SBF: Augmenting Skeleton for Effective Video-based Human Action Recognition

Zhuoxuan Peng[1] Yiyi Ding[2] Yang Lin[1] S.-H. Gary Chan[1]
[1] The Hong Kong University of Science and Technology
[2] The Hong Kong University of Science and Technology (Guangzhou)
zpengac@cse.ust.hk, ydingaz@connect.hkust-gz.edu.cn, {lyangbe, gchan}@cse.ust.hk

Abstract

Many modern video-based human action recognition (HAR) approaches use 2D skeleton as the intermediate representation in their prediction pipelines. Despite overall encouraging results, these approaches still struggle in many common scenes, mainly because the skeleton misses critical action-related information pertaining to the depth of the joints, contour of the human body, and interaction between the human and objects. To address this, we augment skeleton with a novel and effective representation that captures action-related information in the pipeline of HAR without any extra annotation overhead beyond the existing skeleton extraction. The representation, termed Scale-Body-Flow (SBF), consists of three distinct components, namely a map volume given by the scale (and hence depth information) of each joint, a body map outlining the human subject, and a flow map given by pixel-wise optical flow values due to human-object interaction. To predict SBF, we further present SFSNet, a novel ***s****egmentation* ***net****work supervised by the optical* ***f****low and* ***s****keleton. Extensive experiments across different datasets demonstrate that our pipeline based on SBF and SFSNet achieves significantly higher HAR accuracy with similar compactness and efficiency as compared with the state-of-the-art skeleton-only approaches. Code is available at* https://github.com/Shimmer93/EMDUL.

1. Introduction

Video-based human action recognition (HAR) is to classify human actions from a video sequence. It has wide applications in robotics, healthcare, surveillance, and human-computer interaction. The most intuitive approach for HAR is direct end-to-end classification from input frames. Despite its impressive performance, it is too computationally intensive (due to the large model structure and volume of video pixels involved) to be practically deployed.

To reduce computational overhead, HAR based on intermediate representation has been adopted, where an action representation is first extracted from the video frames and then processed by a downstream network. Among various representations studied, the 2D human skeleton has emerged as the most popular due to its compactness and efficiency. The skeleton, represented by a set of joint locations connected via limbs, is used to depict human pose. This focuses exclusively on the human action without irrelevant details such as human appearance and environments. While encouraging results have been reported in the literature, the approach still struggles or fails in some common simple scenarios, mainly because it misses crucial action-related information pertaining to joint depth, body contour, and human-object interaction, as demonstrated in Fig. 2.

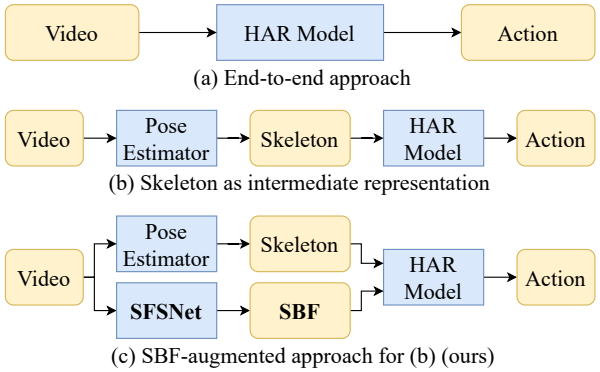
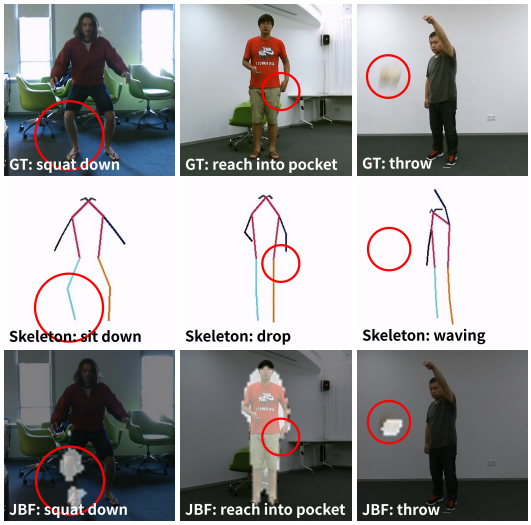


Figure 1. Comparison of video-based HAR pipelines. Our proposed pipeline (c) employs SBF predicted by SFSNet to augment skeleton for effective HAR, addressing the limitations of the skeleton-only approach (b).

To address the issue, several approaches have been proposed to extend skeleton by integrating object information [14, 40, 44]. Despite their improved performance, critical information for HAR such as joint depth and body contour has not been considered. Moreover, object detection demands large amounts of additional annotations beyond skeleton extraction, hampering their deployability.

In this paper, we propose a simple yet effective representation called SBF to augment skeleton for video-based



(a) Joint depth (b) Body contour (c) HOI

Figure 2. Failure cases of HAR based on extracting 2D skeletons from videos, addressed with our SBF representation predicted via SFSNet. Top row: original video frame; middle row: skeleton with prediction error; bottom row: correct prediction using skeleton+SBF, owing to the capturing of the action-related information. Red circles highlight key regions possibly leading to the error. (a) **Joint depth:** The 2D skeleton's flat nature causes the ambiguity of the depth of each joint, e.g., confusing "squatting down" and "sitting down" from a front view. (b) **Body contour:** The contour of human body, an encircling of all body parts, provides richer features than skeleton. E.g., "reaching into a pocket" is mis-predicted as "drop" because the subject's left hand is not shown overlapping with his body in the skeleton. (c) **Human-object interaction (HOI):** The skeleton misses interactions between the subject and objects, e.g., "throwing" misinterpreted as "waving" due to the ignored motion between the subject and the thrown object.

HAR. SBF is composed of three distinct binary maps, each capturing different action-related information:

- *Scale map volume* to capture joint depth: The scale (or size) of a joint on the camera plane provides valuable information on its depth. We propose a map volume for this scale, thereby capturing the depth information.
- *Body map* to capture body contour: We use the body map given by the contour of the entire human subject to cover body parts missing in skeleton.
- *Flow map* to capture human-object interaction: The interaction between the human subject and object provides crucial clues on the human action. We propose using a flow map based on the optical flow values to reflect such interaction.

By augmenting the skeleton with all the above action-related information given by these maps, SBF offers a rich representation to achieve effective HAR.

To extract SBF from video, we present SFSNet, a novel **S**egmentation **Net**work supervised by the optical **F**low and **S**keleton without additional labeling overhead. In SFSNet, the predictions of the scale map volume and body map are trained using the skeleton, while the optical flow predicted by a flow estimator supervises the flow map. By employing an unsupervised flow estimator, SFSNet requires no extra annotations for training beyond the skeleton approach, hence attaining training efficiency and facilitating wide deployment of SBF-based HAR pipeline.

Our contributions are summarized as follows:

- *SBF: A novel representation to augment skeleton for HAR:* We propose a novel representation, SBF, to augment 2D skeleton for video-based HAR. With its rich action-related information, SBF combined with skeleton achieves higher HAR accuracy with similar compactness and efficiency as previous skeleton-only approaches.
- *SFSNet: An effective network to extract SBF:* We introduce SFSNet, which effectively predicts SBF from video utilizing only the point annotations from skeleton and optical flow without any additional annotation overhead.
- *Extensive experiments and validation:* To validate our proposed pipeline based on SBF and its extractor SFSNet, we conduct extensive experiments on several commonly used datasets, namely NTU RGB+D [29], NTU RGB+D 120 [23], UCF101 [33], and HMDB51 [19]. Our results show that SBF-augmented method significantly outperforms state-of-the-art skeleton-only approaches with comparable efficiency. Specifically, it achieves a substantial 2.2% improvement in the X-Sub setting and a 1.6% increase in the X-Set setting on NTU RGB+D 120.

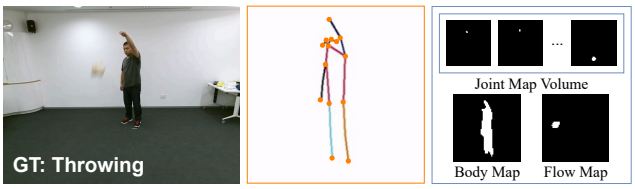
2. Related Works on Video-based HAR

2.1. End-to-end Approaches

The most straightforward HAR approach is to directly classify each input video into a specific human action category. Early works adopt spatiotemporal CNNs [38] or 3D-CNNs [3, 12, 13] as backbone, while transformer-based structure has become increasingly popular in recent years due to its ability to model long-term spatiotemporal dependency [1, 26, 34, 35, 42]. Although current video-based HAR methods achieve state-of-the-art performance on various datasets, their large amount of parameters, computational overhead and need for large amounts of data hamper their deployability in resource-constrained environment.

2.2. Skeleton-based HAR

Compared with end-to-end methods, approaches using skeleton as an intermediate representation are considerably more compact and efficient, since the input size is reduced and noise from background and human appearance is eliminated. Recent years have witnessed the rapid evolution of skeleton-based HAR methods[8, 22, 32, 43, 45, 46, 51]. Notably, PoseConv3D [11] transforms skeletons into



Video frame Skeleton SBF

Figure 3. An example video frame, its extracted skeleton, and our proposed SBF.

heatmaps and employs 3D-CNNs for prediction, achieving state-of-the-art accuracy and efficiency across multiple datasets. As part of the video-based HAR pipeline, these methods typically rely on 2D skeletons extracted from videos as input. However, they suffer from significantly lower performance compared to end-to-end approaches in challenging cases, primarily due to the absence of crucial action-related information such as joint depth, body contour and human-object interaction.

Some approaches incorporate object information into skeleton with additional annotations [14, 40, 44]. Despite promising results, these methods require extra object detection, leading to significant additional annotation overhead and making them labor-intensive in practical applications. In contrast, our proposed SBF not only integrates more information, but also does not require annotations beyond skeleton extraction, enabling broader application scenarios.

2.3. Optical Flow

Optical flow is a dense vector field that describes motion on the camera plane between consecutive frames in a video. Flow estimators can be trained via unsupervised learning [27, 36, 48], requiring no annotation overhead.

Early video-based HAR methods often leverage optical flow to extract motion information [9, 28, 31]. A more recent approach [2] combines skeletons with joint-aligned optical flow patches to enhance understanding of human motion. Despite the improved performance, only optical flow around joints is used in this method, while valuable information from surrounding objects is not considered. Our SBF, in contrast, effectively utilizes optical flow to capture human-object interaction, thereby significantly improving the HAR performance.

3. Scale-Body-Flow Representation

We provide in this section a detailed description of Scale-Body-Flow (SBF) as a novel representation to augment 2D skeleton in the pipeline of video-based HAR. We begin with the definition of SBF in Sec. 3.1, as illustrated in Fig. 3. Next, we introduce the three SBF components, joint map volume (Sec. 3.2), body map (Sec. 3.3) and flow map (Sec. 3.4). Finally, we explain how to combine SBF with skeleton for HAR and introduce our SBF-augmented HAR method, SBFConv3D, in Sec. 3.5.

3.1. Definition

Given two consecutive video frames $\mathcal{I}_{t-1}$ and $\mathcal{I}_t$ of size $H_0 \times W_0$, their SBF is composed of one scale map volume $\mathcal{S}_{t-1} \in \{0,1\}^{H\times W\times J}$, one body map $\mathcal{B}_{t-1} \in \{0,1\}^{H\times W}$ and one flow map $\mathcal{F}_t \in \{0,1\}^{H\times W}$, where J is the number of joints, $H = H_0/4$ and $W = W_0/4$ empirically chosen to ensure compactness. We provide detailed description for these components in the following sections.

3.2. Scale Map Volume

The scale map volume $\mathcal{S}$ consists of J binary maps of size $H \times W$. $\mathcal{S}$ has two variants, the joint scale map volume $\mathcal{S}^J$ and the limb scale map volume $\mathcal{S}^L$.

In $\mathcal{S}^J$, a pixel x on the i-th map is assigned a value of 1 if x belongs to the i-th joint in the skeleton, and 0 otherwise. The area of pixels with value 1 belonging to the joint, or "scale", is negatively correlated with joint depth, as objects closer to the camera appear larger on the image plane. Although predicting depth from video is challenging, $\mathcal{S}^J$ can be estimated using only joint locations, as we will elaborate in Sec. 4.2.

$\mathcal{S}^L$ is derived from $\mathcal{S}^J$. Given the i-th limb connecting joints a_i and b_i, the i-th map in $\mathcal{S}^L$ is defined as

$$\mathcal{S}^L_i(x) = \max(\mathcal{S}^J_{a_i}(x), \mathcal{S}^J_{b_i}(x)). \tag{1}$$

3.3. Body Map

Inspired by early works [4, 39], we introduce body map $\mathcal{B}$, a binary map segmenting the human body from the background. While its ground truth is not easily available, we will demonstrate how to approximate it using only skeleton during training in Sec. 4.2.

3.4. Flow Map

We observe from methods in unsupervised video object segmentation [7, 20, 25] that pixels with optical flow distinct from the background often represent moving human body parts and interacting objects. Hence, human-object interaction can be inferred from optical flow values.

To capture this interaction, we introduce the flow map $\mathcal{F}$, a binary map derived from the optical flow $\mathcal{O} \in \mathbb{R}^{H\times W\times 2}$. $\mathcal{F}$ is defined as follows:

$$\Omega(x) = \|\mathcal{O}(x) - \bar{\mathcal{O}}\|_2, \tag{2}$$

$$\mathcal{F}(x) = \begin{cases} 1, & \text{if } \Omega(x) > \epsilon \max \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Here, $\bar{\mathcal{O}}$ is the average flow value representing the background, and $\epsilon > 0$ is a constant that sets the threshold between static and moving objects. As we will discuss

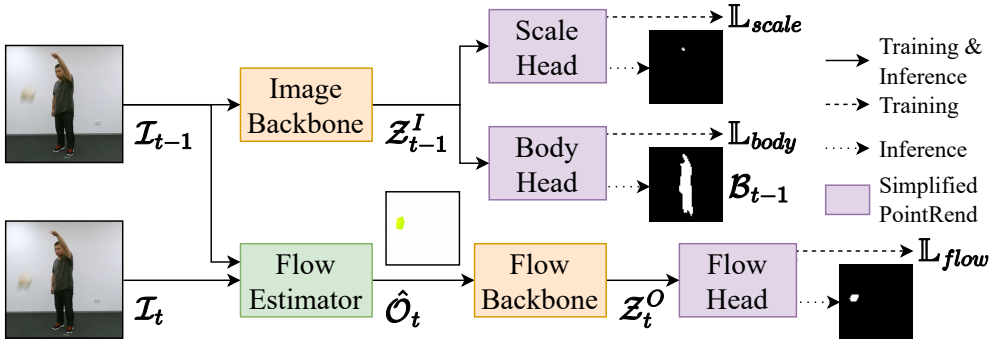


Figure 4. The overall structure of SFSNet. The flow estimator is pretrained via unsupervised learning.

in Sec. 4.2, optical flow can be predicted through unsupervised learning, allowing the estimation of $\mathcal{F}$ without requiring ground-truth optical flow $\mathcal{O}$.

3.5. SBFConv3D: SBF-augmented HAR

Before combining SBF and skeleton, we first transform the skeleton into either a joint heatmap volume $\mathcal{H}^J$ or limb heatmap volume $\mathcal{H}^L$, following PoseConv3D [11]. Both $\mathcal{H}^J$ and $\mathcal{H}^L$ are heatmaps of size $H \times W \times J$, generated using a 2D Gaussian filter with standard deviation σ. Using $\mathcal{H}^J$ and $\mathcal{S}^J$ results in the "joint" variant of our method, while $\mathcal{H}^L$ and $\mathcal{S}^L$ produce the "limb" variant. In the following, we use $\mathcal{H}$ and $\mathcal{S}$ to generically represent either the joint ($\mathcal{H}^J$, $\mathcal{S}^J$) or limb ($\mathcal{H}^L$, $\mathcal{S}^L$) variants.

The same Gaussian filter with standard deviation σ is applied to the three components in SBF to obtain their smoothed versions: $\mathcal{S}'$, $\mathcal{B}'$ and $\mathcal{F}'$. The integrated volume $\mathcal{V}$ is calculated as a weighted sum of $\mathcal{H}$ and $\mathcal{S}'$:

$$\mathcal{V} = \mathcal{H} + \mu\mathcal{S}', \tag{4}$$

where μ is the weighting parameter. Finally, $\mathcal{V}$, $\mathcal{B}'$ and $\mathcal{F}'$ are concatenated to form a tensor of size $H \times W \times (J+2)$, which is fed into a downstream HAR model. Notably, we combine $\mathcal{H}$ and $\mathcal{S}'$ via a weighted sum because $\mathcal{H}_i$ and $\mathcal{S}'_j$ are highly correlated when $i = j$ but largely independent otherwise. $\mathcal{V}$, $\mathcal{B}'$ and $\mathcal{F}'$ are concatenated because they provide complementary information for HAR.

Due to its similarity in size to $\mathcal{H}$, we employ the same 3D CNNs as PoseConv3D for our SBF-augmented HAR method, **SBFConv3D**, to validate the effectiveness of SBF independent of model architecture. Like previous multi-stream skeleton methods, SBFConv3D can integrate the "joint" and "limb" streams using both SBF variants.

4. Extracting SBF with SFSNet

In this section, we introduce SFSNet, an effective segmentation network for SBF prediction requiring no extra annotation overhead beyond the existing skeleton extraction. First, Sec. 4.1 presents a comprehensive overview of SFSNet. We then explain our Simplified PointRend (SPR) module for point-supervised segmentation and our proposed process for point annotation generation in Sec. 4.2. Finally, the training details of SFSNet are provided in Sec. 4.3.

4.1. Overview

The overall structure of our SFSNet is illustrated in Fig. 4. The image backbone encodes an $\mathcal{I}_{t-1}$ into the image feature $\mathcal{Z}^I_{t-1}$, which is then fed into the scale head and body head to predict the joint scale volume $\mathcal{S}_{t-1}$ and the body map $\mathcal{B}_{t-1}$, respectively. Meanwhile, the optical flow $\mathcal{O}_t$ is predicted from $\mathcal{I}_{t-1}$ and $\mathcal{I}_t$ using an off-the-shelf flow estimator pretrained via unsupervised learning. The flow feature $\mathcal{Z}^O_t$ is then encoded from $\mathcal{O}_t$ with the flow backbone and subsequently input into the flow head to predict the flow map $\mathcal{F}_t$. The entire training process is supervised under three segmentation losses, namely $\mathbb{L}_{scale}$, $\mathbb{L}_{body}$ and $\mathbb{L}_{flow}$.

4.2. Simplified PointRend (SPR)

We treat the prediction of each binary map of size $H \times W$ in SBF as a single-class segmentation task. To avoid the labor-intensive annotation process to obtain pixel-level ground truth for segmentation, we introduce our Simplified PointRend (SPR) module, which can utilize sparse point annotations for supervision instead of complete ground truths. It employs a multi-layer perceptron that incorporates positional encoding and adaptive subdivision upsampling from Implicit PointRend (IPR) [6, 18], while excluding elements unnecessary for single-class segmentation. The scale head, body head and flow head are three instances of the SPR module.

All of our SPR modules are trained using point annotations derived from skeleton and optical flow, without requiring any additional annotation overhead beyond the existing skeleton extraction. The generation process for each SPR module begins by constructing positive and negative pixel sets, $\mathcal{P}^{pos}$ and $\mathcal{P}^{neg}$, with labels derived from skeleton or optical flow, as illustrated in Fig. 5. We then randomly

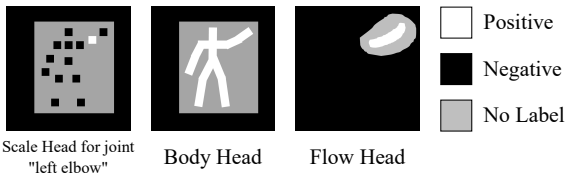


Figure 5. A conceptual example of the "waving" action for our annotation generation method in SPR. In each segmentation task, the white region denotes $\mathcal{P}^{pos}$ (positive labels), the black region represents $\mathcal{P}^{neg}$ (negative labels), and the grey region indicates areas excluded from point sampling.

sample a fixed number of positive and negative points from $\mathcal{P}^{pos}$ and $\mathcal{P}^{neg}$, respectively, as the point annotations used in training. Detailed explanations of the annotation generation process for the scale head, body head and flow head are provided below.

Scale Head: We define the **joint set** of joint i, denoted as $\mathcal{P}_i^{joint}$, to be the set of points containing the 3×3 pixels centered on its exact coordinate. The **background set** $\mathcal{P}^{bg}$ comprises all pixels outside the bounding box of the target individual with a padding of ρ. For each joint i, the positive set $\mathcal{P}_i^{pos}$ equals its joint set $\mathcal{P}_i^{joint}$, and the negative set is the union of all other joint sets and the background set:

$$\mathcal{P}_i^{neg} = \left(\bigcup_{j \neq i} \mathcal{P}_j^{joint} \right) \cup \mathcal{P}^{bg}. \tag{5}$$

We sample N^{pos} positive points from $\mathcal{P}_i^{pos}$ and N^{neg} negative points from $\mathcal{P}_i^{neg}$. The training loss is calculated using the predicted binary classification scores $\mathcal{C}_i^{pos}$ at positive points and $\mathcal{C}_i^{neg}$ at negative points:

$$\mathbb{L}_{scale} = \sum_{i}^{J} (BCE(\mathcal{C}_i^{pos}, \mathbf{I}) + \alpha \cdot BCE(\mathcal{C}_i^{neg}, \mathbf{O})), \tag{6}$$

where α is a balancing constant, and $\mathbf{I}$ and $\mathbf{O}$ respectively denote the identity and zero tensor.

The i-th prediction of the scale head aggregates all pixels highly corresponding to the joint i while excluding other body parts and the background. This ensures that the area of 1's in the i-th prediction roughly covers the entire joint i, thus estimating the i-th scale map $\mathcal{S}_i$ defined in Sec. 3.2.

Body Head: We use Bresenham's line algorithm to connect joints on the image plane according to the graph structure of the skeleton. $\mathcal{P}^{pos}$ comprises all connected line segments and their adjacent pixels, while $\mathcal{P}^{neg}$ equals the background set as noted in the paragraph of Scale Head. A constant of N^{body} points are randomly sampled from either set. The loss function for the body head $\mathbb{L}_{body}$ is a BCE loss directly computed on all sampled points.

The body head prediction aims to distinguish pixels inside and outside the human body, with the boundary representing an approximate human body contour.

Flow Head: Points for the flow head are selected based on the predicted optical flow $\hat{\mathcal{O}}$ from an off-the-shelf unsupervised flow estimator. To generate $\mathcal{P}^{pos}$ and $\mathcal{P}^{neg}$, we first estimate $\Omega(x)$ from Eq. (2) by replacing the ground-truth $\mathcal{O}$ with its prediction $\hat{\mathcal{O}}$ and $\overline{\mathcal{O}}$ with $\overline{\hat{\mathcal{O}}}$:

$$\hat{\Omega}(x) = \|\hat{\mathcal{O}}(x) - \bar{\hat{\mathcal{O}}}\|_2, \tag{7}$$

where $\bar{\hat{\mathcal{O}}}$ is the mean flow vector of $\hat{\mathcal{O}}$. Given constants $\beta, \gamma \in (0, 1)$, $\mathcal{P}^{pos}$ and $\mathcal{P}^{neg}$ are defined as follows:

$$\mathcal{P}^{pos} = \{x : \hat{\Omega}(x) > \beta \max \hat{\Omega}\}, \tag{8}$$

$$\mathcal{P}^{neg} = \{x : \hat{\Omega}(x) < \gamma \max \hat{\Omega}\}. \tag{9}$$

N^{flow} points are randomly sampled from either set and BCE is adopted as the loss function $\mathbb{L}_{flow}$.

The flow head prediction is clearly an estimate of $\mathcal{F}$ defined in Sec. 3.4, with the threshold ϵ lying between β and γ.

4.3. Model Training

We first train the flow backbone and flow head on a small video dataset with $\mathbb{L}_{flow}$ as the total loss. Then the remaining components of SFSNet are trained on a large-scale image dataset with the loss changed to

$$\mathbb{L} = \mathbb{L}_{scale} + \lambda_{body} \mathbb{L}_{body}, \tag{10}$$

where λ_{body} is the weighting parameter.

5. Experimental Evaluation

5.1. Datasets

We evaluate our method on four mainstream action recognition datasets: NTU RGB+D [29], NTU RGB+D 120 [23], UCF101 [33] and HMDB51 [19].

NTU RGB+D (NTU) is a large-scale action recognition dataset collected in lab environment, with 56,880 action sequences performed by 40 subjects categorized into 60 classes. There are two standard benchmarks for this dataset, Cross-Subject (X-Sub) and Cross-View (X-View).

NTU RGB+D 120 (NTU120) is an extended version of NTU RGB+D with 113,945 action sequences performed by 106 subjects categorized into 120 classes. There are two standard evaluation protocols most commonly used, namely Cross-Subject (X-Sub) and Cross-Setup (X-Set).

UCF101 and HMDB51 are two video-based HAR datasets sourced from the Internet. UCF101 (UCF) contains 13,320 videos across 101 action categories, and HMDB51 (HMDB) includes 6,849 videos across 51 actions. In our evaluation, models trained on these datasets are pretrained on Kinetics 400 [17], a large-scale video action dataset.

Table 1. Comparison of model size, computational cost and accuracy (%) with the state-of-the-art HAR methods on NTU and NTU120. All methods shown in the table extract intermediate representations (skeleton, object and SBF) from videos. Number of parameters and FLOPs are calculated for the downstream HAR model alone. Both 1-clip and 10-clip testing are applied to PoseConv3D and SBFConv3D. "J" and "L" indicate the joint and limb variants of SBF defined in Sec. 3.5.

Method	Category	Add. Anno.	Params (M)	FLOPs (G)	Clips	NTU		NTU120	
						X-Sub	X-View	X-Sub	X-Set
ST-GCN [46]	Skeleton	✗	12.3	15.3	10	92.4	98.3	84.7	89.0
AA-GCN [30]	Skeleton	✗	15.0	17.4	10	93.0	98.2	85.5	89.9
MS-G3D [24]	Skeleton	✗	11.8	27.4	10	94.1	98.3	87.4	90.9
CTR-GCN [5]	Skeleton	✗	5.7	7.8	10	93.6	98.4	86.6	90.1
ST-GCN++ [10]	Skeleton	✗	5.6	11.2	10	93.2	98.5	86.4	90.3
BlockGCN [50]	Skeleton	✗	5.2	8.4	10	93.9	98.2	87.3	90.7
ProtoGCN [22]	Skeleton	✗	24.9	36.9	10	94.1	**98.8**	87.5	90.9
IOL [44]	Skl+Obj	✓	-	-	-	90.0	95.7	-	-
MSI [40]	Skl+Obj	✓	-	-	-	91.5	96.5	88.2	89.4
PoseConv3D [11]	Skeleton	✗	4.0	31.6	1	94.0	96.6	86.3	89.9
					10	94.1	97.1	86.9	90.3
SBFConv3D (Ours)	Skl+SBF(J)	✗	2.0	16.2	1	94.6	97.8	89.0	91.5
					10	94.6	98.0	89.1	91.7
	Skl+SBF(J+L)	✗	4.0	32.4	1	94.8	97.9	89.4	92.2
					10	**95.0**	98.1	**89.6**	**92.3**

5.2. Implementation Details

SFSNet: We use HRNet-W32 [41] as the image backbone, LiteHRNet-18 [47] as the flow backbone, and SMURF [36] is selected as the flow estimator. The data augmentation pipeline follows [18] , with each bounding box padded to a square and resized to 256×256. During the process of point annotation generation, we use padding $\rho = 10$, $N_{pos} = 32, N_{neg} = 128, N_{body} = N_{flow} = 256$, $\alpha = 19$, $\beta = 0.8$, and $\gamma = 0.2$. The loss weights are set to $\lambda_{joint} = 0.025$ and $\lambda_{body} = 1$. We first train the flow backbone and flow head in SFSNet on J-HMDB [15] for 100 epochs and then train the other components on COCO17 [21] for 210 epochs with the flow backbone and flow head frozen.

SBFConv3D: We employ SlowOnly-R50 (SO-R50) [13] as the 3D CNN backbone and adhere to the training pipeline in [11] for our SBF-augmented HAR. The length of each SBF sequence is 48, and the crop size is 56×56 during training. μ is set to 0.1, and a Gaussian filter with $\sigma = 0.4$ is applied. The models are trained for 240 epochs in total. In our experiments, we apply 10-clip testing, which aggregates results of 10 distinct samples from a video, unless otherwise specified. When fusing the two streams of SBFConv3D, we average their prediction scores to produce the final results.

5.3. Comparison with State of the Art

In this section, we compare our SBF-augmented HAR (Skl+SBF) with state-of-the-art HAR approaches based on intermediate representations extracted from videos. Selected baselines belong to two categories:

Skeleton methods , including ST-GCN [46], AA-GCN [30], MS-G3D [24], CTR-GCN [5], ST-GCN++ [10], BlockGCN [50], PoTion [8], PA3D [45] and PoseConv3D [11]. For a fair comparison, we evaluate their performance using 2D skeleton extracted from videos using HRNet-w32 [41] with a clip length of 100. BlockGCN is trained by ourselves, while the results for other such methods are sourced from [10]. For methods involving multi-stream fusion, we report results using the most streams.

Skeleton-object fusion methods (Skl+Obj) which integrate object information into 2D skeleton using additional annotations, including IOL [44], MSI [40], and SKP [14]. We present the best results reported in their papers.

As shown in Tab. 1, our SBFConv3D with 1-clip testing already outperforms all other methods in 3 out of 4 settings on NTU and NTU120. Notably, SBFConv3D's performance even exceeds skeleton-object fusion methods, which rely on additional annotations, across all settings. With 10-clip testing, SBFConv3D achieves a 2.2% accuracy increase on NTU120 X-Sub and a 1.6% increase on NTU120 X-Set compared to skeleton methods. Results in Tab. 2 also demonstrate that our SBFConv3D attains higher performance than skeleton methods on UCF and HMDB and even surpasses SKP on HMDB, indicating its effectiveness in more challenging scenarios. Even in the NTU X-view setting, where SBFConv3D shows lower accuracy than state-of-the-art skeleton approaches, it still substantially outper-

Table 2. Comparison of accuracy (%) with the state-of-the-art HAR methods on UCF and HMDB.

Method	Category	UCF	HMDB
PoTion [8]	Skeleton	65.2	43.7
PA3D [45]	Skeleton	-	55.3
SKP [14]	Skl+Obj	**87.8**	70.9
PoseConv3D [11]	Skeleton	87.0	69.7
SBFConv3D	Skl+SBF	87.5	**71.2**

Table 3. Comparison of efficiency of entire video-based HAR pipelines on NTU X-Sub. Number of parameters and FLOPs are calculated for both the skeleton/SBF extraction network and the downstream HAR model. 1-clip testing is used here. E. V. stands for "extractor variant".

Method	Category	E. V.	Params	FLOPs	Acc (%)
CTR-GCN [5]	Skeleton	Base	36.3M	1.1T	93.5
Proto-GCN [22]	Skeleton	Large	90.6M	3.6T	94.2
SBFConv3D	Skl+SBF	Small	**25.7M**	**617G**	93.8
		Base	67.4M	3.5T	**94.8**

Table 4. Comparison of accuracy (%) on videos captured from different viewing angles on NTU120 X-Sub.

Method	Category	Front	Side	45°
PoseConv3D [11]	Skeleton	87.4	83.5	86.9
SBFConv3D	Skl+SBF	**90.1**	**87.3**	**89.9**
Δ	-	+2.7	+3.8	+3.0

Table 5. Comparison of accuracy (%) on actions with different difficulty levels on NTU120 X-Sub.

Method	Category	Hard	Medium	Easy
PoseConv3D [11]	Skeleton	52.2	82.2	96.5
SBFConv3D	Skl+SBF	**61.1**	**87.1**	**97.4**
Δ	-	+8.9	+4.9	+0.9

forms PoseConv3D, which utilizes the same model structure and training pipeline, underscoring that SBF effectively complements the skeleton with rich action representation.

Beyond performance, we assess the efficiency of SFS-Net and SBFConv3D in comparison with state-of-the-art skeleton approaches. Tab. 1 presents the statistics for the downstream HAR model alone. The results reveal that SBFConv3D achieves the highest accuracy with the smallest model size and comparable computational cost in most settings on NTU and NTU120. Tab. 3 further exhibits the parameter count and FLOPs for the entire video-based HAR pipeline, including both the skeleton/SBF extraction and the downstream HAR model, with 1-clip testing. In addition to the Base extractor variants used in Tab. 1, we employ a larger skeleton extractor for ProtoGCN (Large) and a more compact skeleton+SBF extractor for SBFConv3D (Small) to further highlight the efficiency of SBF. The former adopts HRNet-w48 [41] with 4x higher input resolution; the latter uses RTM-Pose [16] for skeleton extraction, LiteHRNet-30 [47] as the image backbone, LiteHRNet-18 [47] as the flow backbone, and downsamples the input size to $1/16$ in the flow estimator. Results show that Small SBFConv3D achieves higher accuracy than Base CTR-GCN with lower computational cost, while Base SBFConv3D outperforms Large ProtoGCN with comparable efficiency.

5.4. Study on Challenging Scenarios

To more clearly demonstrate the advantage of SBF-augmented HAR, we conduct further analysis on the recognition results in challenging scenarios characterized by difficult viewing angles or hard action categories. Only the joint stream is used for SBFConv3D and PoseConv3D.

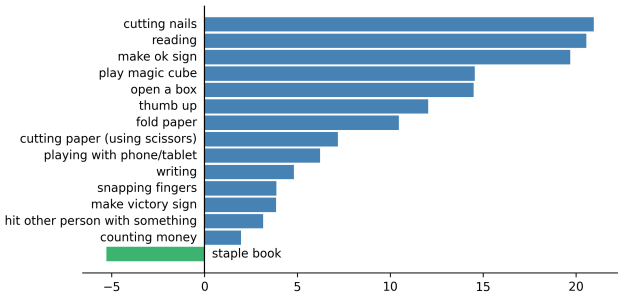


Figure 6. The accuracy difference (%) between our SBFConv3D and PoseConv3D for hard actions on NTU120 X-Sub.

Viewing angles: We divide the testing data in NTU120 X-Sub into three splits according to their viewing angles, namely the Front, Side and 45° view. Intuitively, data from the Side view suffers the most from lack of depth information, as it provides the least amount of spatial differentiation. Results in Tab. 4 confirm this intuition, showing lowest accuracy for the Side view. Our SBFConv3D attains greater performance improvement on the Side view (+3.8%) than on Front (+2.7%) and 45° (+3.0%) views, which demonstrates that the joint scale map volume effectively provides depth information to enhance the action representation.

Action Categories: Following [49], we classify the action categories in NTU120 into three difficulty levels indicated by the action-specific accuracy of PoseConv3D. Actions with accuracy above 90% are considered as Easy, those within 70-90% as Medium, and those below 70% as Hard. As demonstrated in Tab. 5, our SBFConv3D significantly boosts the accuracy for Hard (+8.9%) and Medium (+4.9%)

Table 6. Ablation Study on SBF components on NTU X-Sub.

Skeleton	$\mathcal{S}$	$\mathcal{B}$	$\mathcal{F}$	Acc (%)
✓				93.5
	✓			90.3
		✓		86.7
			✓	87.8
✓	✓			93.6
✓	✓	✓		93.8
✓	✓		✓	94.5
✓	✓	✓	✓	**94.6**

actions. Fig. 6 lists all the Hard actions and shows accuracy improvements in 14 out of the 15 actions, including those highly related to body contour (*e.g.* "make OK sign") or human-object interaction (*e.g.* "cutting nails"). This suggests that the body map and flow map in SBF effectively capture the relevant information, substantially enhancing HAR performance in challenging scenarios.

5.5. Ablation Studies and Qualitative Analysis

This section presents ablation studies and qualitative analysis using the joint stream of SBFConv3D and PoseConv3D.

Effects of SBF Components: We validate the effectiveness of all three components of SBF, scale map volume $\mathcal{S}$, body map $\mathcal{B}$ and flow map $\mathcal{F}$ on NTU X-Sub. Results in Tab. 6 demonstrate that all components collectively enhance the performance of our SBF-augmented HAR. Notably, the flow map contributes the most, possibly because NTU includes various types of objects.

Comparison with PoseConv3D with different 3D CNNs: In addition to SO-R50, we also compare the performance of our SBFConv3D with PoseConv3D using two other 3D CNNs, namely C3D-s [37] and X3D-s [12], on NTU X-Sub. As shown in Tab. 7, SBFConv3D outperforms PoseConv3D across all three model types, with only a marginal increase in model size and computational cost. This compelling evidence proves that SBF improves HAR accuracy with comparable efficiency regardless of the model structure.

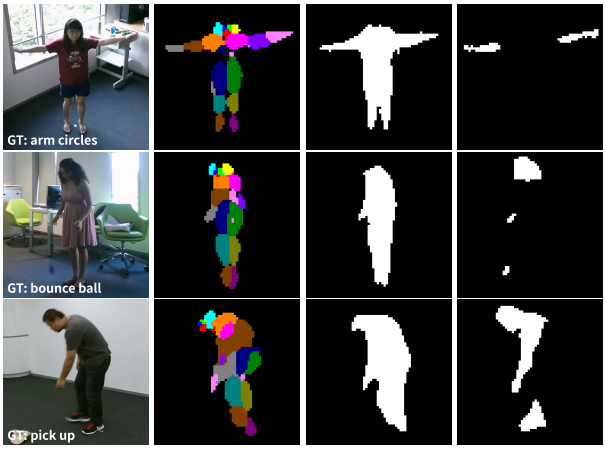
Visualization of Predicted SBF: We present visualization of SBF samples predicted by SFSNet in Fig. 7. The results demonstrate that $\mathcal{S}$ effectively reflects the scale of each joint, $\mathcal{B}$ successfully encircles various body parts, and $\mathcal{F}$ clearly captures the moving body parts and interacting objects. This validates that SBF effectively incorporates the action-related information.

6. Conclusion

Previous works on video-based human action recognition (HAR) are usually based on 2D skeleton as an intermediate representation, which struggles with common scenes due to its lack of action-related information, such as joint depth, body contour and human-object interaction. We propose a novel representation called SBF to augment skeleton for video-based HAR. SBF consists of a scale map volume with the scale (relating to depth) of each joint, a body map outlining the human contour, and a flow map based on optical flow values to capture the interaction between the human and the object. We further present the extractor SFSNet, which predicts SBF with point annotations generated from skeleton and unsupervised optical flow. Without additional annotation overhead beyond existing skeleton extraction, SBF effectively integrates the action-related information with similar compactness and efficiency. Extensive experimental results on commonly used datasets show that our SBF-augmented HAR pipeline outperforms state-of-the-art skeleton methods with similar efficiency. Future work may explore more advanced designs for SFSNet to improve the quality of predicted SBF, and develop more powerful downstream models for SBF-augmented HAR.

Table 7. Comparison with PoseConv3D using different downstream networks on NTU X-Sub.

Method	Network	Acc (%)	Params	FLOPs
PoseConv3D [11]	SO-R50	93.7	2.0M	15.8G
	C3D-s	92.9	3.4M	16.8G
	X3D-s	92.3	241K	0.6G
SBFConv3D	SO-R50	**94.6**	2.0M	16.2G
	C3D-s	**93.4**	3.4M	17.1G
	X3D-s	**92.4**	242K	0.7G



(a) Image (b) $\mathcal{S}$ (c) $\mathcal{B}$ (d) $\mathcal{F}$

Figure 7. Visualization of SBF components predicted by SFSNet on NTU120 X-Sub. Each joint's scale map has a distinct color.

7. Acknowledgment

This work was supported,in part, by RGC-General Research Fund (under grant number 16201625), and Smart Traffic Fund (under grant number STF26EG01) of Hong Kong.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A Video Vision Transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6816–6826, 2021. 2

[2] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2735–2744, 2021. 3

[3] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2

[4] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, 34(15):1799–1807, 2013. Smart Approaches for Human Action Recognition. 3

[5] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13339–13348, 2021. 6, 7, 1, 2

[6] Bowen Cheng, Omkar Parkhi, and Alexander Kirillov. Pointly-Supervised Instance Segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2607–2616, 2022. 4, 1

[7] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Chaewon Park, Donghyeong Kim, and Sangyoun Lee. Treating Motion as Option to Reduce Motion Dependency in Unsupervised Video Object Segmentation. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5129–5138, 2023. 3

[8] Vasileios Choutas, Philippe Weinzaepfel, Jerome Revaud, and Cordelia Schmid. PoTion: Pose MoTion Representation for Action Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7024–7033, 2018. 2, 6, 7

[9] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. MARS: Motion-Augmented RGB Stream for Action Recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7874–7883, 2019. 3

[10] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. PYSKL: Towards Good Practices for Skeleton Action Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7351–7354, New York, NY, USA, 2022. Association for Computing Machinery. 6

[11] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting Skeleton-based Action Recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2959–2968, 2022. 2, 4, 6, 7, 8, 1

[12] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 200–210, 2020. 2, 8

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6201–6210, 2019. 2, 6

[14] Ryo Hachiuma, Fumiaki Sato, and Taiki Sekii. Unified Keypoint-Based Action Recognition Framework via Structured Keypoint Pooling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22962–22971, 2023. 1, 3, 6, 7

[15] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards Understanding Action Recognition. In *2013 IEEE International Conference on Computer Vision*, pages 3192–3199, 2013. 6

[16] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose. 2023. 7

[17] Will Kay, João Carreira, K. Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, T. Back, A. Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *ArXiv*, 2017. 5

[18] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image Segmentation As Rendering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2020. 4, 6

[19] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. *2011 International Conference on Computer Vision*, pages 2556–2563, 2011. 2, 5, 3

[20] Minhyeok Lee, Suhwan Cho, Seunghoon Lee, Chaewon Park, and Sangyoun Lee. Unsupervised Video Object Segmentation via Prototype Memory Network. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5913–5923, 2023. 3

[21] 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 *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 6

[22] Hongda Liu, Yunfan Liu, Min Ren, Hao Wang, Yunlong Wang, and Zhenan Sun. Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29248–29257, 2025. 2, 6, 7

[23] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10):2684–2701, 2020. 2, 5, 1, 3

[24] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 140–149, 2020. 6

[25] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli. See More, Know More: Unsupervised Video Object Segmentation With Co-Attention Siamese Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3618–3627, 2019. 3

[26] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Rethinking Video ViTs: Sparse Video Tubes for Joint Image and Video Learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2214–2224, 2023. 2

[27] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised Deep Learning for Optical Flow Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 3

[28] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J. Black. On the Integration of Optical Flow and Action Recognition. In *Pattern Recognition*, pages 281–297, Cham, 2019. Springer International Publishing. 3

[29] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. 2, 5, 1

[30] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Skeleton-Based Action Recognition With Multi-Stream Adaptive Graph Convolutional Networks. *IEEE Transactions on Image Processing*, 29:9532–9545, 2020. 6

[31] K. Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. *ArXiv*, 2014. 3

[32] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1474–1488, 2023. 2

[33] K. Soomro, Amir Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *ArXiv*, 2012. 2, 5, 3

[34] Siddharth Srivastava and Gaurav Sharma. OmniVec2 - A Novel Transformer based Network for Large Scale Multimodal and Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27412–27424, 2024. 2

[35] Siddharth Srivastava and Gaurav Sharma. OmniVec: Learning robust representations with cross modal sharing. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1225–1237, 2024. 2

[36] Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. SMURF: Self-Teaching Multi-Frame Unsupervised RAFT with Full-Image Warping. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3886–3895, 2021. 3, 6

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 8

[38] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[39] Md. Zia Uddin. Human activity recognition using segmented body part and body joint features with hidden Markov models. *Multimedia Tools and Applications*, 76(11):13585–13614, 2017. 3

[40] Haoran Wang, Baosheng Yu, Jiaqi Li, Linlin Zhang, and Dongyue Chen. Multi-Stream Interaction Networks for Human Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):3050–3060, 2022. 1, 3, 6

[41] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021. 6, 7

[42] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023. 2

[43] Xinghan Wang, Xin Xu, and Yadong Mu. Neural Koopman Pooling: Control-Inspired Temporal Dynamics Encoding for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10597–10607, 2023. 2

[44] Liang Xu, Cuiling Lan, Wenjun Zeng, and Cewu Lu. Skeleton-Based Mutually Assisted Interacted Object Localization and Human Action Recognition. *IEEE Transactions on Multimedia*, 25:4415–4425, 2023. 1, 3, 6

[45] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. PA3D: Pose-Action 3D Machine for Video Recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7914–7923, 2019. 2, 6, 7

[46] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2, 6

[47] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-HRNet: A Lightweight High-Resolution Network. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10435–10445, 2021. 6, 7

[48] Shuai Yuan, Lei Luo, Zhuo Hui, Can Pu, Xiaoyu Xiang, Rakesh Ranjan, and Denis Demandolx. UnSAMFlow: Unsupervised Optical Flow Guided by Segment Anything Model. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19027–19037, 2024. 3

[49] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning Discriminative Representations for Skeleton Based Action Recognition. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10608–10617, 2023. 7

[50] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. BlockGCN: Redefine Topology Awareness for Skeleton-Based Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2049–2058, 2024. 6

[51] Youwei Zhou, Tianyang Xu, Cong Wu, Xiaojun Wu, and Josef Kittler. Adaptive hyper-graph convolution network for skeleton-based human action recognition with virtual connections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12648–12658, 2025. 2