Heatmap Pooling Network for Action Recognition From RGB Videos

Mengyuan Liu, Jinfu Liu, Yongkang Jiang, and Bin He

***Abstract*—Human action recognition (HAR) in videos has garnered widespread attention due to the rich information in RGB videos. Nevertheless, existing methods for extracting deep features from RGB videos face challenges such as information redundancy, susceptibility to noise and high storage costs. To address these issues and fully harness the useful information in videos, we propose a novel heatmap pooling network (HP-Net) for action recognition from videos, which extracts information-rich, robust and concise pooled features of the human body in videos through a feedback pooling module. The extracted pooled features demonstrate obvious performance advantages over the previously obtained pose data and heatmap features from videos. In addition, we design a spatial-motion co-learning module and a text refinement modulation module to integrate the extracted pooled features with other multimodal data, enabling more robust action recognition. Extensive experiments on several benchmarks namely NTU RGB+D 60, NTU RGB+D 120, Toyota-Smarthome and uncrewed aerial vehicles (UAV)-Human consistently verify the effectiveness of our HP-Net, which outperforms the existing human action recognition methods.**

***Index Terms*—Action recognition, heatmap pooling, multimodal data.**

I. Introduction

HUMAN action recognition (HAR) [1], [2], [3], [4] is an active research task in video understanding, as human actions convey essential information like body movements and tendencies, providing valuable insights into understanding individuals in videos [5], [6], [7], [8]. The HAR also holds significant research value and has broad applications in fields like virtual reality, human-robot interaction, and content-based video retrieval. To achieve accurate action recognition, various video modalities have been extensively studied, including RGB frames [3], [4], depth images [9], pose sequences [10], [11], [12], textual descriptions [13], [14], human parsing [15] and point cloud [16]. Specifically, RGB video frames contain abundant information, encompassing body actions, physical appearances and objects interacting with individuals, making them a highly favored modality for action recognition. Consequently, a series of classic action recognition methods [1], [3], [17] based on RGB videos have been developed, leveraging this rich visual information to enhance recognition accuracy and robustness.

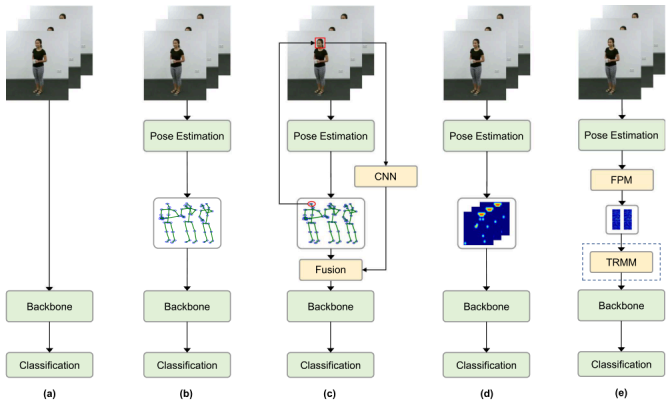


Fig. 1. *Comparison between existing methods and ours.* Existing methods for modeling features from RGB videos have several limitations. (a) RGB-based methods rely on RGB images that are prone to environmental interference and noise. (b) Pose-based methods use pose data with limited information, making it difficult to support fine-grained action recognition. (c) Pose-guided feature aggregation methods require additional modeling and fusion networks, resulting in higher computational cost. (d) Heatmap-based methods often involve redundant data and incur high storage costs. (e) Our proposed HP-Net incorporates a Feedback Pooling Module (FPM) to obtain more efficient heatmap representations by reducing redundancy while retaining relevant information. In addition, the Text Refinement Modulation Module (TRMM) is used to refine the visual features with textual guidance, helping to further improve representation quality.

Existing methods [1], [3], [10], [18] for modeling features from RGB videos can be broadly classified into four main categories. A straightforward approach involves using different visual backbones (e.g., CNNs and ViTs) to directly model video frames [1], [19], [20], as illustrated in Fig. 1(a), ultimately relying on the extracted deep features for action recognition. Although these RGB-based methods that directly utilize raw video frames as input can almost use all the information in videos, they are susceptible to environmental interference and other intrinsic noise in images, which may hinder recognition performance. The second common approach involves using pose estimation to extract 2D/3D pose data from videos and performing action recognition based on the estimated poses [6],

Received 1 July 2025; revised 26 October 2025; accepted 30 November 2025. Date of publication 5 December 2025; date of current version 4 February 2026. This work was supported in part by the National Key Research and Development Program of China under Grant 2024YFB4709800, in part by Guangdong S&T Program under Grant 2024B0101050002, and in part by Shenzhen Innovation in Science and Technology Foundation for the Excellent Youth Scholars under Grant RCYX20231211090248064. Recommended for acceptance by L. Wang. *(Corresponding author: Jinfu Liu.)*

Mengyuan Liu is with the State Key Laboratory of General Artificial Intelligence, Peking University, Shenzhen Graduate School, Shenzhen, Guangdong 518055, China (e-mail: liumengyuan@pku.edu.cn).

Jinfu Liu is with Imaging Department, DJI Technology Company Ltd., Shenzhen, Guangdong 518055, China (e-mail: goch.liu@dji.com).

Yongkang Jiang and Bin He are with Tongji University, Shanghai 200070, China (e-mail: jiangyongkang@tongji.edu.cn; hebin@tongji.edu.cn).

Our code is publicly available at: https://github.com/liujf69/HPNet-Action.

Digital Object Identifier 10.1109/TPAMI.2025.3640697

0162-8828 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

[10], [21], as illustrated in Fig. 1(b). Although these 2D/3D pose-based methods can significantly alleviate environmental interference [10], [11], [12], [22], [23], they omit a wealth of valuable information (e.g., appearance and objects) and are challenging to achieve fine-grained action recognition. The third common approach leverages pose information to guide the extraction of RGB patches, which are modeled using CNNs and then fused with pose features for human action recognition, as illustrated in Fig. 1(c). However, these pose-guided feature aggregation methods require additional RGB patch modeling and multimodal fusion modules, often resulting in higher computational costs.

To address the issues present in the above three approaches, the fourth effective approach is to retain the heatmap features obtained from human pose estimation [24], [25], [26] in videos, as illustrated in Fig. 1(d), and further model them using different visual backbones [3], [18] for action recognition. These heatmap-based methods [3], [18] combine the advantages of the three approaches mentioned above and have been proven to be an effective solution for action recognition. Nevertheless, existing heatmap-based methods typically store heatmaps in image format and model them using image-based visual backbones, which still face challenges including information redundancy, high storage costs, and limitations imposed by downstream modeling backbones tailored to specific input formats. The aforementioned heatmap-based methods naturally lead to a question: *How to better model heatmaps to obtain more concise features, reduce redundancy and support a wider range of visual backbones?* Motivated by the above question, we propose a novel heatmap pooling network (HP-Net) for action recognition from videos, which extracts information-rich, robust and concise heatmap pooled features from videos via a feedback pooling module (FPM), as illustrated in Fig. 1(e).

To better and more intuitively explain the motivation of our work, we begin by reviewing the classical heatmap-based pose estimation methods [24], [25]. These methods use neural networks to model raw video frames in a layer-by-layer manner, which inevitably leads to both information compression and information loss. Throughout this process, heatmaps of various stages and scales are generated, which retain the spatial structure of the human body and represent the probability distributions of human poses using 2D Gaussian functions. To effectively capture human motion in videos, heatmap-based pose estimation methods typically output simplified human pose coordinates, which are then used for various downstream tasks in human motion understanding, such as human action recognition in this paper. In detailed implementation, these pose estimation methods project multi-scale heatmaps into a fixed number of channels through convolutional operations, with each channel corresponding to a specific pose to be estimated. The values within each channel represent the likelihood of the pose occurring at each position, while the final pose locations are then determined using an Argmax operation to select the coordinates with the highest probability. Although these pose estimation methods produce concise human poses to describe human motion in videos, they suffer from three critical information loss issues for action recognition tasks: (1) Feature Reduction: The final human pose coordinates discard appearance features (e.g., color and texture) that could enhance motion understanding, retaining only sparse positional data. (2) Channel Compression: heatmaps with diverse scales and channel counts are forcibly projected to a fixed channel number (e.g., 17 for the COCO dataset), which may lead to the loss of many valuable channel-wise features. (3) Spatial Oversimplification: The Argmax operation retains only the highest-probability coordinate per heatmap, ignoring neighboring regions that may contain valid cues for human motion.

To mitigate information loss, prior heatmap-based action recognition methods [3], [18] directly feed entire heatmaps into downstream visual networks, introducing severe information redundancy, as most heatmap pixels represent irrelevant or noisy coordinate probabilities. In contrast, our proposed heatmap pooling network (HP-Net) rethinks the core principles of heatmap-based pose estimation. By introducing a feedback pooling mechanism to generate heatmap pooled features, our HP-Net strategically aggregates features from multi-stage and multi-scale heatmaps guided by estimated pose positions. Intuitively, the estimated poses inherently localize the most informative regions in heatmaps. Centering pooling operations around these poses maximizes the retention of discriminative local features, while preserving all channel dimensions across heatmaps in different scales. This design in our HP-Net offers two key advantages: (1) Targeted Feature Preservation: Human pose positions, being probability maxima, anchor regions rich in motion-relevant features. Pooling around these pose points focuses on meaningful areas, suppressing noise while retaining critical spatial-contextual details. (2) Multi-Channel Integrity: By avoiding channel compression and leveraging full-channel heatmaps, the HP-Net captures latent features across hierarchical representations, essential for modeling complex human motions. Notably, our HP-Net utilizes a feedback pooling mechanism, where the estimated concise pose coordinates are used to reversely pool the information-redundant heatmaps from different stages and scales, enabling the extraction of full-channel and information-rich heatmap-pooled features. Thus, the HP-Net fundamentally reconciles the trade-off between information loss and redundancy, enabling efficient yet comprehensive motion representation for human motion analysis.

Compared to conventional pose data and image-based heatmap features extracted from videos, our heatmap pooled features offer more useful information, less redundancy and lower storage costs, while being compatible with a wider range of visual backbones, such as graph convolutional networks (GCNs) and Transformers. Although modeling heatmap pooled features through visual networks outperforms approaches that directly process raw heatmaps or pose data, it still falls short of current state-of-the-art (SOTA) multimodal action recognition methods [3], [4], [6], [14]. In essence, the heatmap-pooled features obtained through the feedback pooling mechanism are still unimodal in nature and thus inherently subject to the limitations of single-modality representations. Therefore, the recognition performance only based on the heatmap pooled features is inevitably inferior to that of existing multimodal approaches. We also observe that these leading multimodal

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

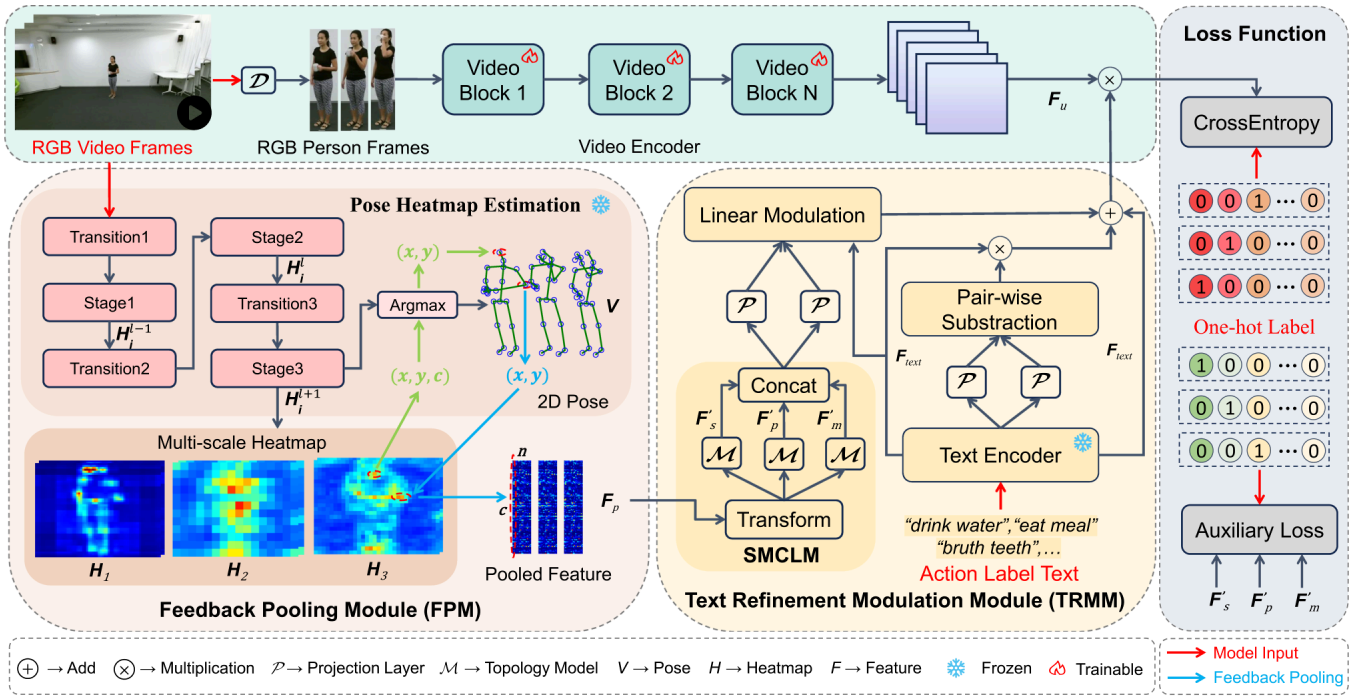


Fig. 2. *Framework of our proposed Heatmap Pooling Network (HP-Net).* We utilize a feedback pooling module (FPM) to extract information-rich, robust and concise heatmap pooled features from videos. The spatial-motion co-learning module (SMCLM) is designed to model spatial and motion dynamics within heatmap pooled features, enabling the generation of spatiotemporal representations that seamlessly integrate with other modalities. The text refinement modulation module (TRMM) aims to enhance and modulate the basic text features of action labels, transforming them into enriched semantic representations that effectively support human action recognition.

methods [3], [4], [6], [14] typically design specialized modules to fuse multimodal data (e.g., human pose, RGB video frames or text descriptions), leveraging cross-modality complementarity to enhance recognition accuracy. To validate the transferability of our heatmap-pooled features and demonstrate their potential to complement and integrate with other modalities, our proposed HP-Net futher integrates the extracted pooled features with other modalities, such as deep features from RGB video frames and text features generated by large language models (LLMs), enabling more robust and accurate action recognition. In the detailed implementation, we propose two modules to enable feature-level multimodal fusion of heatmap pooled features, RGB features and textual features, while also exploring the effectiveness of decision-level fusion through a multi-stream ensemble strategy.

In our heatmap pooling network (HP-Net), we employ a feedback pooling module (FPM) to perform pooling operation on heatmap features using 2D pose data obtained from pose estimation. These pooled features can be effectively modeled by using backbones that accept topology graph inputs (e.g., GCNs), thereby avoiding the use of complex visual backbones designed exclusively for heatmap images (e.g., 3D CNNs). Meanwhile, we designed a text refinement modulation module (TRMM) to adapt the pooled features for refining and modulating text features. These refined and modulated text features are then further fused with video features, demonstrating that heatmap pooled features can be effectively integrated with multimodal features to achieve more robust action recognition. In addition, we designed a spatial-motion co-learning module (SMCLM) to transform heatmap pooled features into spatiotemporal modalities, which are modeled by using different graph convolutional networks (GCNs), further enhancing the recognition performance.

The framework of our heatmap pooling network (HP-Net) is shown in Fig. 2 and our contributions are summarized as follows:

- We propose a novel heatmap pooling network (HP-Net) for action recognition from videos, which extracts concise, robust and transferable heatmap features through a feedback pooling mechanism. The proposed feedback pooling module (FPM) in our HP-Net is orthogonal to the heatmap-based human pose estimation methods, providing a new connection bridge between the fields of human pose estimation and action recognition.
- Our HP-Net further enhances image-based human heatmap features, making them adaptable to a wider range of visual backbone networks. Simultaneously, the extracted heatmap pooled features can be smoothly integrated with other multimodal features, facilitating more accurate and robust human action recognition. Compared to traditionally acquired 2D pose data, the proposed heatmap pooled feature achieves consistent performance gains when integrated with distinct GCN/Transformer architectures.
- Our HP-Net conduct the most extensive evaluation of heatmap-based human action recognition, spanning three major benchmarks with radically different characteristics: controlled laboratory settings (i.e., NTU-60 and NTU-120), home environments (i.e., Toyota-Smarthome) and

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

aerial surveillance [i.e., uncrewed aerial vehicles (UAV)-Human]. The proposed HP-Net achieves state-of-the-art performance in all the above scenarios, providing compelling evidence of its effectiveness and generalizability.

The remainder of this paper is organized as follows. Section II briefly reviews related work. Section III describes the structure of the heatmap pooling network. Section IV reports the experimental results and analysis. Section V discusses the generalizability of our work. Section VI concludes the paper.

II. Related Work

The task of human action recognition from videos has attracted significant attention from researchers due to the availability of rich multimodal data. As a result, a series of mainstream methods have emerged based on different modalities. In this section, we will first discuss the related works in both RGB-based methods and pose-based methods. Following that, we will introduce heatmap-based methods, which are more relevant to the work of this paper. Finally, we will provide an overview of the classic upstream task in human action recognition, namely human pose estimation from RGB videos.

A. RGB-Based Human Action Recognition

RGB-based human action recognition have garnered widespread attention due to the RGB videos contain rich information and are relatively simple to obtain in daily life [27], [28], [29], [30], [31]. Due to the unique advantages of convolutional neural networks (CNNs) in processing euclidean data, most previous works [1], [4], [17], [32] have used CNNs to model the spatiotemporal features of RGB videos. For instance, Carreira et al. [17] introduced the inflated 3D CNN (I3D) that leveraging both RGB and optical flow streams for human action recognition. Feichtenhofer et al. [1] presented the SlowFast networks for video action recognition, which involves a slow pathway to capture spatial semantics and a fast pathway to capture motion of RGB videos. Additionally, some works [4], [32], [33], [34] have fused features from RGB video frames with pose features to achieve more robust human action recognition. Das et al. [32] proposed the VPN, which models spatiotemporal features of RGB video frames using a visual backbone and integrates RGB features with pose features through spatial embedding and an attention network. Yu et al. [4] transformed RGB video into a single-image ST-ROI and proposed the MM-Net, which uses pretrained CNNs to model ST-ROI and fuses it with pose features for video action recognition. Reilly et al. [6] introduce the first Pose Induced Video Transformer (π ViT), which augments the RGB representations learned by video transformers with pose information for action recognition. The aforementioned methods directly use RGB video frames as input to visual networks, making them susceptible to the inherent noise in RGB frames, which reduces the accuracy of human action recognition.

B. Pose-Based Human Action Recognition

Human skeletons/poses have become a commonly used modality in action recognition due to their simplicity and resistance to environmental interference [35], [36], [37]. Generally, human poses can be obtained from videos through sensors [38], [39], [40] or pose estimation [24], [25], [26]. These human poses extracted from videos is a structured natural topology graph, making them well-suited for modeling with graph convolutional networks (GCNs) in pose-based action recognition. For instance, Yan et al. [10] were the first to introduce GCNs for pose-based action recognition, proposing the ST-GCN to capture spatiotemporal features from pose data. Shi et al. [41] proposed a novel two-stream adaptive graph convolutional network (2s-AGCN), which model both the first-order and the second-order information of human pose simultaneously. Chen et al. [11] enhanced the design of GCNs by introducing the channel-wise topology refinement graph convolutional network (CTR-GCN), which efficiently aggregates joint features within each channel. Liu et al. [12] proposed the temporal decoupling graph convolutional network (TD-GCN), which innovatively utilizes a temporal-dependent adjacency matrix to effectively extract high-level features from pose data. Since human pose is merely a set of simple coordinate data with limited information, the aforementioned pose-based methods have drawbacks in fine-grained action recognition.

C. Heatmap-Based Human Action Recognition

Human heatmap as a feature representation in the human pose estimation process [24], [25], can filter environmental interference while containing most information about human action. Some previous works [3], [18] have used heatmaps for video action recognition and achieved significant success. For instance, to eliminate the limitations of cluttered backgrounds and non-action motions for video action recognition, Liu et al. [18] were the first to introduce heatmaps as a human cue for action recognition. To address the sparse nature of heatmaps obtained from the pose estimation process, Liu et al. [18] also developed a spatial rank pooling method to aggregate the evolution of heatmaps into body shape evolution images. In addition, Duan et al. [3] proposed the PoseConv3D, which relies on 3D heatmap volumes as input and uses the pretrained 3D CNNs to effectively model spatiotemporal features. The above methods typically convert heatmaps into image formats as model inputs to facilitate effective modeling with CNNs, which to some extent results in model limitations and information redundancy. Unlike the aforementioned methods that convert heatmaps into image formats, our HP-Net employs feedback pooling to transform redundant heatmaps into topological pooled features, enabling efficient modeling with compact networks that accept topological inputs.

D. Human Pose Estimation From RGB Videos

Human pose estimation (HPE) [43], [44] aims to localize body joints from a single RGB image and has garnered significant attention as an upstream task for various vision applications. The pose data obtained from videos through human pose estimation can be used for pose-based action recognition. Additionally, heatmap-based methods [24], [25] is a key branch of human pose estimation, which produce heatmaps that can be effectively used

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

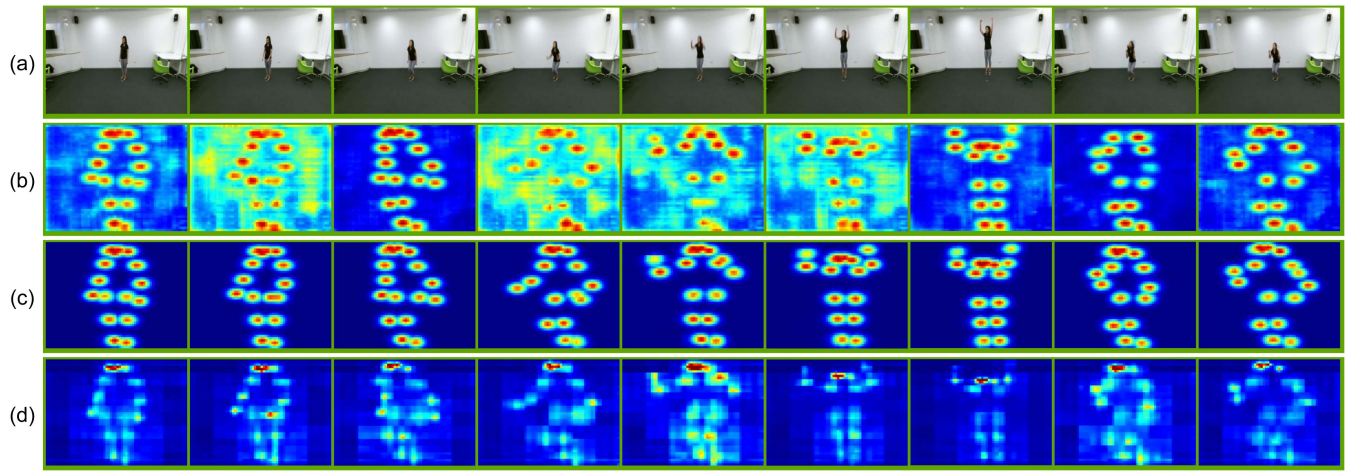


Fig. 3. Visualization of human heatmaps obtained by different human pose estimation models. (a) Visualization of the video frame sequence. (b) Visualization of human heatmaps obtained using ResNet [42]. (c) Visualization of human heatmaps obtained using HR-Net [24]. (d) Visualization of human heatmaps obtained using SimCC [25].

for video action recognition [3], [18]. Some previous works have explored how to effectively estimate human pose from videos based on heatmaps. For instance, Sun et al. [24] proposed the HRNet, which learns reliable high-resolution representations and integrates features of different scales to predict accurate keypoint heatmaps. Li et al. [25] proposed the SimCC, which reinterprets the traditional 2D heatmap-based pose estimation approach as two classification tasks for horizontal and vertical coordinates, achieving better prediction performance. The heatmaps generated by the aforementioned pose estimation methods effectively eliminate environmental interference while preserving critical cues about human motion. Consequently, these heatmaps can be readily applied to the human action recognition and seamlessly integrated with other modalities to enhance recognition performance. In Fig. 3, we visualize the human heatmaps for one action sample using different pose estimation models [24], [25], [42]. In our proposed heatmap pooling network (HP-Net), we employ a feedback pooling mechanism to transform heatmaps obtained from pose estimation into concise and robust pooled features. By effectively integrating these pooled features with RGB video and text features, our HP-Net achieve more accurate video action recognition.

III. Heatmap Pooling Network

In this section, we present our proposed Heatmap Pooling Network (HP-Net) in detail. HP-Net aims to extracts information-rich, robust and concise heatmap pooled features in videos for human action recognition. Meanwhile, the extracted heatmap pooled features are adaptable to different network backbones and can be effortlessly integrated with other modalities, enabling more precise recognition performance. Below we will introduce our HP-Net in five parts. In the first part, we will introduce the detailed implementation of effectively extracting heatmap pooled features from videos using the Feedback Pooling Module (FPM). Then in the second part, we introduce the method of using lightweight networks for topological modeling of heatmap pooled features. Additionally, we provide a detailed explanation of how heatmap pooled features are transferred and integrated with other modalities through the SMCLM and TRMM in the third and fourth parts. Finally, in the fifth section, we introduce the method for obtaining robust classification scores through multimodal fusion and define the loss function of our proposed HP-Net.

A. Feedback Pooling Module

In our proposed Heatmap Pooling Network (HP-Net), we employ a Feedback Pooling Module (FPM) to extract heatmap pooled features from videos. The FPM aims to leverage the compact pose data obtained from pose estimation to acquire robust and information-rich human heatmap pooled features by reflectively pooling the human heatmap containing human motion cues. In terms of implementation details, we use a pose estimation model ϕ to obtain multi-scale human heatmaps based on the following:

$$\begin{aligned} \boldsymbol{H}^t &= \left\{\boldsymbol{H}^t_{s_1}, \boldsymbol{H}^t_{s_2}, \ldots, \boldsymbol{H}^t_{s_n}\right\} \\ &= \phi(\boldsymbol{U}^t, \{s_1, s_2, \ldots, s_n\}), \quad \forall t \in \{1, 2, \ldots, T\}, \end{aligned} \tag{1}$$

where $\boldsymbol{H}^t$ and $\boldsymbol{U}^t$ represent the obtained human heatmap and the original video frame at frame t, respectively. The $\{s_1, s_2, \ldots, s_n\}$ represent parameter sets of different scales. Then, as is common in most heatmap-based pose estimation methods [24], [25], [45], we apply the $argmax$ operation along the channel dimension of the heatmap $\boldsymbol{H} \in \mathbb{R}^{h \times w \times c}$ to obtain the concise human pose coordinates, which is formulated as:

$$\boldsymbol{V}_i = \arg\max_{(x,y)} \boldsymbol{H}_i(x, y), \quad \forall i \in \{1, 2, \ldots, n\}, \tag{2}$$

where $\boldsymbol{V} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_c, y_c)\}$ represents the 2D coordinate set of the human pose. The h, w and c represent the height, width and channels of the heatmap $\boldsymbol{H} \in \mathbb{R}^{h \times w \times c}$ respectively. Finally, we perform a feedback pooling operation on the human heatmaps of different scales using the extracted 2D

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

pose to obtain concise heatmap pooled features. The feedback pooling operation $\mathcal{G}$ is formulated as:

$$\boldsymbol{F}_{\mathrm{p}}(x,y) = \mathcal{G}\left(\boldsymbol{H}_{(x-\frac{R}{2}):(x+\frac{R}{2}),(y-\frac{R}{2}):(y+\frac{R}{2})}\right), \tag{3}$$

where $\boldsymbol{F}_{\mathrm{p}}(x,y)$ represents the heatmap pooled features corresponding to the pose coordinates (x,y) and R represents the size of the pooled region. Typically, we pool the human heatmap using n key pose coordinates, resulting in heatmap pooled features $\boldsymbol{F}_{\mathrm{p}} \in \mathbb{R}^{n\times c}$ corresponding to n human joints. Based on the above formulas, our proposed FPM can extract compact heatmap pooled features $\boldsymbol{F}_{\mathrm{p}}$ related to human action from videos. The extracted heatmap pooled features maintain the simplicity of human pose while significantly reducing the information redundancy in the human heatmaps, preserving rich and effective features related to human action. It is worth emphasizing that in the proposed FPM, the final obtained pose data can be used to reverse-pool heatmaps from different stages and scales, enabling the extraction of diverse human motion cues.

B. Topology Modeling

The heatmap pooled features extracted by the FPM have the same topological structure as human pose, allowing for effective modeling using classical topological models like GCNs, while avoiding the use of heavy CNNs. Meanwhile, many previous works [5], [24], [25] have used GCNs to model the human poses with topological structure, achieving great success in pose-based action recognition. Therefore, to ensure a fair comparison with human poses and to demonstrate the superiority of the heatmap pooled features, our HP-Net employs a lightweight GCN to perform topological modeling of the extracted heatmap pooled features. Here we select GCN as the topological model $\mathcal{M}$ and the normal graph convolution of GCN can be formulated as:

$$\boldsymbol{F}^{l+1} = \sigma\left(\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{F}^{l}\boldsymbol{W}^{l}\right), \tag{4}$$

where $\boldsymbol{F}^{l}$ is the heatmap pooled features at layer l and σ is the activation function. $\boldsymbol{A}$ is the adjacency matrix representing topological connections. $\boldsymbol{D}$ is the degree matrix and $\boldsymbol{W}^{l}$ is the learnable parameter of the l-th layer. In our HP-Net, the topologically modeled heatmap pooled features can be directly used for action classification through a simple classification head or transferred for fusion with other modalities to achieve multimodal-based action recognition. It is worth emphasizing that the heatmap pooled features we propose are orthogonal to the backbone networks and thus can be coupled with various topological models.

C. Spatial-Motion Co-Learning Module

Human action recognition is a task with strong spatiotemporal correlations and it is a helpful approach to addressing this task effectively via generating spatiotemporal features. In our HP-Net, we utilize a Spatial-motion Co-learning Module (SMCLM) to perform spatial and motion modeling on the heatmap pooled features, facilitating deeper integration with other modalities via generating spatiotemporal features. First, we employ a transform module $\mathcal{T}$ to convert the raw heatmap pooled features $\boldsymbol{F}_{\mathrm{p}}$ into spatial features $\boldsymbol{F}_{\mathrm{s}}$ and motion features $\boldsymbol{F}_{\mathrm{m}}$. The SMCLM fully leverages the spatial and motion information inherent in the heatmap pooled features by converting the features into spatial and motion modalities, providing a more comprehensive representation for human action. Below we describes the implementation process of the transform module $\mathcal{T}$ based on [12]. Given two heatmap pooled features $\boldsymbol{F}_{\mathrm{p}}^{i} = \{c_1^i, c_2^i, \ldots, c_n^i\}$ and $\boldsymbol{F}_{\mathrm{p}}^{j} = \{c_1^j, c_2^j, \ldots, c_n^j\}$ with natural topological connections, the spatial features are transformed into $\boldsymbol{F}_{\mathrm{s}}^{ij} = \{c_1^i - c_1^j, c_2^i - c_2^j, \ldots, c_n^i - c_n^j\}$. Given two heatmap pooled features $\boldsymbol{F}_{\mathrm{tp}}^{i} = \{c_{t1}^i, c_{t2}^i, \ldots, c_{tn}^i\}$ and $\boldsymbol{F}_{(\mathrm{t+1})\mathrm{p}}^{i} = \{c_{(t+1)1}^i, c_{(t+1)2}^i, \ldots, c_{(t+1)n}^i\}$ from two consecutive frames t and $t+1$, the motion features are transformed into $\boldsymbol{F}_{\mathrm{m}}^{i} = \boldsymbol{F}_{(\mathrm{t+1})\mathrm{p}}^{i} - \boldsymbol{F}_{\mathrm{tp}}^{i}$. Then, we use three independent topological models to effectively model the heatmap pooled features, spatial features and motion features. These well-modeled features are then integrated through concatenation and passed to the text refinement modulation module (TRMM), which is formulated as:

$$\boldsymbol{F}_{\mathrm{c}} = \mathcal{M}_1(\boldsymbol{F}_{\mathrm{p}}) \oplus \mathcal{M}_2(\boldsymbol{F}_{\mathrm{s}}) \oplus \mathcal{M}_3(\boldsymbol{F}_{\mathrm{m}}), \tag{5}$$

where $\boldsymbol{F}_{\mathrm{c}}$ represents the features obtained after concatenation and $\oplus$ represents the concatenation operation. The symbols $\mathcal{M}_1, \mathcal{M}_2$ and $\mathcal{M}_3$ denote three independent topological models. Meanwhile, to constrain the effective training of the above three independent topology models, we calculate the cross-entropy loss between the output features of these three models and the ground truth action labels.

Algorithm 1: Text Refinement Modulation Module.

Input: Action label text $\in \mathbb{R}^N$, Heatmap pooled features $F_{\mathrm{c}} \in \mathbb{R}^D$

Output: Information-rich action description features $F'_{\mathrm{text}} \in \mathbb{R}^{N\times C'}$

1: **for** action label index $i = 1$ in N **do**
2: $\quad F_{\mathrm{text}_i} = \mathcal{E}_t(\mathrm{text}_i) \in \mathbb{R}^C$
3: **end for**
4: Obtain text features: $F_{\mathrm{text}} \in \mathbb{R}^{N\times C}$
5: Calculate scale factor and shift factor: $\gamma \leftarrow MLP(F_{\mathrm{c}}), \beta \leftarrow MLP(F_{\mathrm{c}})$
6: Modulate text features: $\psi(F_{\mathrm{c}}, F_{\mathrm{text}}) \leftarrow \gamma \times F_{\mathrm{text}} + \beta$
7: Refine and project features: $\eta(F_{\mathrm{c}}) \leftarrow Tanh(MLP(F_{\mathrm{c}}).unsqueeze(-1) - MLP(F_{\mathrm{c}}).unsqueeze(-2)) \circ F_{\mathrm{c}}$
8: Aggregate features: $F'_{\mathrm{text}} \leftarrow F_{\mathrm{text}} + \eta(F_{\mathrm{c}}) + \eta(F_{\mathrm{c}})$
9: **return** $F'_{\mathrm{text}} \in \mathbb{R}^{N\times C'}$

D. Text Refinement Modulation Module

Some previous studies have demonstrated that incorporating text features into human action recognition can significantly enhance performance [5], [13], [14], [22]. However, most previous works directly use the encoded text features of action labels, which suffer from the issue of limited information since action labels are inherently simple words. To address this issue, in

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

our proposed Text Refinement Modulation Module (TRMM), we use the concatenated pooled features $\boldsymbol{F}_{\mathrm{c}}$ output from the SMCLM to refine and modulate the text features of action labels. By leveraging the powerful generalization ability of the pretrained text model and the concise heatmap pooled features, we generate action text features with richer semantic information to assist in human action recognition. First, we use a pretrained text encoding model to encode the action texts of all categories (e.g., *drink water*, *eat meal*, etc. in the NTU 60 dataset), obtaining text features that describe the action categories by:

$$\boldsymbol{F}_{\text{text}} = [\mathcal{E}_t(\text{text}_i)]_{i=1}^{N} \in \mathbb{R}^{N\times C}, \tag{6}$$

where $\mathcal{E}_t$ represents the text encoder. The N represents the number of action categories and the C denotes the channel size of the encoded text features. Then, we modulate and refine the encoded text features $\boldsymbol{F}_{\text{text}}$ based on the pooled features $\boldsymbol{F}_{\mathrm{c}}$, which is formulated as:

$$\boldsymbol{F}'_{\text{text}} = \boldsymbol{F}_{\text{text}} + \psi(\boldsymbol{F}_{\mathrm{c}}, \boldsymbol{F}_{\text{text}}) + \eta(\boldsymbol{F}_{\mathrm{c}}), \tag{7}$$

$$\psi(\boldsymbol{F}_{\mathrm{c}}, \boldsymbol{F}_{\text{text}}) = \gamma \boldsymbol{F}_{\mathrm{c}} \boldsymbol{F}_{\text{text}} + \beta = \mathcal{P}_1(\boldsymbol{F}_{\mathrm{c}})\boldsymbol{F}_{\text{text}} + \mathcal{P}_2(\boldsymbol{F}_{\mathrm{c}}), \tag{8}$$

$$\eta(\boldsymbol{F}_{\mathrm{c}}) = [\mathcal{P}_3(\boldsymbol{F}_{\mathrm{c}}) - \mathcal{P}_4(\boldsymbol{F}_{\mathrm{c}})] \circ \boldsymbol{F}_{\mathrm{c}}, \tag{9}$$

where γ and β represent the scale factor and shift factor respectively. The $\circ$ represents the element-wise multiplication and the $\mathcal{P}$ represents the projection layer, which is implemented as an MLP network layer in our HP-Net. In detail, each projection layer in TRMM is composed of two Linear layers followed by a ReLU activation function. It is worth emphasizing that the four projection layers used in TRMM are implemented as independent MLP networks, without parameter sharing. To provide a clearer understanding of the workflow of the TRMM, we show its pseudocode in Algorithm 1. In addition, the information-rich action description features $\boldsymbol{F}'_{\text{text}}$ obtained through the TRMM can be further fused with RGB video features, enabling more accurate multimodal-based human action recognition.

E. Multimodal Fusion and Loss Function

In video-based human action recognition, previous methods have demonstrated that integrating features from multiple modalities leads to more accurate recognition performance [3], [4], [6]. Therefore, our proposed HP-Net also fuses RGB video features, textual features and heatmap pooled features to achieve multimodal action recognition. Specifically, in addition to utilizing the compact heatmap pooled features produced by the FPM and the semantically rich text features generated by the TRMM, we also extract features from RGB video frames using a pre-trained image encoder. In detail, given an RGB video stream $\boldsymbol{U} = \{u_1, u_2, \ldots, u_n\}$ with n frames, we extract the RGB video features by:

$$\boldsymbol{F}_{\mathrm{u}} = \mathcal{E}_u\left[\mathcal{U}_n^m(\mathcal{D}(u_i) \overset{(\mathrm{h}',\mathrm{w}')}{\mapsto} u'_i)\right], \quad \forall i \in \{1, 2, \ldots, n\}, \tag{10}$$

where $\mathcal{D}$ denotes a human detector. Inspired by the top-down pose estimation, we use a human detector $\mathcal{D}$ to filter out most ambient noise and focus on the human body, which has already been validated in some previous works [5]. The $\mathcal{D}(u_i) \overset{(\mathrm{h}',\mathrm{w}')}{\mapsto} u'_i$ represents the resizing operation applied to the detected human frames to save computational resources and adapt to the downstream RGB visual encoder. Then, we sample m frames from n detected human frames using $\mathcal{U}_n^m$ and input them into the visual encoder $\mathcal{E}_u$ to obtain the RGB video features. Finally, we fuse the text features $\boldsymbol{F}'_{\text{text}}$ refined and modulated by the heatmap pooled features $\boldsymbol{F}_{\mathrm{c}}$ with the video features $\boldsymbol{F}_{\mathrm{u}}$ to obtain the classification scores $\boldsymbol{S}$ for each action category. In summary, the fusion of multi-modal features in our proposed HP-Net can be outlined as the following mapping process:

$$\boldsymbol{U} \overset{\text{FPM}}{\mapsto} \boldsymbol{F}_{\mathrm{p}} \overset{\text{SMCLM}}{\mapsto} \boldsymbol{F}_{\mathrm{c}} \overset{\text{TRMM}}{\mapsto} \boldsymbol{F}'_{\text{text}}, \tag{11}$$

$$\boldsymbol{S} = \boldsymbol{F}'_{\text{text}} \circ \boldsymbol{F}_{\mathrm{u}}. \tag{12}$$

In our proposed Heatmap Pooling Network (HP-Net), we calculate the cross-entropy loss between the ground truth action labels and the classification scores as the classification loss by:

$$\mathcal{L}_{\text{cls}} = \mathcal{L}_{\text{CE}}(\boldsymbol{Y}, \boldsymbol{S}) = -\sum_{i}^{N} \boldsymbol{Y}^i log(\boldsymbol{S}^i), \tag{13}$$

where N is the number of samples in a batch and $\boldsymbol{Y}^i$ is the one-hot presentation of the ground truth action label about data sample i. The $\boldsymbol{S}^i$ is the classification score of data sample i. Besides, to constrain the effective training of three independent topology models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ in the Spatial-motion Co-learning Module (SMCLM), we also calculate the cross-entropy loss between the output features of these three models and the ground truth action labels as the auxiliary loss. Therefore, the training loss function for the entire Heatmap Pooling Network (HP-Net) is defined as:

$$\mathcal{L}_{\text{loss}} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\mathrm{p}} + \lambda_2 \mathcal{L}_{\mathrm{s}} + \lambda_3 \mathcal{L}_{\mathrm{m}}, \tag{14}$$

where $\mathcal{L}_{\mathrm{p}} = \mathcal{L}_{\text{CE}}(\boldsymbol{Y}, \varphi(\boldsymbol{F}'_{\mathrm{p}}))$, $\mathcal{L}_{\mathrm{s}} = \mathcal{L}_{\text{CE}}(\boldsymbol{Y}, \varphi(\boldsymbol{F}'_{\mathrm{s}}))$ and $\mathcal{L}_{\mathrm{m}} = \mathcal{L}_{\text{CE}}(\boldsymbol{Y}, \varphi(\boldsymbol{F}'_{\mathrm{m}}))$. The φ represents the classification head used to convert features into classification scores. The λ_1, λ_2 and λ_3 represent three hyper-parameters that regulate the weights of the different auxiliary loss terms.

IV. Experiments

A. Datasets

We conduct experiments on four well-established large-scale human action recognition datasets: *NTU-RGB+D 60* [38], *NTU-RGB+D 120* [39], *Toyota-Smarthome* [67], [68] and *UAV-Human* [69]. The following sections provide a detailed introduction to these datasets and their evaluation strategies.

NTU-RGB+D 60 dataset [38] (NTU 60) is a prominent benchmark for action recognition, featuring 56,880 skeleton/pose sequences and videos of human actions. These actions are grouped into 60 distinct categories and performed by 40 unique individuals. The original paper introduces two evaluation protocols: (1) *Cross-view* (X-View), where training data is collected from cameras positioned at 0° (view 2) and 45° (view 3) and testing data comes from a -45° (view 1) perspective. (2) *Cross-subject* (X-Sub), where 20 subjects' data is used for training while the other 20 subjects are reserved for testing.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

NTU-RGB+D 120 dataset [39] (NTU 120) is an extension of the NTU-RGB+D dataset, which consists of 114,480 video samples spanning 120 action classes and is performed by 106 participants. The evaluation follows two benchmarks introduced in the original paper [39]: (1) *Cross-subject* (X-Sub), where data from 53 participants is used for training and the remaining 53 for testing. (2) *Cross-setup* (X-Set), where training samples are drawn from even-numbered setups, while odd-numbered setups are reserved for testing.

Toyota-Smarthome dataset [67], [68] is a video dataset recorded in an apartment equipped with Kinect cameras. It contains 31 daily living activities and 18 subjects. The videos were clipped per activity, resulting in a total of 16,115 video samples. We follow the protocol provided in the original paper [67] and use the same mean per-class accuracy for evaluation: *Cross-subject* (CS), where the training group consists of 11 subjects with IDs (i.e., 3, 4, 6, 7, 9, 12, 13, 15, 17, 19, 25) and the remaining 7 subjects are reserved for testing. *Cross-view* (CV), is divided into CV_1 and CV_2, where the camera 2 is used for testing and camera 5 for validation. The CV_1 selects all samples from camera 1 for training, while the CV_2 includes samples from multiple cameras: 1, 3, 4, 6 and 7 for the training set.

UAV-Human dataset [69] is a comprehensive dataset for action recognition, consisting of 22,476 video clips spanning 155 unique classes. Captured by a UAV navigating diverse urban and rural environments under both day and night conditions, the dataset provides a vivid portrayal of human activities. It features 119 individuals performing 155 different actions across 45 distinct settings, offering a rich and diverse representation of human behavior. For evaluation, the original study proposes two benchmark scenarios (CSv1 and CSv2), with data from 89 participants used for training and 30 for testing, ensuring robust and reliable analysis.

B. Implementation Details

All experiments are conducted on four NVIDIA GeForce RTX 3080 Ti GPUs. Our proposed Heatmap Pooling Network (HP-Net) uses YoloV5 [70] as the human detector for videos and utilizes SimCC [25] as the human pose estimator. In detail, we initialize the SimCC model with the pretrained weights of an HR-Net backbone trained on the COCO dataset, using an input resolution of 256×192. Besides, our HP-Net uses the visual encoder provided in XCLIP [19] and the text encoder from CLIP [71], initializing both with pre-trained parameters on Kinetics-600 [40] as provided in [19]. To effectively model the heatmap pooled features, our HP-Net selected the lightweight CTR-GCN [11] as the topology model backbone. We use weight decay with adam (AdamW) to train our model for a total number of 30 epochs with batch size 4, while keeping the text encoder fixed. Inspired by [3], [4], in addition to using the single-stream HP-Net shown in Fig. 2, we also employ CTR-GCN [11] to perform multi-stream modeling on the extracted heatmap pooled features. We retain classification scores from all streams and perform the same late fusion strategy [3], [4], [15] to achieve multi-stream ensemble. More detailed implementation can be found in our open-source code project.

C. Comparison With State-of-the-Art Methods

In Tables I and II, we compare our Heatmap Pooling Network (HP-Net) with the state-of-the-art methods on the NTU RGB+D 60, NTU RGB+D 120, Toyota-Smarthome and UAV-Human datasets. On four datasets, our model outperforms all existing methods under nearly all evaluation benchmarks. Our HP-Net using only a single stream, achieves recognition accuracies of 96.3% and 96.5% on the X-Sub and X-Set benchmarks of the NTU RGB+D 120 dataset, which outperforms the π-ViT [6] by 1.2% and 0.4% respectively. When adopting the same late-fusion strategy for multi-stream ensemble, our HP-Net achieves accuracies of 96.8% and 97.5% on the aforementioned benchmarks, which outperforms the PoseConv3D [3] by 1.5% and 1.1% respectively. Notably, unlike PoseC3D [3] that uses heavy CNNs to model 3D heatmaps, our HP-Net employs a lightweight GCN to model heatmap pooled features and achieves higher recognition accuracy when fused with RGB data.

In addition, our HP-Net achieves recognition accuracies of 97.9% and 99.7% on the X-Sub and X-View benchmarks of the NTU RGB+D 60 dataset, which outperforms the DVANet by 4.5% and 1.5% respectively. Compared to MMNet [4] that performs late fusion of 3D pose and RGB video, our HP-Net surpasses it by 1.9% and 0.9% on the above benchmarks respectively. On the Toyota-Smarthome dataset, our HP-Net also achieved state-of-the-art action recognition performance, with mean per-class accuracy of 81.6%, 57.9% and 67.4% across three benchmarks. In the UAV-Human dataset, our HP-Net achieves accuracy rates of 53.6% and 80.2% on two benchmarks, outperforming existing methods. The state-of-the-art recognition accuracy across the four datasets above fully demonstrates the effectiveness and superiority of our proposed HP-Net, as it outperforms existing methods.

D. Ablation Studies

In this section, we conduct ablation experiments on the most challenging X-Sub benchmark from NTU-RGB+D 120 dataset to demonstrate the effectiveness, robustness and simplicity of the individual modules in our proposed Heatmap Pooling Network (HP-Net). First, we thoroughly explored the effectiveness of the proposed Feedback Pooling Module (FPM). Then, we investigated the transferability of the heatmap pooled features generated based on the FPM. Finally, we examined the impact of the settings of some hyperparameters and modules in the HP-Net on the performance.

1) Effectiveness of FPM: In Table III, we model the outputs from the same pose estimation process using different visual backbones (e.g., GCN and Transformer) to demonstrate the effectiveness and robustnes of our proposed feedback pooling module (FPM). The 2D Pose in Table III represents the human skeleton data directly obtained from videos using pose estimation, while the Pooled Feature refers to the concise features obtained by pooling heatmaps using 2D pose data based on our proposed FPM. Meanwhile, we convert the 2D pose and heatmap pooled features into four modalities (i.e., Joint, Bone, Joint Motion and Bone Motion) to comprehensively demonstrate

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

TABLE I
ACCURACY COMPARISON WITH STATE-OF-THE-ART METHODS. IN THE NTU RGB+D 60 AND NTU RGB+D 120 DATASETS, WE REPORT THE TOP-1 ACCURACY, WHILE IN THE TOYOTA-SMARTHOME DATASET, WE REPORT THE MEAN PER-CLASS ACCURACY. THE SYMBOLS R, P AND D DENOTE THE RGB, POSE AND DEPTH MODALITIES, RESPECTIVELY. THE SYMBOL ⋆ DENOTES THE RERUNNING RESULT OF AUTHORS' ORIGINAL CODE.

Method	Source	Visual Modality	NTU RGB+D 60 (%)		NTU RGB+D 120 (%)		Toyota-Smarthome (%)		
			X-Sub	X-View	X-Sub	X-Set	CS	CV_1	CV_2
VPN [32]	ECCV'20	R+P	95.5	98.0	87.8	86.3	60.8	43.8	53.5
VPN++ [46]	TPAMI'21	R+P	96.6	99.1	90.7	92.5	71.0	-	58.1
TSMF [34]	AAAI'21	R+P	92.5	97.4	87.0	89.1	-	-	-
DRDIS [9]	TCSVT'22	R+D	91.1	94.3	81.3	83.4	-	-	-
MMNet [4]	TPAMI'22	R+P	96.0	98.8	92.9	94.4	70.1	37.4	46.6
VideoMAE⋆ [47]	NIPS'22	R	89.3	91.2	86.5	88.2	-	-	-
PoseConv3D [3]	CVPR'22	R+P	97.0	99.6	95.3	96.4	53.8	21.5	33.4
ViewCon [48]	WACV'23	R+P	93.7	98.9	85.6	87.5	-	-	-
STAR-Transformer [49]	WACV'23	R+P	92.0	96.5	90.3	92.7	-	-	-
3DA [21]	ICCV'23	R+P	94.3	97.9	90.5	91.4	-	-	-
MMCL [5]	MM'24	R+P	93.5	97.4	90.3	91.7	-	-	-
DVANet [50]	AAAI'24	R	93.4	98.2	90.4	91.6	-	-	-
π-ViT [6]	CVPR'24	R+P	96.3	99.0	95.1	96.1	73.1	55.6	65.0
VA-AR [51]	AAAI'25	P	93.1	97.2	90.3	91.5	-	-	-
Hulk [52]	TPAMI'25	P	94.3	-	-	-	-	-	-
LA-GCN [14]	TMM'25	P	93.5	97.2	90.7	91.8	-	-	-
Heterogeneous [53]	CVPR'25	R+P	87.8	93.7	78.9	82.2	-	-	-
ProtoGCN [54]	CVPR'25	P	93.8	97.8	90.9	92.2	-	-	-
HP-Net (1s)	**Ours**	R	**97.0**	**99.0**	**96.3**	**96.5**	**81.3**	**51.7**	**66.4**
HP-Net	**Ours**	R	**97.9**	**99.7**	**96.8**	**97.5**	**81.6**	**57.9**	**67.4**

TABLE II
ACCURACY COMPARISON WITH STATE-OF-THE-ART METHODS ON THE UAV-HUMAN DATASET. THE SYMBOLS R AND P DENOTE THE RGB AND POSE MODALITIES, RESPECTIVELY. THE SYMBOL ⋆ DENOTES THE RERUNNING RESULT OF AUTHORS' ORIGINAL CODE.

Method	Source	Visual Modality	UAV-Human (%)	
			CSv1	CSv2
ST-GCN [10]	AAAI'18	P	30.3	56.1
DGNN [55]	CVPR'19	P	29.9	-
2s-AGCN [41]	CVPR'19	P	34.8	66.7
HARD-Net [56]	ECCV'20	P	37.0	-
Shift-GCN [57]	CVPR'20	P	38.0	67.0
CTR-GCN [11]	ICCV'21	P	45.6	73.5
FAR [58]	ECCV'21	R	39.1	-
FR-AGCN [59]	Neuroc'22	P	44.0	69.5
ACFL-CTR [60]	ACM MM'22	P	45.3	-
MixFormer [61]	ACM MM'23	P	47.2	73.5
AZTR [62]	ICRA'23	R	47.4	-
TD-GCN [12]	TMM'24	P	45.4	72.9
HDBN [63]	ICME'24	P	48.0	75.4
MITFAS [64]	WACV'24	R	50.8	-
STRN [65]	TIM'25	P	45.3	72.0
$2D^3$-SkelAct [66]	TCSVT'25	P	32.5	56.2
ProtoGCN⋆ [54]	CVPR'25	P	46.1	73.5
HP-Net (1s)	**Ours**	R	**52.0**	**77.8**
HP-Net	**Ours**	R	**53.6**	**80.2**

TABLE III
COMPARISON BETWEEN 2D POSE AND HEATMAP POOLED FEATURES. WE COMPARED THE PERFORMANCE UNDER FOUR MODALITIES: JOINT (J), BONE (B), JOINT MOTION (JM) AND BONE MOTION (BM). ENSEMBLE REFERS TO COMBINING THE CLASSIFICATION SCORES OF THE ABOVE FOUR MODALITIES.

Methods	Visual Modality	Acc. (%)
CTR-GCN [11]	J 2D Pose	83.0
CTR-GCN	**J Pooled Feature**	**86.1**$^{\uparrow 3.1}$
CTR-GCN	JM 2D Pose	80.3
CTR-GCN	**JM Pooled Feature**	**83.2**$^{\uparrow 2.9}$
CTR-GCN	B 2D Pose	83.9
CTR-GCN	**B Pooled Feature**	**85.3**$^{\uparrow 1.4}$
CTR-GCN	BM 2D Pose	80.0
CTR-GCN	**BM Pooled Feature**	**82.0**$^{\uparrow 2.0}$
CTR-GCN	Ensemble	86.2
CTR-GCN	**Ensemble**	**88.7**$^{\uparrow 2.5}$
Skeleton-MixFormer [61]	J 2D Pose	82.9
Skeleton-MixFormer	**J Pooled Feature**	**85.9**$^{\uparrow 3.0}$
Skeleton-MixFormer	JM 2D Pose	79.2
Skeleton-MixFormer	**JM Pooled Feature**	**82.0**$^{\uparrow 2.8}$
Skeleton-MixFormer	B 2D Pose	84.0
Skeleton-MixFormer	**B Pooled Feature**	**84.3**$^{\uparrow 0.3}$
Skeleton-MixFormer	BM 2D Pose	80.1
Skeleton-MixFormer	**BM Pooled Feature**	**81.0**$^{\uparrow 0.9}$
Skeleton-MixFormer	Ensemble	86.2
Skeleton-MixFormer	**Ensemble**	**88.4**$^{\uparrow 2.2}$

the effectiveness of the heatmap pooled features extracted by the FPM.

In addition, we use two classic network backbones (e.g., CTR-GCN [11] and Skeleton-MixFormer [61]) to model the heatmap pooled features, demonstrating that the pooled features extracted by our FPM can be easily transferred to different visual backbones. Unlike precious methods that require heavy CNNs to model heatmaps, the heatmap pooled features obtained by our FPM can be effectively modeled using lightweight GCNs and Transformers. The results in Table III show that the heatmap pooled features extracted using FPM outperform the directly generated 2D pose data in terms of recognition performance, across different modalities and network backbones. Compared to the conventional 2D poses estimated directly from heatmaps, our FPM improves recognition accuracy in the Joint modality by 3.1% and 3.0% when modeled with different network backbones.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
COMPARISON OF DIFFERENT MODALITIES USING THE SAME MODEL BACKBONE. HERE WE USE THE SAME POSE ESTIMATION MODEL AND ACTION RECOGNITION BACKBONE TO ENSURE A FAIR COMPARISON.

Visual Modality	Source	Acc. (%)
2D Pose Data	Pose Estimation [25]	83.0
3D Pose Data	Kinect Sensor [39]	85.0
Heatmap Coordinate Vector	Pose Estimation [25]	83.9
Heatmap Pooled Feature	**Ours**	**86.1**

To thoroughly demonstrate the effectiveness of FPM, we compare the extracted heatmap pooled features with 3D pose data collected from sensors and heatmap coordinate vectors obtained by pose estimation, which is shown in Table IV. To ensure a fair comparison, we use CTR-GCN [11] as the consistent modeling backbone. The results in Table IV demonstrate that our heatmap pooled features also exhibit outstanding performance advantages over 3D pose data and heatmap coordinate vectors. The results in Tables III and IV fully validate the effectiveness of our FPM, as the heatmap pooled features it extracts outperform other common features obtained from videos in recognition performance, while requiring only lightweight downstream networks for modeling.

In Table V, we further analyze the effectiveness of the proposed heatmap pooled features for some action categories in the NTU120 dataset, where the action labels start from 1. To ensure a fair comparison, we use the same backbone modeling network to compare the proposed heatmap pooled features with the 2D pose data obtained from the same pose estimation network. Additionally, we include a comparison with the 3D pose data collected by Kinect cameras provided by the original dataset. Among the 35 action categories in Table IV, our proposed heatmap pooled features significantly improve recognition accuracy compared to both 2D and 3D pose data. Notably, for challenging actions involving environmental context or objects, our proposed heatmap pooled features demonstrate more significant performance advantages. For instance, in the action category *play with phone/tablet* (action label is 29), the accuracy of heatmap pooled features surpasses that of 2D and 3D poses by 20.4% and 12.4%, respectively. Similar examples include fine-grained action categories such as *cutting nails* (action label is 75) and *blow nose* (action label is 105).

In Table VI, we explored the performance of heatmap pooled features obtained using different human pose estimation models. In terms of implementation, we first selected two pose estimation models, HR-Net [24] and SimCC [25], to extract heatmap pooled features from the videos. Then, we utilized CTR-GCN [11], Skeleton-MixFormer [61] and the proposed HP-Net to model the extracted heatmap pooled features and compared their recognition performance. The result in Table VI indicates that the heatmap pooled features extracted using SimCC demonstrate better performance, with recognition accuracy higher than HR-Net by 0.8% and 1.9% under the same network modeling. Therefore, in our HP-Net, we selected the SimCC model as the human pose estimation network. It is worth emphasizing that our HP-Net is orthogonal to the pose estimation model, allowing it to be easily transferred to different heatmap-based human pose estimation networks.

In Table VII, we also explore the recognition performance when using the FPM to pool heatmaps of different scales and modeling them with the same network backbone. The second column in Table VII represents the size of the heatmaps, while the third column indicates the size of the heatmap pooled features, where n denotes the number of joints used for feedback pooling module (FPM). We found that the heatmap pooled features derived from human heatmap H_2 achieved better recognition performance, with an accuracy of 86.2% on the NTU-120 X-Sub benchmark. Therefore, in our proposed FP-Net, we exclusively use heatmap H_2 to derive the heatmap pooled features. In the specific pose estimation network we used, the size of heatmap H_2 is $[c_1 = 96, h_1 = 32, w_1 = 24]$, and the resulting heatmap feature's size is $[n = 17, c_1 = 96]$. In Table VII, we compare the recognition performance of the three-scale heatmaps (H_1, H_2 and H_3), where the resolution decreases while the channel dimensionality increases progressively. In image modeling, high-resolution but low-dimensional representations tend to preserve richer spatial details (e.g., , edges, contours, and textures as shown in Fig. 4), which primarily correspond to low-level visual features. In contrast, low-resolution but high-dimensional representations capture more abstract semantic information while losing a considerable amount of spatial detail. Heatmap H_2 strikes a balance between resolution and channel dimensionality, preserving sufficient spatial details while simultaneously extracting richer semantic information, which explains its superior recognition performance. To further validate the suitability of H_2 as the heatmap for extracting pooled features, we conducted a quantitative analysis based on the Fisher criterion [72]. Specifically, we computed the Fisher scores of H_1, H_2 and H_3 on the NTU-RGB+D dataset, where the Fisher score is derived from intra-class and inter-class scatter, and serves as a widely adopted metric for evaluating feature discriminability in classification tasks. As illustrated in Fig. 7, heatmap H_2 achieves the highest Fisher score, indicating stronger discriminability and greater suitability for tasks such as action recognition. This finding is consistent with the experimental results reported in Table VII.

In Table VIII, we also explore the recognition performance when using the FPM to pool heatmaps of different stages in the human pose estimation process. In Table IX, we explore the recognition performance when performing pooling over heatmaps using regions of varying sizes. Additionally, we analyze the parameters and computational cost required for modeling different forms of heatmap features in Table X. Unlike traditional methods [3], [18] that convert heatmap features into image format and rely on heavy CNNs for modeling, our heatmap pooled features can be efficiently modeled using a lightweight GCN, demonstrating superior performance in both parameters and computational cost. In Table XI, we investigate the performance of using pose-guided regional sampling on heatmaps and RGB images. The results in Table XI demonstrate that our proposed feedback pooling mechanism for heatmap aggregation achieves superior performance compared to directly applying regional sampling on RGB images.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

TABLE V
ACCURACY IMPROVEMENT OF HEATMAP POOLED FEATURES OVER 2D/3D POSE DATA FOR SOME ACTION CATEGORIES. THE BLUE AND RED FONTS INDICATE THE IMPROVEMENT RATES RELATIVE TO 2D POSE AND 3D POSE, RESPECTIVELY.

Action Label	**2**	**3**	**4**	**5**	**11**	**12**	**18**
Acc. (2D Pose data) (%)	80.7	83.5	87.2	85.8	56.4	64.0	93.8
Acc. (3D Pose data) (%)	72.7	83.5	86.4	84.4	64.5	53.7	91.6
Acc. (Pooled Feature) (%)	82.9$^{\uparrow 2.2}\uparrow^{10.2}$	90.1$^{\uparrow 6.6}\uparrow^{6.6}$	89.4$^{\uparrow 2.2}\uparrow^{3.0}$	90.9$^{\uparrow 5.1}\uparrow^{6.5}$	70.3$^{\uparrow 13.9}\uparrow^{5.8}$	71.3$^{\uparrow 7.3}\uparrow^{17.6}$	96.0$^{\uparrow 2.2}\uparrow^{4.4}$
Action Label	**29**	**30**	**31**	**37**	**44**	**47**	**50**
Acc. (2D Pose data) (%)	53.1	78.9	71.4	86.2	87.7	90.2	86.5
Acc. (3D Pose data) (%)	61.1	70.5	76.1	87.3	78.3	86.2	87.6
Acc. (Pooled Feature) (%)	73.5$^{\uparrow 20.4}\uparrow^{12.4}$	84.0$^{\uparrow 5.1}\uparrow^{13.5}$	80.8$^{\uparrow 9.4}\uparrow^{4.7}$	94.6$^{\uparrow 8.4}\uparrow^{7.3}$	92.0$^{\uparrow 4.3}\uparrow^{13.7}$	95.7$^{\uparrow 5.5}\uparrow^{9.5}$	91.6$^{\uparrow 5.1}\uparrow^{4.0}$
Action Label	**54**	**56**	**57**	**58**	**61**	**62**	**67**
Acc. (2D Pose data) (%)	90.6	89.9	88.0	91.7	89.8	90.3	71.0
Acc. (3D Pose data) (%)	90.2	87.7	92.0	95.7	92.7	89.8	78.9
Acc. (Pooled Feature) (%)	93.5$^{\uparrow 2.9}\uparrow^{3.3}$	92.4$^{\uparrow 2.5}\uparrow^{4.7}$	95.3$^{\uparrow 7.3}\uparrow^{3.3}$	98.9$^{\uparrow 7.2}\uparrow^{3.2}$	94.8$^{\uparrow 5.0}\uparrow^{2.1}$	92.9$^{\uparrow 2.6}\uparrow^{3.1}$	85.9$^{\uparrow 14.9}\uparrow^{7.0}$
Action Label	**68**	**71**	**75**	**79**	**86**	**93**	**105**
Acc. (2D Pose data) (%)	81.9	44.2	67.8	80.3	76.0	77.8	68.7
Acc. (3D Pose data) (%)	79.8	45.7	68.9	76.5	78.6	80.0	64.5
Acc. (Pooled Feature) (%)	85.7$^{\uparrow 3.8}\uparrow^{5.9}$	53.9$^{\uparrow 9.7}\uparrow^{8.2}$	78.9$^{\uparrow 11.1}\uparrow^{10.0}$	86.3$^{\uparrow 5.9}\uparrow^{9.7}$	81.9$^{\uparrow 5.9}\uparrow^{3.3}$	82.3$^{\uparrow 4.5}\uparrow^{2.3}$	83.3$^{\uparrow 14.6}\uparrow^{18.8}$
Action Label	**106**	**108**	**110**	**111**	**116**	**117**	**118**
Acc. (2D Pose data) (%)	60.7	88.5	72.5	92.5	95.0	90.1	90.3
Acc. (3D Pose data) (%)	62.4	91.8	70.8	93.0	91.7	91.8	94.4
Acc. (Heatmap Pooling) (%)	66.1$^{\uparrow 5.4}\uparrow^{3.7}$	94.8$^{\uparrow 6.3}\uparrow^{3.0}$	74.8$^{\uparrow 2.3}\uparrow^{4.0}$	96.9$^{\uparrow 4.4}\uparrow^{3.9}$	98.6$^{\uparrow 3.6}\uparrow^{6.9}$	95.7$^{\uparrow 5.6}\uparrow^{3.8}$	97.6$^{\uparrow 7.3}\uparrow^{3.2}$

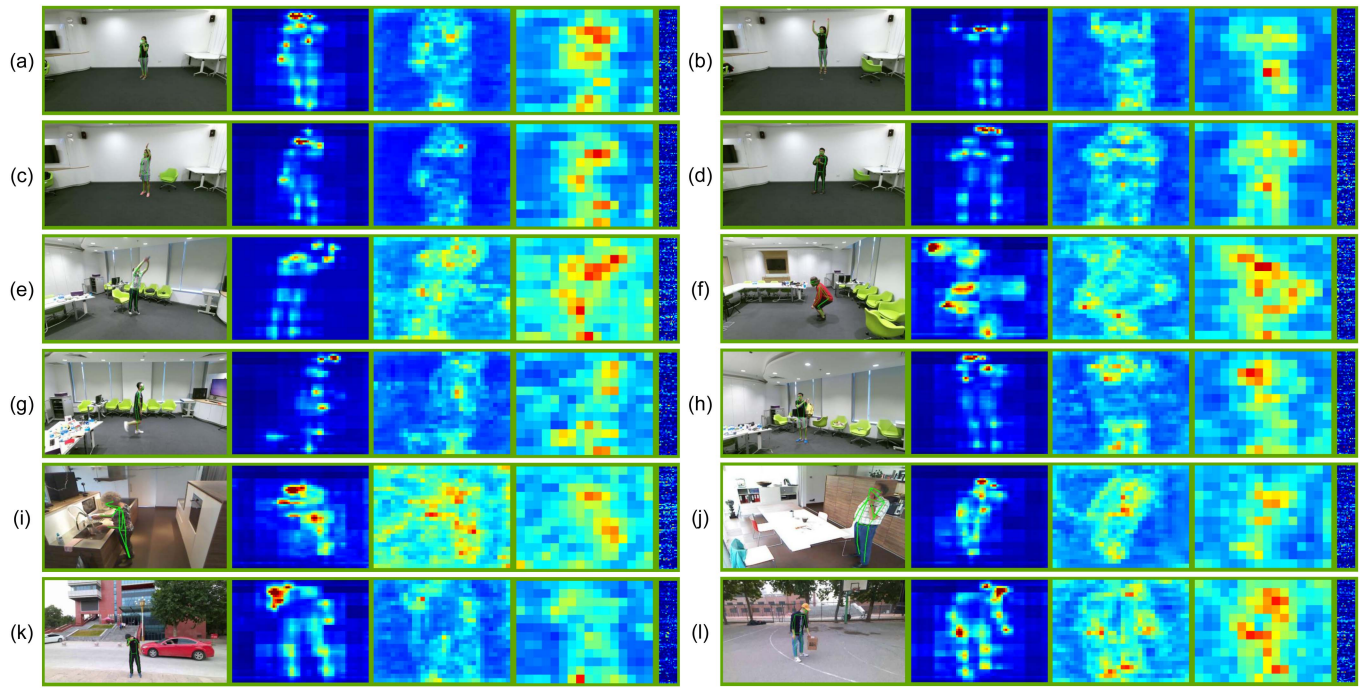


Fig. 4. Visualization of different action samples. In each subplot, the first column represents the visualization of the 2D pose data obtained from human pose estimation on the original RGB video frame. The second, third and fourth columns correspond to the human heatmaps H_1, H_2 and H_3, respectively. The fifth column shows the heatmap pooled features output by our Feedback Pooling Module (FPM).

2) Transferability of FPM: In our heatmap pooling network (HP-Net), we designed two modules (i.e., TRMM and SMCLM) to effectively fuse the heatmap pooled features generated by the FPM with the text and RGB modalities. In Table XII, we conducted ablation experiments on the designed TRMM and SMCLM, where the baseline retains only the text encoder and visual encoder. Integrating the heatmap pooled features generated by FPM via TRMM improved recognition accuracy by 0.4%, while the addition of the SMCLM further boosted accuracy by 1.0%. Additionally, we fused the heatmap pooled features

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

TABLE VI
COMPARISON OF ACCURACY WHEN OBTAIN HEATMAP POOLED FEATURES USING DIFFERENT HUMAN POSE ESTIMATION METHODS. THE HP STANDS FOR HEATMAP POOLED FEATURES.

Modality	Source	Backbone	Acc. (%)
HP	HR-Net [24]	CTR-GCN [11]	85.3
HP	SimCC [25]	CTR-GCN	86.1
HP	HR-Net	Skeleton-MixFormer [61]	84.0
HP	SimCC	Skeleton-MixFormer	85.9
HP+RGB+Text	HR-Net	HP-Net (Ours)	95.8
HP+RGB+Text	SimCC	HP-Net (Ours)	96.3

TABLE VII
COMPARISON OF THE FPM WHEN POOLING ON HEATMAPS AT DIFFERENT SCALES. h, w AND c REPRESENT THE HEIGHT, WIDTH AND CHANNEL SIZE OF THE HEATMAPS, WHILE n DENOTES THE NUMBER OF JOINTS USED IN THE FEEDBACK POOLING MODULE (FPM).

Heatmap	Heatmap Shape	Feature Shape	Acc. (%)
H_1	$[c_1, h_1, w_1]$	$[n, c_1]$	81.4
H_2	$[c_2, h_2, w_2]$	$[n, c_2]$	**86.2**
H_3	$[c_3, h_3, w_3]$	$[n, c_3]$	85.8

TABLE VIII
COMPARISON OF THE FPM WHEN POOLING ON HEATMAPS AT DIFFERENT STAGES

Heatmap	Stages	Heatmap Shape	Backbone	Acc. (%)
H_2^1	Stage1	$[c_2, h_2, w_2]$	CTR-GCN	69.4
H_2^2	Stage2	$[c_2, h_2, w_2]$	CTR-GCN	75.8
H_2^3	Stage3	$[c_2, h_2, w_2]$	CTR-GCN	**86.2**
H_2^1	Stage1	$[c_2, h_2, w_2]$	HP-Net	95.8
H_2^2	Stage2	$[c_2, h_2, w_2]$	HP-Net	95.9
H_2^3	Stage3	$[c_2, h_2, w_2]$	HP-Net	**96.3**

TABLE IX
COMPARISON OF THE FPM WHEN PERFORMING POOLING OVER HEATMAPS USING REGIONS OF VARYING SIZES

Heatmap	Stages	Region Size	Backbone	Acc. (%)
H_3	Stage3	1×1	CTR-GCN	**86.2**
H_3	Stage3	3×3	CTR-GCN	82.0
H_3	Stage3	5×5	CTR-GCN	81.9
H_3	Stage3	1×1	HP-Net	**96.3**
H_3	Stage3	3×3	HP-Net	96.2
H_3	Stage3	5×5	HP-Net	96.1

TABLE X
COMPARISON IN PARAMETERS AND COMPUTATION COST WHEN MODELING HEATMAP FEATURES

Method	Backbone	Param.	FLOPs
Heatmap Evolution [18]	2D CNN	140.1M	19.6G
Heatmap Volume [3]	3D CNN	2.0M	20.6G
Heatmap Pooling (Ours)	**GCN**	**1.4M**	**1.8G**

generated by the FPM with text features using Cross Attention [73] and Feature-wise Linear Modulation (FiLM) [74], achieving improvements in recognition performance, which demonstrates the versatility and transferability of the heatmap pooled features. The results in Table XII fully validate the effectiveness of the text refinement modulation module (TRMM) and spatial-motion co-learning module (SMCLM). By effectively integrating with text and RGB modalities, our HP-Net achieved more accurate action recognition and demonstrated the transferability of the heatmap pooled features.

TABLE XI
COMPARISON OF POSE-GUIDED REGIONAL SAMPLING ACROSS DIFFERENT MODALITIES. THE HP STANDS FOR HEATMAP POOLED FEATURES.

Method	Modality	Backbone	Acc. (%)
RGB Regional Sampling	RGB+Pose	CTR-GCN	84.8
Heatmap Regional Sampling	**HP**	**CTR-GCN**	**86.1**

TABLE XII
COMPARISON OF ACCURACY WHEN TRANSFER HEATMAP POOLED FEATURES USING DIFFERENT MODULES. THE HP STANDS FOR HEATMAP POOLED FEATURES

Method	Modality	Acc. (%)
Baseline	Text+RGB	95.3
w/ FPM+Cross Attention	Text+RGB+HP	**95.4**$^{\uparrow 0.1}$
w/ FPM+FiLM	Text+RGB+HP	**95.6**$^{\uparrow 0.3}$
w/ FPM+TRMM	Text+RGB+HP	**95.7**$^{\uparrow 0.4}$
w/ FPM+TRMM+SMCLM	Text+RGB+HP	**96.3**$^{\uparrow 1.0}$

TABLE XIII
COMPARISON OF HEATMAP POOLING NETWORK WITH AND WITHOUT USING A HUMAN DETECTOR

Benchmark	Setting	Acc. (%)
NTU 120 X-Sub	W/o Human Detector	87.0
NTU 120 X-Sub	W/ Human Detector	**96.3**$^{\uparrow 9.3}$
NTU 120 X-Set	W/o Human Detector	88.0
NTU 120 X-Set	W/ Human Detector	**96.5**$^{\uparrow 8.5}$
Smarthome CS	W/o Human Detector	80.7
Smarthome CS	W/ Human Detector	**81.3**$^{\uparrow 0.6}$

TABLE XIV
COMPARISON OF HEATMAP POOLING NETWORK WITH DIFFERENT HYPERPARAMETER SETTINGS

Method	Benchmark	λ_1	λ_2	λ_3	Acc. (%)
HP-Net	NTU 120 X-Sub	0.1	0.1	0.1	96.1
HP-Net	NTU 120 X-Sub	0.5	0.5	0.5	96.2
HP-Net	NTU 120 X-Sub	1.0	1.0	1.0	**96.3**
HP-Net	NTU 120 X-Sub	2.0	2.0	2.0	96.1

3) Setting of HP-Net: In our Heatmap Pooling Network (HP-Net), we use a human detector to extract human frames from RGB videos, aiming to filter out irrelevant environmental noise and focus on modeling human body parts. In Table XIII, we present the performance differences with and without using the human detector on the NTU 120 and Smarthome datasets. On the X-Sub benchmark of the NTU120 dataset, our HP-Net achieved a 9.3% increase in recognition accuracy when using a human detector compared to not using one. We believe that the performance difference is due to the large pixel space occupied by environmental factors in the original NTU120 dataset, which leads the visual encoder to model too much irrelevant environmental information. When a human detector is used to filter out most of the environmental factors, the visual encoder can focus more on modeling the pixel space where the human body is located, thereby improving the accuracy of human action recognition. Meanwhile, in Fig. 8, we show RGB images of different action samples with and without using the human detector. Besides, in Table XIV, we also explored the impact of different hyperparameters in the loss function on model performance.

4) Robustness of HP-Net: The UAV-Human dataset contains numerous action samples captured under extreme conditions,

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

TABLE XV
COMPARISON OF RECOGNITION ACCURACY FOR ACTION SAMPLES UNDER EXTREME CONDITIONS. THE HP STANDS FOR HEATMAP POOLED FEATURES.

Method	Modality	Acc. (%)
Baseline	Text+RGB	42.3
w/ FPM+TRMM+SMCLM	Text+RGB+HP	**51.5**$^{\uparrow 9.2}$

such as challenging weather, low-light night scenes, overexposure, and occlusion, making it highly challenging. In Fig. 9, we present several action samples captured under extreme scenarios. To further investigate the robustness of our HP-Net in recognizing actions under these extreme conditions, we conducted experiments on these samples using HP-Net, as reported in Table XV. When only standard RGB and text features were employed, the recognition accuracy was 42.3%. By incorporating our HP-Net with heatmap pooled features, the accuracy increased to 51.5%, resulting in an overall improvement of 9.2%. These results in Table XV demonstrate that our HP-Net, along with the extracted heatmap pooled features, achieves good performance gains even under extreme conditions, thereby validating the effectiveness and robustness of the proposed framework.

E. Multi-Stream Ensemble

Multi-stream ensemble is a commonly used technique in human action recognition [3], [4], [11], [12], [61], [75], [76], which achieves better performance by integrating classification scores from multiple modalities through a late-fusion strategy. In Tables XVII and XVIII, we present the recognition accuracy of multi-modality fusion using multi-stream ensemble on the NTU 60, NTU 120, Smarthome and UAV-Human datasets. In the implementation, we follow the approach in [12] to convert the heatmap pooled features into four modalities: HP_J, HP_B, HP_{JM} and HP_{BM} as shown in the Tables XVII and XVIII. We then model each of these modalities separately using the same network (i.e., CTR-GCN [11]), preserving their classification scores. Finally, we use a late-fusion strategy [3], [4], [32], [46] to combine the classification scores from the four modalities with the classification scores obtained from the HP-Net in Fig. 2, aiming to achieve better recognition performance. The experimental results in Tables XVII and XVIII show that the heatmap pooled features we proposed also exhibit strong transferability. By using a late fusion strategy for multi-stream ensemble, they enable more robust action recognition. In Table XVI, we have displayed the top-10 action samples with the highest and lowest accuracy in the single-stream HP-Net (i.e., HP-Net (1s)). Meanwhile, we have shown the improvement in the recognition accuracy of these action samples after using multi-stream ensemble (i.e., HP-Net (ms)). The results in Table XVI indicate that the accuracy of both simple and difficult actions has improved after using multi-stream ensemble, which also demonstrates that multi-stream ensemble based on our proposed heatmap pooled features will achieve better performance.

F. Visualization

In Table VII, we compare the recognition performance of heatmap pooled features obtained from human heatmaps of different scales. Meanwhile, as shown in Fig. 4, we present the visualizations of the above three-scale heatmaps from Table VII for various action samples, corresponding to the second, third and fourth columns of each subplot, respectively. Specifically, we selected 12 action samples from the NTU60 (e.g., (a) *drink water*, (b) *jump up*, (c) *hand waving*, (d) *clapping*), NTU120 (e.g., (e) *shoot at basket*, (f) *squat down*, (g) *butt kicks*, (h) *put on bag*), Toyota-Smarthome ((i) *Cook.Cleandishes*, (j) *Usetelephone*) and UAV-Human ((k) *cough*, (l) *move a box*) datasets for visualization. In each subplot, the first column represents the original RGB image and the 2D pose data obtained through pose estimation, while the fifth column illustrates the heatmap pooled features proposed in our Heatmap Pooling Network (HP-Net).

In heatmap-based pose estimation, models typically extract 2D pose data from videos, which is then used for human action recognition. Unlike previous methods that directly rely on 2D pose data, our HP-Net innovatively incorporates a feedback pooling mechanism to extract compact, informative and robust heatmap pooled features from the pose estimation process. These extracted features offer superior robustness and accuracy in action recognition, outperforming traditional 2D pose representations. In Fig. 5, we use the same network backbone to model both the 2D pose and the heatmap pooled features derived from the same pose estimation model. We then visualize the top-8 action categories on benchmarks NTU 120 XSub, Toyota-Smarthome CS and UAV-Human CSv1 where the heatmap pooled features outperform the 2D pose in recognition accuracy. We also visualize the overall recognition accuracy of the two modalities across all action categories in Fig. 5.

In Fig. 6, we showcase the differences between 2D/3D pose data and heatmap pooled features from a feature perspective. Specifically, we use the same network to model the three types of data, where 2D pose data and heatmap pooled features come from the same pose estimation network, while 3D pose data is provided by the original dataset. In detail implementation, we selected samples from 60 actions in the NTU60 X-Sub benchmark for T-distributed stochastic neighbor embedding (T-SNE) visualization. On the premise of ensuring fairness, we found that the modeled heatmap pooled features exhibit better feature discriminability. The sample features from different action categories are more dispersed and distinguishable in the feature space, which also demonstrates the superiority of our proposed heatmap pooled features.

V. DISCUSSION

In this paper, we propose a Heatmap Pooling Network (HP-Net), whose core motivation is to leverage a feedback mechanism to extract heatmap pooled features that can effectively represent human motion from videos. Extensive experiments conducted in controlled laboratory settings, home environments, and aerial surveillance scenarios comprehensively validate the effectiveness, robustness, and transferability of the heatmap pooled features. It is worth emphasizing that the core idea of HP-Net is not limited to the three aforementioned scenario-specific datasets, but rather represents a general approach to human motion analysis. For example, HP-Net can also be applied

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

TABLE XVI
PERFORMANCE IMPROVEMENTS BY AGGREGATING THE RESULTS OF MULTI-STREAM ENSEMBLE COMPARED WITH THE TOP-10 ACCURATE AND CONFUSED ACTION SAMPLES OF THE HP-NET (1S). THE HP-NET (1S) REPRESENTS l_5 IN TABLES XVII AND XVIII, WHILE HP-NET (MS) REPRESENTS $l1$+$l2$+$l3$+$l4$+$l5$ IN TABLES XVII AND XVIII.

Benchmark	Accurate Actions	HP-Net (1s) (%)	HP-Net (ms) (%)	Confused Actions	HP-Net (1s) (%)	HP-Net (ms) (%)
NTU 120 X-Sub	carry object	100.0	100.0	nausea/vomiting	84.4	86.9$^{\uparrow 2.5}$
	high-five	100.0	100.0	sneeze/cough	89.5	89.5
	cheers and drink	100.0	100.0	rub two hands	89.9	91.3$^{\uparrow 1.4}$
	cross arms	100.0	100.0	writing	91.2	93.4$^{\uparrow 2.2}$
	run on the spot	100.0	100.0	put on a shoe	91.2	93.0$^{\uparrow 1.8}$
	squat down	100.0	100.0	clapping	91.6	93.0$^{\uparrow 1.4}$
	follow	99.8	100.0$^{\uparrow 0.2}$	eat meal	91.6	92.0$^{\uparrow 0.4}$
	rock-paper-scissors	99.7	99.8$^{\uparrow 0.1}$	take off a shoe	93.1	94.9$^{\uparrow 1.8}$
	put on bag	99.7	99.8$^{\uparrow 0.1}$	hand waving	93.8	94.2$^{\uparrow 0.4}$
	take off bag	99.5	99.8$^{\uparrow 0.3}$	reach into pocket	95.3	96.4$^{\uparrow 1.1}$
UAV-Human CSv1	put on a coat	79.5	81.8$^{\uparrow 2.3}$	hit someone with something	21.2	27.3$^{\uparrow 6.1}$
	eat snacks	76.1	76.1	throw away something	23.4	27.6$^{\uparrow 4.2}$
	take off a coat	75.0	77.3$^{\uparrow 2.3}$	walk toward someone	25.0	28.1$^{\uparrow 3.1}$
	fishing	75.0	75.0	open an umbrella	28.6	42.9$^{\uparrow 14.3}$
	put hands on hips	73.3	73.3	close an umbrella	30.0	45.0$^{\uparrow 15.0}$
	sit down	72.7	72.7	walk	31.8	36.4$^{\uparrow 4.6}$
	take off a hat	68.9	68.9	give something to someone	35.5	35.5
	stand up	68.2	79.5$^{\uparrow 11.3}$	stab someone with a knife	35.5	48.4$^{\uparrow 12.9}$
	open a box	66.7	66.7	threat some with a knife	36.4	39.4$^{\uparrow 3.0}$
	apply cream to face	65.9	68.2$^{\uparrow 2.3}$	run	36.7	36.7

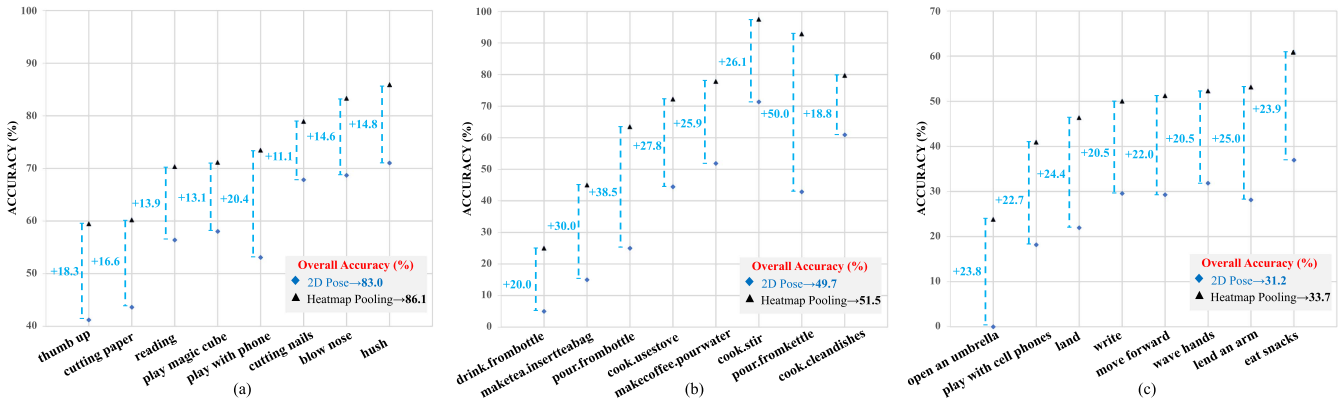


Fig. 5. Visualization of the recognition accuracy for some action categories. (a) The top-8 action categories where the use of heatmap pooled features outperforms 2D pose in the NTU 120 dataset. (b) The top-8 action categories where the use of heatmap pooled features outperforms 2D pose in the Toyota-Smarthome dataset. (c) The top-8 action categories where the use of heatmap pooled features outperforms 2D pose in the UAV-Human dataset.

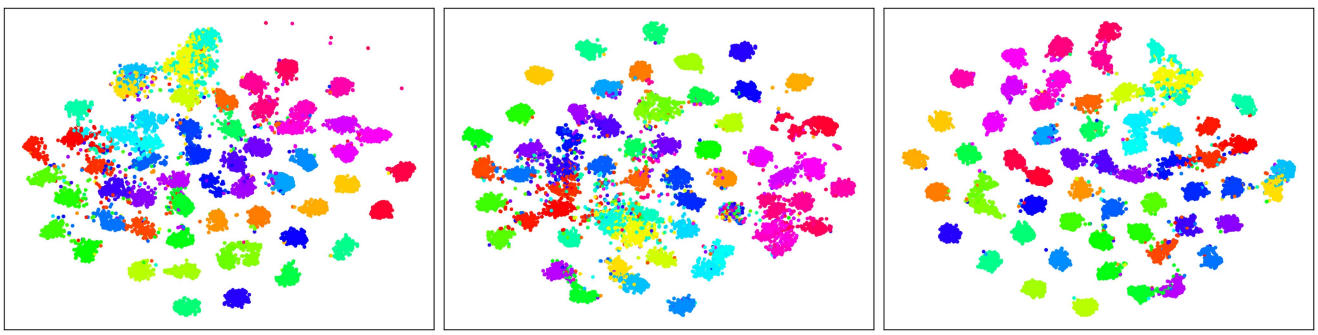


Fig. 6. T-distributed stochastic neighbor embedding (T-SNE) visualization of features. We use the same network backbone to model three types of data (i.e., 2D pose data, 3D pose data and heatmap pooled features) and visualize the modeled features. The 2D pose data and heatmap pooled features come from the same pose estimation network (i.e., SimCC [25]), while the 3D pose data is provided by the original NTU-60 dataset.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

TABLE XVII
ACCURACY OF MULTI-STREAM ENSEMBLE WHEN USING DIFFERENT MODALITY. IN THE NTU 60 AND NTU 120 DATASETS, WE REPORT THE TOP-1 ACCURACY. THE HP REPRESENTS HEATMAP POOLED FEATURES.

Modality	NTU 60 (%)		NTU 120 (%)	
	X-Sub	X-View	X-Sub	X-Set
l_1: HP_J	93.8	94.5	86.1	88.8
l_2: HP_B	92.8	92.8	85.3	87.4
l_3: HP_{JM}	91.5	93.3	83.2	85.7
l_4: HP_{BM}	90.7	92.8	82.0	84.3
l_5: HP+RGB+Text	97.0	99.0	96.3	96.5
$l1$+$l5$	97.3	99.2	96.4	96.8
$l2$+$l5$	97.1	99.1	96.4	96.7
$l1$+$l2$	94.3	95.3	87.8	90.2
$l1$+$l2$+$l5$	97.5	99.4	96.5	97.0
$l1$+$l2$+$l3$+$l4$+$l5$	**97.9**	**99.7**	**96.8**	**97.5**

TABLE XVIII
ACCURACY OF MULTI-STREAM ENSEMBLE WHEN USING DIFFERENT MODALITY. IN THE UAV-HUMAN DATASET, WE REPORT THE TOP-1 ACCURACY, WHILE IN THE TOYOTA-SMARTHOME DATASET, WE REPORT THE MEAN PER-CLASS ACCURACY. THE HP REPRESENTS HEATMAP POOLED FEATURES.

Modality	Toyota-Smarthome (%)			UAV-Human (%)	
	CS	CV_1	CV_2	CSv1	CSv2
l_1: HP_J	51.5	30.9	37.0	33.7	56.5
l_2: HP_B	51.3	29.9	30.3	31.3	52.6
l_3: HP_{JM}	45.2	29.5	34.5	28.2	49.3
l_4: HP_{BM}	43.2	28.6	33.9	25.5	44.3
l_5: HP+RGB+Text	81.3	51.7	66.4	52.0	77.8
$l1$+$l5$	81.4	53.0	66.5	53.2	79.2
$l2$+$l5$	81.1	53.9	66.3	52.9	78.8
$l1$+$l2$	55.7	32.1	37.1	36.6	60.5
$l1$+$l2$+$l5$	81.7	53.6	66.8	53.4	79.5
$l1$+$l2$+$l3$+$l4$+$l5$	**81.6**	**57.9**	**67.4**	**53.6**	**80.2**

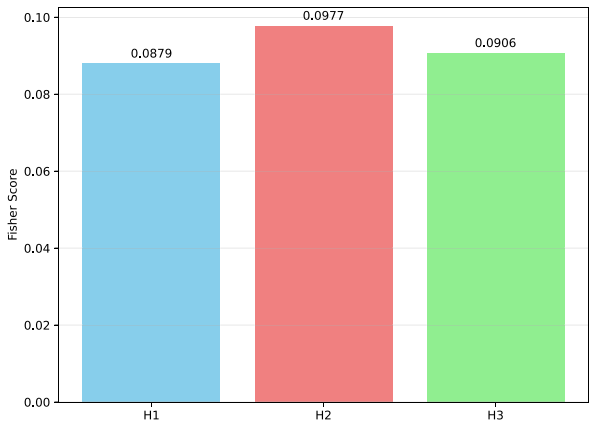


Fig. 7. Visualization of the Fisher scores of heatmaps at different scales. A higher Fisher score indicates greater feature discriminability.



Fig. 8. Visualization of different action samples with and without using a human detector. (a) We present the RGB images of datasets NTU-RGB+D 120, UAV-Human, and Toyota-Smarthome without using a human detector. (b) We present the RGB images of datasets NTU-RGB+D 120, UAV-Human, and Toyota-Smarthome with a human detector filtering environmental noise.

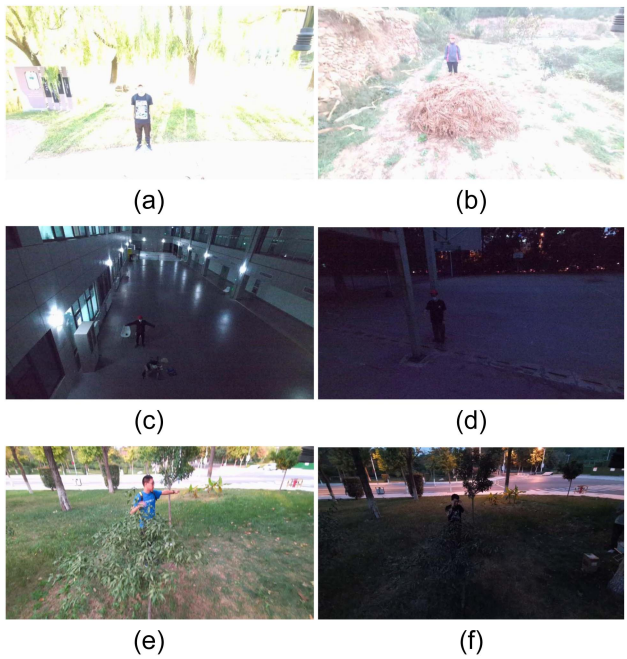


Fig. 9. Visualization of some action samples under extreme scenarios. The action samples in subfigures (a) and (b) are captured under overexposed scenarios. The action samples in subfigures (c) and (d) are captured under low-light night scenes. The action samples in subfigures (e) and (f) are captured under occlusion scenes.

to generic action recognition benchmarks such as Something-Something [77] to enhance action recognition performance. Below, we discuss the potential, adaptability and challenges of HP-Net on general action recognition benchmarks.

First, the feedback pooling mechanism employed by HP-Net does not rely on datasets from specific scenarios. Whenever heatmap-based pose or gesture estimation methods are used to extract video features for human motion analysis, the feedback pooling mechanism of HP-Net can be applied. For instance, the Something-Something dataset is an action dataset that involves interactions between human hands and objects. By employing a heatmap-based gesture estimation model to obtain gesture poses and gesture heatmaps, the estimated gesture poses can be used to feedback-pool the gesture heatmaps, thereby generating gesture-specific heatmap pooled features that help improve recognition performance. Second, the heatmap pooled features extracted by HP-Net are influenced by different heatmap-based pose estimation models. This conclusion that has been thoroughly demonstrated in the ablation experiments presented above (i.e., in Table VI). Therefore, when the heatmap-based pose estimation model fails to accurately estimate poses and heatmaps from a general action dataset, the extracted heatmap pooled features may not provide significant performance improvements for the corresponding tasks. However, this potential

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

risk does not constrain the feedback pooling concept proposed by the HP-Net, as the aforementioned issue can be effectively addressed by employing more advanced heatmap-based pose estimation networks.

In conclusion, the feedback pooling mechanism of HP-Net presents a promising and generalizable solution for action recognition from RGB videos. It can efficiently extract compact and robust heatmap pooled features, making it a strong candidate for human motion analysis. Exploring the applicability of the feedback pooling mechanism across different human motion analysis tasks represents a clear and valuable direction for our future work, and we are optimistic about the results based on its principled design.

VI. Conclusion

We present a novel Heatmap Pooling Network (HP-Net) for action recognition from RGB videos, which extracts information-rich, robust and concise heatmap pooled features through a Feedback Pooling Module (FPM). The innovative FPM is compatible with various human pose estimation methods, forging a new link between human pose estimation and video action recognition. The HP-Net extends image-based heatmap features to support a wider range of visual backbone networks while enabling seamless integration with other multimodal features for more robust and precise human action recognition. Compared to traditional modality data, the heatmap pooled features in the HP-Net establish a new modality benchmark for video action recognition. The effectiveness of our proposed HP-Net is verified on four benchmark datasets namely NTU RGB+D 60, NTU RGB+D 120, Toyota-Smarthome and UAV-Human, where it outperforms state-of-the-art methods.

References

[1] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast networks for video recognition," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.

[2] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, 2017.

[3] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2969–2978.

[4] B. X. Yu, Y. Liu, X. Zhang, S.-H. Zhong, and K. C. Chan, "MMNet: A model-based multimodal network for human action recognition in RGB-D videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3522–3538, Mar. 2023.

[5] J. Liu, C. Chen, and M. Liu, "Multi-modality co-learning for efficient skeleton-based action recognition," in *Proc. ACM Int. Conf. Multimedia*, 2024, pp. 4909–4918.

[6] D. Reilly and S. Das, "Just add π! Pose induced video transformers for understanding activities of daily living," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 18340–18350.

[7] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1474–1488, Feb. 2023.

[8] Q. Cheng, J. Cheng, Z. Liu, Z. Ren, and J. Liu, "A dense-sparse complementary network for human action recognition based on RGB and skeleton modalities," *Expert Syst. Appl.*, vol. 244, 2024, Art. no. 123061.

[9] H. Wu, X. Ma, and Y. Li, "Spatiotemporal multimodal learning with 3D CNNs for video action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1250–1261, Mar. 2022.

[10] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, Art. no. 912.

[11] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng, and W. Hu, "Channel-wise topology refinement graph convolution for skeleton-based action recognition," in *Proc. Int. Conf. Comput. Vis.*, 2021, pp. 13359–13368.

[12] J. Liu, X. Wang, C. Wang, Y. Gao, and M. Liu, "Temporal decoupling graph convolutional network for skeleton-based gesture recognition," *IEEE Trans. Multimedia*, vol. 26, pp. 811–823, 2024.

[13] M. Wang, J. Xing, and Y. Liu, "ActionCLIP: A new paradigm for video action recognition," 2021, *arXiv:2109.08472*.

[14] H. Xu, Y. Gao, Z. Hui, J. Li, and X. Gao, "Language knowledge-assisted representation learning for skeleton-based action recognition," *IEEE Trans. Multimedia*, vol. 27, pp. 5784–5799, 2025.

[15] J. Liu et al., "Explore human parsing modality for action recognition," *CAAI Trans. Intell. Technol.*, vol. 9, no. 6, pp. 1623–1633, 2024.

[16] Z. Deng, X. Li, X. Li, Y. Tong, S. Zhao, and M. Liu, "VG4D: Vision-language model goes 4D video recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 5014–5020.

[17] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[18] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1159–1168.

[19] B. Ni et al., "Expanding language-image pretrained models for general video recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 1–18.

[20] Y. Wang et al., "InternVideo: General video foundation models via generative and discriminative learning," 2022, *arXiv:2212.03191*.

[21] S. Kim, D. Ahn, and B. C. Ko, "Cross-modal learning with 3D deformable attention for action recognition," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 10265–10275.

[22] W. Xiang, C. Li, Y. Zhou, B. Wang, and L. Zhang, "Generative action description prompts for skeleton-based action recognition," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 10276–10285.

[23] H. Zhou, Q. Liu, and Y. Wang, "Learning discriminative representations for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10608–10617.

[24] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[25] Y. Li et al., "SimCC: A simple coordinate classification perspective for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 89–106.

[26] W. Li, M. Liu, H. Liu, P. Wang, J. Cai, and N. Sebe, "Hourglass tokenizer for efficient transformer-based 3D human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 604–613.

[27] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:1703.07475*.

[28] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2649–2656.

[29] D. Shao, Y. Zhao, B. Dai, and D. Lin, "FineGym: A hierarchical video dataset for fine-grained action understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2616–2625.

[30] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[31] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.

[32] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 72–90.

[33] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13289–13299.

[34] X. Bruce, Y. Liu, and K. C. Chan, "Multimodal fusion via teacher-student network for indoor action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 3199–3207.

[35] S. Chi, H.-G. Chi, Q. Huang, and K. Ramani, "InfoGCN++: Learning representation by predicting the future for online skeleton-based action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 1, pp. 514–528, Jan. 2025.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.

[36] S. Yang, J. Liu, S. Lu, E. M. Hwa, and A. C. Kot, "One-shot action recognition via multi-scale spatial-temporal skeleton matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 7, pp. 5149–5156, Jul. 2024.
[37] S. Yang, J. Liu, S. Lu, E. M. Hwa, Y. Hu, and A. C. Kot, "Self-supervised 3D action representation learning with skeleton cloud colorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 509–524, Jan. 2024.
[38] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGBD: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1010–1019.
[39] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.
[40] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.
[41] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12026–12035.
[42] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 466–481.
[43] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.
[44] H.-S. Fang et al., "AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7157–7173, Jun. 2023.
[45] H. Qu, L. Xu, Y. Cai, L. G. Foo, and J. Liu, "Heatmap distribution matching for human pose estimation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 24327–24339.
[46] S. Das, R. Dai, D. Yang, and F. Bremond, "VPN++: Rethinking video-pose embeddings for understanding activities of daily living," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9703–9717, Dec. 2022.
[47] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, Art. no. 732.
[48] K. Shah, A. Shah, C. P. Lau, C. M. de Melo, and R. Chellappa, "Multi-view action recognition using contrastive learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3381–3391.
[49] D. Ahn, S. Kim, H. Hong, and B. C. Ko, "Star-transformer: A spatio-temporal cross attention transformer for human action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 3330–3339.
[50] N. Siddiqui, P. Tirupattur, and M. Shah, "DVANet: Disentangling view and action features for multi-view action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 4873–4881.
[51] J. Wei et al., "VA-AR: Learning velocity-aware action representations with mixture of window attention," in *Proc. AAAI Conf. Artif. Intell.*, 2025, pp. 8286–8294.
[52] Y. Wang et al., "Hulk: A universal knowledge translator for human-centric tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 7, pp. 5672–5689, Jul. 2025.
[53] H. Wang, X. Ma, J. Kuang, and J. Gui, "Heterogeneous skeleton-based action representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 19154–19164.
[54] H. Liu, Y. Liu, M. Ren, H. Wang, Y. Wang, and Z. Sun, "Revealing key details to see differences: A novel prototypical perspective for skeleton-based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 1–8.
[55] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7912–7921.
[56] T. Li, J. Liu, W. Zhang, and L. Duan, "HARD-Net: Hardness-aware discrimination network for 3D early activity prediction," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 420–436.
[57] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 183–192.
[58] D. Kothandaraman, T. Guan, X. Wang, S. Hu, M. Lin, and D. Manocha, "FAR: Fourier aerial video recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 657–676.
[59] Z. Hu, Z. Pan, Q. Wang, L. Yu, and S. Fei, "Forward-reverse adaptive graph convolutional networks for skeleton-based action recognition," *Neurocomputing*, vol. 492, pp. 624–636, 2022.
[60] X. Wang, Y. Dai, L. Gao, and J. Song, "Skeleton-based action recognition via adaptive cross-form learning," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 1670–1678.
[61] W. Xin, Q. Miao, Y. Liu, R. Liu, C.-M. Pun, and C. Shi, "Skeleton MixFormer: Multivariate topology representation for skeleton-based action recognition," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 2211–2220.
[62] X. Wang et al., "AZTR: Aerial video action recognition with auto zoom and temporal reasoning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 1312–1318.
[63] J. Liu, B. Yin, J. Lin, J. Wen, Y. Li, and M. Liu, "HDBN: A novel hybrid dual-branch network for robust skeleton-based action recognition," in *Proc. IEEE Int. Conf. Multimedia Expo Workshop*, 2024, pp. 1–6.
[64] R. Xian, X. Wang, and D. Manocha, "MITFAS: Mutual information based temporal feature alignment and sampling for aerial video action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 6625–6634.
[65] R. Wang, J. Jin, Z. Chen, C. Wu, X.-J. Wu, and N. Sebe, "Structural topology refinement network for skeleton-based action recognition," *IEEE Trans. Instrum. Meas.*, vol. 74, 2025, Art. no. 2532716.
[66] X. Zhang, Y. Jia, J. Zhang, Y. Yang, and Z. Tu, "Robust 2D skeleton action recognition via decoupling and distilling 3D latent features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 10, pp. 10410–10422, Oct. 2025.
[67] S. Das et al., "Toyota smarthome: Real-world activities of daily living," in *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 833–842.
[68] R. Dai et al., "Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2533–2550, Feb. 2023.
[69] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-Human: A large benchmark for human behavior understanding with unmanned aerial vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16266–16275.
[70] R. Khanam and M. Hussain, "What is YOLOv5: A deep look into the internal features of the popular object detector," 2024, *arXiv:2407.20892*.
[71] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[72] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," 2012, *arXiv:1202.3725*.
[73] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst. Red Hook*, 2017, pp. 6000–6010.
[74] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI Conf. Artif. Intell.*, 2018, Art. no. 483.
[75] J. Lee, M. Lee, D. Lee, and S. Lee, "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 10444–10453.
[76] W. Myung, N. Su, J.-H. Xue, and G. Wang, "DeGCN: Deformable graph convolutional networks for skeleton-based action recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 2477–2490, 2024.
[77] R. Goyal et al., "The "something something" video database for learning and evaluating visual common sense," in *Proc. Int. Conf. Comput. Vis.*, 2017, pp. 5842–5850.



Mengyuan Liu received the PhD degree from the School of Electrical Engineering and Computer Science, Peking University. He was a research fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University. Currently, he is an assistant professor with Peking University, Shenzhen Graduate School. His research focuses on human-centric perception and robot learning. His work has been published in leading conferences and journals, including CVPR and *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He also serves as an associate editor of the *Pattern Recognition*.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.



Jinfu Liu received the MSc degree from Sun Yat-sen University, Shenzhen, China, in 2025. He is currently an algorithm engineer with DJI Technology Company Ltd. His work have been published in *IEEE Transactions on Multimedia* and ACM MM. His research interests include video understanding, content generation, and on-device AI. He also serves as a reviewer of *IEEE Transactions on Circuits and Systems for Video Technology* and ICME.



Yongkang Jiang received the BSc and PhD degrees in mechanical and electronic engineering from Beihang University, Beijing, China, in 2015 and 2021, respectively. He is currently an assistant professor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His research interests include computer vision and physical interaction.



Bin He received the PhD degree in mechanical and electronic control engineering from Zhejiang University, Hangzhou, China, in 2001. He held his postdoctoral research appointments with the State Key Laboratory of Fluid Power Transmission and Control, Zhejiang University, from 2001 to 2003. He is currently a professor with the College of Electronics and Information Engineering, Tongji University, Shanghai, China. His current research interests include computer vision and intelligent robot.

Authorized licensed use limited to: Peking University. Downloaded on October 21,2026 at 04:11:39 UTC from IEEE Xplore. Restrictions apply.