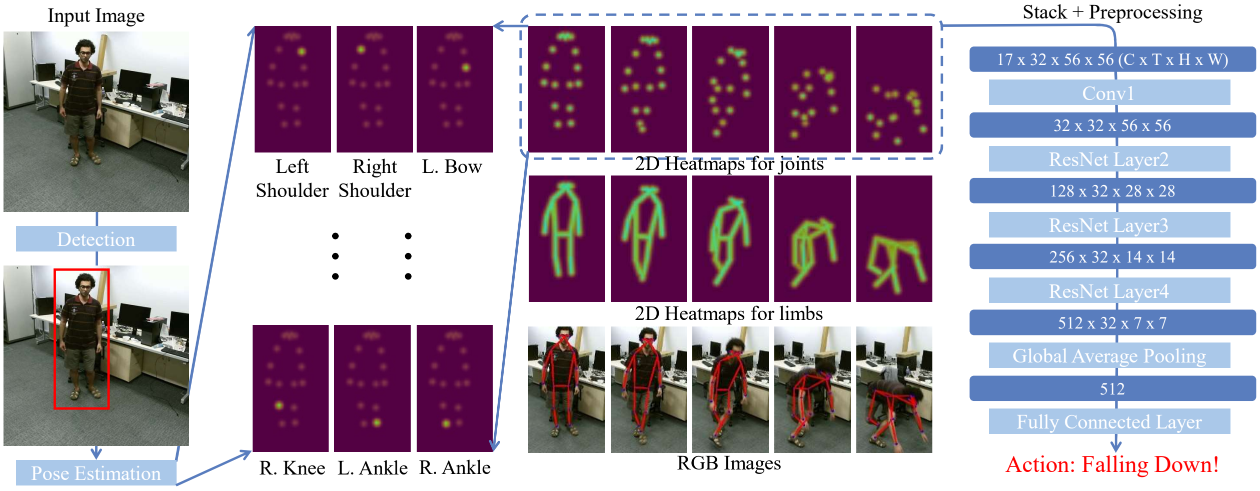

Figure 2. **Our Framework.** For each frame in a video, we first use a two-stage pose estimator (detection + pose estimation) for 2D human pose extraction. Then we stack heatmaps of joints or limbs along the temporal dimension and apply pre-processing to the generated 3D heatmap volumes. Finally, we use a 3D-CNN to classify the 3D heatmap volumes.

Table 2. **Evalution of PoseConv3D variants.** 's' indicates shallow (fewer layers); 'HR' indicates high-resolution (double height & width); 'wd' indicates wider network with double channel size.

Backbone	Variant	NTU60-XSub	FLOPs	Params
SlowOnly	-	93.7	15.9G	2.0M
SlowOnly	HR	93.6	73.0G	8.0M
SlowOnly	wd	93.7	54.9G	7.9M
C3D	-	93.0	25.2G	6.9M
C3D	s	92.9	16.8G	3.4M
X3D	-	92.6	1.1G	531K
X3D	s	92.3	0.6G	241K

from each segment. The uniform sampling strategy is better at maintaining the global dynamics of the video. Our empirical studies show that the uniform sampling strategy is significantly beneficial for skeleton-based action recognition. More illustration about generating 3D heatmap volumes is provided in Appendix Sec 2.

3.3. 3D-CNN for Skeleton-based Action Recognition

For skeleton-based action recognition, GCN has long been the mainstream backbone. In contrast, 3D-CNN, an effective network structure commonly used in RGB-based action recognition [6, 18, 20], is less explored in this direction. To demonstrate the power of 3D-CNN in capturing spatiotemporal dynamics of skeleton sequences, we design two families of 3D-CNNs, namely **PoseConv3D** for the *Pose* modality and **RGBPose-Conv3D** for the *RGB+Pose* dual-modality.

PoseConv3D. PoseConv3D focuses on the modality of human skeletons, which takes 3D heatmap volumes as input and can be instantiated with various 3D-CNN backbones. Two modifications are needed to adapt 3D-CNNs to skeleton-based action recognition: (1) down-sampling operations in early stages are removed from the 3D-CNN since the spatial resolution of 3D heatmap volumes does not need to be as large as RGB clips (4× smaller in our setting); (2) a shallower (fewer layers) and thinner (fewer channels) network is sufficient to model spatiotemporal dynamics of human skeleton sequences since 3D heatmap volumes are already mid-level features for action recognition. Based on these principles, we adapt three popular 3D-CNNs: C3D [54], SlowOnly [18], and X3D [17], to skeleton-based action recognition (Appendix Table 11 demonstrates the architectures of the three backbones as well as their variants). The different variants of adapted 3D-CNNs are evaluated on the NTURGB+D-XSub benchmark (Table 2). Adopting a lightweight version of 3D-CNNs can significantly reduce the computational complexity at the cost of a slight recognition performance drop ($\leq 0.3\%$ for all 3D backbones). In experiments, we use SlowOnly as the default backbone, considering its simplicity (directly inflated from ResNet) and good recognition performance. PoseConv3D can outperform representative GCN / 2D-CNN counterparts across various benchmarks, both in accuracy and efficiency. More importantly, the interoperability between PoseConv3D and popular networks for RGB-based action recognition makes it easy to involve human skeletons in multi-modality fusion.

RGBPose-Conv3D. To show the interoperability of PoseConv3D, we propose RGBPose-Conv3D for the early fusion of human skeletons and RGB frames. It is a two-stream 3D-CNN with two pathways that respectively process RGB modality and Pose modality. While a detailed instantiation of RGBPose-Conv3D is included in Appendix Sec 3.2, the architecture of RGBPose-Conv3D follows several principles in general: (1) the two pathways are asymmetrical due to the different characteristics of the two modalities:

Compared to the RGB pathway, the pose pathway has a smaller channel width, a smaller depth, as well as a smaller input spatial resolution. (2) Inspired by SlowFast [18], bidirectional lateral connections between the two pathways are added to promote early-stage feature fusion between two modalities. To avoid overfitting, RGBPose-Conv3D is trained with two individual cross-entropy losses respectively for each pathway. In experiments, we find that early-stage feature fusion, achieved by lateral connections, leads to consistent improvement compared to late-fusion only.

4. Experiments

4.1. Dataset Preparation

We use six datasets in our experiments: FineGYM [43], NTURGB+D [34, 42], Kinetics400 [6, 64], UCF101 [51], HMDB51 [26] and Volleyball [22]. Unless otherwise specified, we use the Top-Down approach for pose extraction: the detector is Faster-RCNN [40] with the ResNet50 backbone, the pose estimator is HRNet [53] pre-trained on COCO-keypoint [31]. For all datasets except FineGYM, 2D poses are obtained by directly applying Top-Down pose estimators to RGB inputs. We report the **Mean Top-1** accuracy for FineGYM and **Top-1** accuracy for other datasets. We adopt the 3D ConvNets implemented in MMAction2 [11] in experiments.

FineGYM. FineGYM is a fine-grained action recognition dataset with 29K videos of 99 fine-grained gymnastic action classes. During pose extraction, we compare three different kinds of person bounding boxes: 1. Person bounding boxes predicted by the detector (**Detection**); 2. GT bounding boxes for the athlete in the first frame, tracking boxes for the rest frames (**Tracking**). 3. GT bounding boxes for the athlete in all frames (**GT**). In experiments, we use human poses extracted with the third kind of bounding boxes unless otherwise noted.

NTURGB+D. NTURGB+D is a large-scale human action recognition dataset collected in the lab. It has two versions, namely NTU-60 and NTU-120 (a superset of NTU-60): NTU-60 contains 57K videos of 60 human actions, while NTU-120 contains 114K videos of 120 human actions. The datasets are split in three ways: Cross-subject (**X-Sub**), Cross-view (**X-View**, for NTU-60), Cross-setup (**X-Set**, for NTU-120), for which action subjects, camera views, camera setups are different in training and validation. The 3D skeletons collected by sensors are available for this dataset. Unless otherwise specified, we conduct experiments on the **X-sub** splits for NTU-60 and NTU-120.

Kinetics400, UCF101, and HMDB51. The three datasets are general action recognition datasets collected from the web. Kinetics400 is a large-scale video dataset with 300K videos from 400 action classes. UCF101 and HMDB51 are smaller, contain 13K videos from 101 classes and 6.7K videos from 51 classes, respectively. We conduct experiments using 2D-pose annotations extracted with our Top-Down pipeline.

Volleyball. Volleyball is a group activity recognition dataset with 4830 videos of 8 group activity classes. Each frame contains approximately 12 persons, while only the center frame has annotations for GT person boxes. We use tracking boxes from [41] for pose extraction.

4.2. Good properties of PoseConv3D

To elaborate on the good properties of 3D convolutional networks over graph networks, we compare Pose-SlowOnly with MS-G3D [36], a representative GCN-based approach in multiple dimensions. Two models take exactly the **same** input (coordinate-triplets for GCN, heatmaps generated from coordinate-triplets for PoseConv3D).

Performance & Efficiency. In performance comparison between PoseConv3D and GCN, we adopt the input shape $48\times56\times56$ for PoseConv3D. Table 3 shows that under such configuration, PoseConv3D is lighter than the GCN counterpart, both in the number of parameters and FLOPs. Though being light-weighted, PoseConv3D achieves competitive performance on different datasets. The 1-clip testing result is better than or comparable with a state-of-the-art GCN while requiring much less computation. With 10-clip testing, PoseConv3D consistently outperforms the state-of-the-art GCN. Only PoseConv3D can take advantage of multi-view testing since it subsamples the entire heatmap volumes to form each input. Besides, PoseConv3D uses the same architecture and hyperparameters for different datasets, while GCN relies on heavy tuning of architectures and hyperparameters on different datasets [36].

Robustness. To test the robustness of both models, we can drop a proportion of keypoints in the input and see how such perturbation will affect the final accuracy. Since limb keypoints[3] are more critical for gymnastics than the torso or face keypoints, we test both models by randomly dropping one limb keypoint in each frame with probability p. In Table 4, we see that PoseConv3D is highly robust to input perturbations: dropping one limb keypoint per frame leads to a moderate drop (less than 1%) in Mean-Top1, while for GCN, it's 14.3%. Someone would argue that we can train GCN with the noisy input, similar to the dropout operation [52]. However, even under this setting, the Mean-Top1 accuracy of GCN still drops by 1.4% for the case $p = 1$. Besides, with robust training, there will be an additional 1.1% drop for the case $p = 0$. The experiment results show that PoseConv3D significantly outperforms GCN in terms of robustness for pose recognition.

Generalization. To compare the generalization of GCN and 3D-CNN, we design a cross-model check on FineGYM.

[3] There are eight limb keypoints: bow, wrist, knee, ankle (left/right).

Table 3. **PoseConv3D *v.s.* GCN.** We compare the performance of PoseConv3D and GCN on several datasets. For PoseConv3D, we report the results of 1/10-clip testing. We exclude parameters and FLOPs of the FC layer, since it depends on the number of classes.

	MS-G3D			Pose-SlowOnly			
Dataset	Acc	Params	FLOPs	1-clip	10-clip	Params	FLOPs
FineGYM	92.0	2.8M	24.7G	**92.4**	**93.2**	**2.0M**	**15.9G**
NTU-60	91.9	2.8M	16.7G	**93.1**	**93.7**		
NTU-120	84.8	2.8M	16.7G	**85.1**	**86.0**		
Kinetics400	**44.9**	2.8M	17.5G	44.8	**46.0**		

Table 4. **Recognition performance w. different dropping KP probabilities.** PoseConv3D is more robust to input perturbations.

Method / p	0	1/8	1/4	1/2	1
MS-G3D	92.0	91.0	90.2	86.5	77.7
+ robust training	90.9	91.0	91.0	91.0	90.6
Pose-SlowOnly	**92.4**	**92.4**	**92.3**	**92.1**	**91.5**

Table 5. **Train/Test w. different pose annotations.** PoseConv3D shows great generalization capability in the cross-PoseAnno setting (LQ for low-quality; HQ for high-quality).

	Train → Test		
	HQ → LQ	LQ → HQ	LQ → LQ
MS-G3D	79.3	87.9	89.0
PoseConv3D	**86.5**	**91.6**	**90.7**

(a) Train/Test w. Pose from different estimators.

	Train → Test		
	HQ → LQ	LQ → HQ	LQ → LQ
MS-G3D	78.5	89.1	82.9
PoseConv3D	**82.1**	**90.6**	**85.4**

(b) Train/Test w. Pose extracted with different boxes.

Specifically, we use two models, *i.e.*, HRNet (Higher-Quality, or HQ for short) and MobileNet (Lower-Quality, LQ) for pose estimation and train two PoseConv3D on top, respectively. During testing, we feed LQ input into the model trained with HQ one and vice versa. From Table 5a, we see that the accuracy drops less when using lower-quality poses for both training & testing with PoseConv3D compared to GCN. Similarly, we can also vary the source of person boxes, using either **GT** boxes (HQ) or **tracking** results (LQ) for training and testing. The results are shown in Table 5b. The performance drop of PoseConv3D is also much smaller than GCN.

Scalability. The computation of GCN scales linearly with the increasing number of persons in the video, making it less efficient for group activity recognition. We use an experiment on the Volleyball dataset [22] to prove that. Each video in the dataset contains 13 persons and 20 frames. For GCN, the corresponding input shape will be 13×20×17×3, **13** times larger than the input for one person. Under such configuration, the number of parameters and FLOPs for GCN is 2.8M and 7.2G (13×). For PoseConv3D, we can use one **single** heatmap volume (with shape 17×12×56×56) to represent all 13 persons[4]. The base channel-width of Pose-SlowOnly is set to 16, leading to only 0.52M parameters and 1.6 GFLOPs. Despite the much smaller parameters and FLOPs, PoseConv3D achieves 91.3% Top-1 accuracy on Volleyball-validation, 2.1% higher than the GCN-based approach.

4.3. Multi-Modality Fusion with RGBPose-Conv3D

The 3D-CNN architecture of PoseConv3D makes it more flexible to fuse pose with other modalities via some early fusion strategies. For example, in *RGBPose*-Conv3D, lateral connections between the *RGB*-pathway and *Pose*-pathway are exploited for cross-modality feature fusion in the early stage. In practice, we first train two models for RGB and Pose modalities separately and use them to initialize the *RGBPose*-Conv3D. We continue to finetune the network for several epochs to train the lateral connections. The final prediction is achieved by late fusing the prediction scores from both pathways. *RGBPose*-Conv3D can achieve better fusing results with **early+late** fusion.

Table 6. **The design of RGBPose-Conv3D.** Bi-directional lateral connections outperform uni-directional ones in the early stage feature fusion.

	late fusion	RGB → Pose	Pose → RGB	RGB ↔ Pose
1-clip	92.6	93.0	93.4	**93.6**
10-clip	93.4	93.7	93.8	**94.1**

Table 7. **The universality of RGBPose-Conv3D.** The **early+late** fusion strategy works both on RGB-dominant NTU-60 and Pose-dominant FineGYM.

	RGB	Pose	late fusion	early+late fusion
FineGYM	87.2 / 88.5	91.0 / 92.0	92.6 / 93.4	**93.6 / 94.1**
NTU-60	94.1 / 94.9	92.8 / 93.2	95.5 / 96.0	**96.2 / 96.5**

We first compare uni-directional lateral connections and bi-directional lateral connections in Table 6. The result shows that bi-directional feature fusion is better than uni-directional ones for RGB and Pose. With bi-directional feature fusion in the early stage, the **early+late** fusion with 1-clip testing can outperform the **late** fusion with 10-clip testing. Besides, *RGBPose*-Conv3D also works in situations when the importance of two modalities is different. The pose modality is more important in FineGYM and vice versa in NTU-60. Yet we observe performance improvement by **early+late** fusion on both of them in Table 7. We demonstrate the detailed instantiation of *RGBPose*-Conv3D we used in Appendix Sec 2.

[4] In experiments, we find that using a single heatmap volume to represent all people is the best practice (compared to using one heatmap volume for each person). Please refer to Appendix Sec 4.4 for more details.

Table 8. **PoseConv3D is better or comparable to previous state-of-the-arts.** With estimated high-quality 2D skeletons and the great capacity of 3D-CNN to learn spatiotemporal features, PoseConv3D achieves superior performance across **5 out of 6** benchmarks. $\boldsymbol{J}, \boldsymbol{L}$ means using joint/limb-based heatmaps. ++ denotes using the same human skeletons as ours. Numbers with * are reported by [43].

Method	NTU60-XSub	NTU60-XView	NTU120-XSub	NTU120-XSet	Kinetics	FineGYM
ST-GCN [64]	81.5	88.3	70.7	73.2	30.7	25.2*
AS-GCN [30]	86.8	94.2	78.3	79.8	34.8	-
RA-GCN [48]	87.3	93.6	81.1	82.7	-	-
AGCN [45]	88.5	95.1	-	-	36.1	-
DGNN [44]	89.9	96.1	-	-	36.9	-
FGCN [65]	90.2	96.3	85.4	87.4	-	-
Shift-GCN [9]	90.7	96.5	85.9	87.6	-	-
DSTA-Net [46]	91.5	96.4	86.6	89.0	-	-
MS-G3D [36]	91.5	96.2	86.9	88.4	38.0	-
MS-G3D ++	92.2	96.6	**87.2**	89.0	45.1	92.6
PoseConv3D ($\boldsymbol{J}$)	**93.7**	**96.6**	86.0	**89.6**	**46.0**	**93.2**
PoseConv3D ($\boldsymbol{J}+\boldsymbol{L}$)	**94.1**	**97.1**	86.9	**90.3**	**47.7**	**94.3**

Table 9. **Comparison to the state-of-the-art of Multi-Modality Action Recognition.** Strong recognition performance is achieved on multiple benchmarks with multi-modality fusion. R, F, P indicate RGB, Flow, Pose.

(a) Mulit-modality action recognition with *RGBPose-Conv3D*.

Dataset	Previous state-of-the-art	**Ours**
FineGYM-99	87.7 (R) [27]	**95.6** (R + P)
NTU60 (X-Sub / X-View)	95.7 / 98.9 (R + P) [14]	**97.0 / 99.6** (R + P)
NTU120 (X-Sub / X-Set)	90.7 / 92.5 (R + P) [12]	**95.3 / 96.4** (R + P)

(b) Mulit-modality action recognition with *LateFusion*.[5]

Dataset	Previous state-of-the-art	**Ours (Pose)**	**Ours (Fused)**
Kinetics400	84.9 (R) [35]	47.7	**85.5** (R + P)
UCF101	98.6 (R + F) [16]	87.0	**98.8** (R + F + P)
HMDB51	83.8 (R + F) [16]	69.3	**85.0** (R + F + P)

4.4. Comparisons with the state-of-the-art

Skeleton-based Action Recognition. In Table 8, we compare PoseConv3D with prior works for skeleton-based action recognition. Prior works (Table 8 upper) use 3D skeletons collected with Kinect for NTURGB+D, 2D skeletons extracted with OpenPose for Kinetics (details for FineGYM skeleton data are unknown). PoseConv3D adopts 2D skeletons extracted with good practices introduced in Sec 3.1, which have better quality. We instantiate PoseConv3D with the SlowOnly backbone, feed 3D heatmap volumes of shape $48\times56\times56$ as inputs, and report the accuracy obtained by 10-clip testing. For a fair comparison, we also evaluate the state-of-the-art MS-G3D with our 2D human skeletons (*MS-G3D++*): *MS-G3D++* directly takes the extracted coordinate-triplets (x, y, c) as inputs, while *PoseConv3D* takes pseudo heatmaps generated from the coordinate-triplets as inputs. With high quality 2D human skeletons, *MS-G3D++* and *PoseConv3D* both achieve far better performance than previous state-of-the-arts, demonstrating the **importance** of the proposed practices for pose extraction in skeleton-based action recognition. When both take high-quality 2D poses as inputs, PoseConv3D outperforms the state-of-the-art MS-G3D across **5 of 6** benchmarks, showing its great spatiotemporal feature learning capability. PoseConv3D achieves by far the best results on **3 of 4** NTURGB+D benchmarks. On Kinetics, PoseConv3D surpasses MS-G3D++ by a noticeable margin, significantly outperforming all previous methods. Except for the baseline reported in [43], no work aims at skeleton-based action recognition on FineGYM before, while our work first improves the performance to a decent level.

Multi-modality Fusion. As a powerful representation itself, skeletons are also complementary to other modalities, like RGB appearance. With multi-modality fusion (*RGBPose-Conv3D* or *LateFusion*), we achieve state-of-the-art results across **8** different video recognition benchmarks. We apply the proposed *RGBPose-Conv3D* to FineGYM and 4 NTURGB+D benchmarks, using R50 as the backbone; 16, 48 as the temporal length for *RGB*/*Pose*-Pathway. Table 9a shows that our **early+late** fusion achieves excellent performance across various benchmarks. We also try to fuse the predictions of PoseConv3D directly with other modalities with *LateFusion*. Table 9b shows that late fusion with the Pose modality can push the recognition precision to a new level. We achieve the new state-of-the-art on three action recognition benchmarks: Kinetics400, UCF101, and HMDB51. On the challenging Kinetics400 benchmark, fusing with PoseConv3D predictions increases the recognition accuracy by 0.6% beyond the state-of-the-art [35], which is strong evidence for the complementarity of the Pose modality.

[5] For K400, we fuse PoseConv3D Pose predictions (Top1 acc 47.7%)

4.5. Ablation on Heatmap Processing

Subjects-Centered Cropping. Since the sizes and locations of persons can vary a lot in a dataset, focusing on the action subjects is the key to reserving as much information as possible with a relatively small $H \times W$ budget. To validate this, we conduct a pair of experiments on FineGYM with input size 32×56×56, with or without subjects-centered cropping. We find that subjects-centered cropping is helpful in data preprocessing, which improves the Mean-Top1 by 1.0%, from 91.7% to 92.7%.

Uniform Sampling. The input sampled from a small temporal window may not capture the entire dynamic of the human action. To validate this, we conduct experiments on FineGYM and NTU-60. For fixed stride sampling, which samples from a fixed temporal window, we try to sample 32 frames with the temporal stride 2, 3, 4; for uniform sampling, we sample 32 frames uniformly from the entire clip. In testing, we adopt a fixed random seed when sampling frames from each clip to make sure the test results are reproducible. From Figure 3, we see that uniform sampling consistently outperforms sampling with fixed temporal strides. With uniform sampling, 1-clip testing can even achieve better results than fixed stride sampling with 10-clip testing. Note that the video length can vary a lot in NTU-60 and FineGYM. In a more detailed analysis, we find that uniform sampling mainly improves the recognition performance for longer videos in the dataset (Figure 4). Besides, uniform sampling also outperforms fixed stride sampling on RGB-based recognition on the two datasets[6].

Pseudo Heatmaps for Joints and Limbs. GCN approaches for skeleton-based action recognition usually ensemble results of multiple streams (joint, bone, *etc.*) to obtain better recognition performance [45]. The practice is also feasible for PoseConv3D. Based on the coordinates (x, y, c) we saved, we can generate pseudo heatmaps for joints and limbs. In general, we find that both joint heatmaps and limb heatmaps are good inputs for 3D-CNNs. Ensembling the results from joint-PoseConv3D and limb-PoseConv3D (namely PoseConv3D ($\boldsymbol{J} + \boldsymbol{L}$)) can lead to noticeable and consistent performance improvement.

3D Heatmap Volumes *v.s* 2D Heatmap Aggregations. The 3D heatmap volume is a more 'lossless' 2D-pose representation, compared to 2D pseudo images aggregating heatmaps with colorization or temporal convolutions. PoTion [10] and PA3D [63] are not evaluated on popular benchmarks for skeleton-based action recognition, and there are no public implementations. In the preliminary study, we find that the accuracy of PoTion is much inferior ($\leq$ 85%) to GCN or PoseConv3D (all $\geq$ 90%). For an apple-to-apple comparison, we also re-implement PoTion, PA3D (with higher accuracy than reported) and evaluate them on UCF101, HMDB51, and NTURGB+D. PoseConv3D achieves much better recognition results with 3D heatmap volumes, than 2D-CNNs with 2D heatmap aggregations as inputs. With the lightweight X3D, PoseConv3D significantly outperforms 2D-CNNs, with comparable FLOPs and far fewer parameters (Table 10).

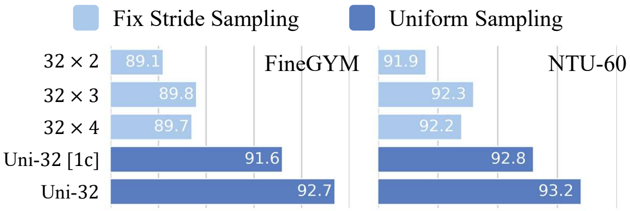

Figure 3. **Uniform Sampling outperforms Fix-Stride Sampling.** All results are for 10-clip testing, except Uni-32[1c], which uses 1-clip testing.

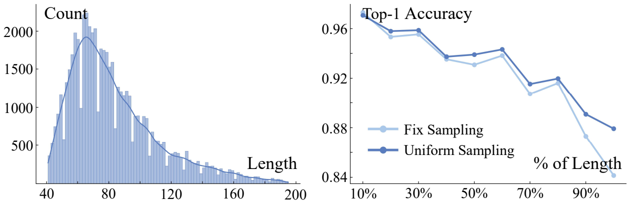

Figure 4. **Uniform Sampling helps in modeling longer videos.** L: The length distribution of NTU60-XSub val videos. R: Uniform Sampling improves the recognition accuracy of longer videos.

Table 10. **An apple-to-apple comparison between 3D heatmap volumes and 2D heatmap aggregations.**

Method	HMDB51	UCF101	NTU60-XSub	FLOPs	Params
PoTion [10]	51.7	67.2	87.8	0.60G	4.75M
PA3D [63]	53.5	69.1	88.6	0.65G	4.81M
Pose-SlowOnly (Ours)	**58.6**	**79.1**	**93.7**	15.9G	**2.0M**
Pose-X3D-s (Ours)	**55.6**	**76.7**	**92.3**	**0.60G**	**0.24M**

5. Conclusion

In this work, we propose **PoseConv3D**: a 3D-CNN-based approach for skeleton-based action recognition, which takes 3D heatmap volumes as input. PoseConv3D resolves the limitations of GCN-based approaches in *robustness*, *interoperability*, and *scalability*. With light-weighted 3D-ConvNets and compact 3D heatmap volumes as input, PoseConv3D outperforms GCN-based approaches in both accuracy and efficiency. Based on PoseConv3D, we achieve state-of-the-art on both skeleton-based and multi-modality-based action recognition across multiple benchmarks.

Acknowledgement. This study is supported by the General Research Funds of Hong Kong (No. 14203518), the RIE2020 Industry Alignment Fund–Industry Collaboration Projects (IAF-ICP) Funding Initiative, and Shanghai Committee of Science and Technology (No. 20DZ1100800).

with VideoSwin [35] RGB predictions. For UCF101 and HMDB51, we fuse PoseConv3D Pose predictions (with K400 PoseConv3D pre-training, 87% Top1 acc on UCF101, 69.3% Top1 acc on HMDB51, details in Appendix Sec 4.2) with OmniSource [16] RGB+Flow predictions.

[6] Please refer to Appendix Sec 4.5 for details and discussions.

References

[1] Sadjad Asghari-Esfeden, Mario Sznaier, and Octavia Camps. Dynamic motion representation for human action recognition. In *WACV*, pages 557–566, 2020. 2

[2] Carlos Caetano, Jessica Sena, François Brémond, Jefersson A Dos Santos, and William Robson Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition. In *AVSS*, pages 1–8. IEEE, 2019. 2

[3] Jinmiao Cai, Nianjuan Jiang, Xiaoguang Han, Kui Jia, and Jiangbo Lu. Jolo-gcn: mining joint-centered light-weight information for skeleton-based action recognition. In *WACV*, pages 2735–2744, 2021. 2

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 3

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 1, 2, 4, 5

[7] 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 *ICCV*, pages 13359–13368, 2021. 2

[8] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 3

[9] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *CVPR*, pages 183–192, 2020. 7

[10] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, pages 7024–7033, 2018. 2, 8

[11] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 5

[12] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *arXiv:2105.08141*, 2021. 7

[13] Srijan Das, Saurav Sharma, Rui Dai, Francois Bremond, and Monique Thonnat. Vpn: Learning video-pose embedding for activities of daily living. In *ECCV*, pages 72–90. Springer, 2020. 2

[14] Mahdi Davoodikakhki and KangKang Yin. Hierarchical action classification with network pruning. In *ISVC*, pages 291–305. Springer, 2020. 7

[15] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, pages 1110–1118, 2015. 1

[16] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *ECCV*, pages 670–688. Springer, 2020. 7, 8

[17] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *CVPR*, pages 203–213, 2020. 2, 4

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019. 2, 3, 4, 5

[19] Pranay Gupta, Anirudh Thatipelli, Aditya Aggarwal, Shubh Maheshwari, Neel Trivedi, Sourav Das, and Ravi Kiran Sarvadevabhatla. Quo vadis, skeleton action recognition? *IJCV*, 129(7):2097–2112, 2021. 2

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, pages 6546–6555, 2018. 4

[21] Alejandro Hernandez Ruiz, Lorenzo Porzi, Samuel Rota Bulò, and Francesc Moreno-Noguer. 3d cnns on distance matrices for human action recognition. In *MM*, pages 1087–1095, 2017. 2

[22] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016. 2, 5, 6

[23] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *TPAMI*, 35(1):221–231, 2012. 2

[24] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *CVPR*, pages 13289–13299, 2020. 2

[25] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 2017. 2

[26] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011. 2, 5

[27] Heeseung Kwon, Manjin Kim, Suha Kwak, and Minsu Cho. Learning self-similarity in space and time as generalized motion for action recognition. *arXiv:2102.07092*, 2021. 7

[28] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. Spatio-temporal graph routing for skeleton-based action recognition. In *AAAI*, volume 33, pages 8561–8568, 2019. 2

[29] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018. 2

[30] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019. 2, 7

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3, 5

[32] Zeyi Lin, Wei Zhang, Xiaoming Deng, Cuixia Ma, and Hongan Wang. Image-based pose representation for action recognition and hand gesture recognition. In *FG*, pages 532–539. IEEE, 2020. 2

[33] Hong Liu, Juanhui Tu, and Mengyuan Liu. Two-stream 3d convolutional neural network for skeleton-based action recognition. *arXiv:1705.08106*, 2017. 2

[34] 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. *TPAMI*, 2019. 2, 3, 5

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv:2106.13230*, 2021. 7, 8

[36] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 2020. 2, 5, 7

[37] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, pages 5137–5146, 2018. 2

[38] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, pages 2277–2287, 2017. 3

[39] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 3

[40] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv:1506.01497*, 2015. 5

[41] Kohei Sendo and Norimichi Ukita. Heatmapping of people involved in group activities. In *ICMVA*, 2019. 5

[42] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, June 2016. 3, 5

[43] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, pages 2616–2625, 2020. 2, 5, 7

[44] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with directed graph neural networks. In *CVPR*, pages 7912–7921, 2019. 7

[45] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, pages 12026–12035, 2019. 2, 7, 8

[46] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Decoupled spatial-temporal attention network for skeleton-based action recognition. *arXiv:2007.03263*, 2020. 7

[47] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv:1406.2199*, 2014. 1

[48] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Richly activated graph convolutional network for robust skeleton-based action recognition. *TSCVT*, 31(5):1915–1925, 2020. 7

[49] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *MM*, pages 1625–1633, 2020. 2

[50] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *arXiv preprint arXiv:2106.15125*, 2021. 2

[51] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv:1212.0402*, 2012. 2, 5

[52] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014. 5

[53] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 3, 5

[54] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1, 2, 4

[55] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *ICCV*, pages 5552–5561, 2019. 2

[56] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 2

[57] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, pages 588–595, 2014. 1

[58] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290–1297. IEEE, 2012. 1

[59] 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 *ECCV*, 2016. 1, 3

[60] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *IJCV*, 129(5):1675–1690, 2021. 1

[61] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 3

[62] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv:2001.08740*, 2020. 1

[63] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *CVPR*, pages 7922–7931, 2019. 2, 8

[64] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, volume 32, 2018. 1, 2, 3, 5, 7

[65] Hao Yang, Dan Yan, Li Zhang, Dong Li, YunDa Sun, ShaoDi You, and Stephen J Maybank. Feedback graph convolutional network for skeleton-based action recognition. *arXiv:2003.07564*, 2020. 7

[66] Dingyuan Zhu, Ziwei Zhang, Peng Cui, and Wenwu Zhu. Robust graph convolutional networks against adversarial attacks. In *KDD*, pages 1399–1407, 2019. 2