Attentive Keypoint Identification: Progressive Spatiotemporal Refinement for Video-based Human Pose Estimation

Sifan Wu[1,2], Haipeng Chen[1,2], Yingda Lyu[1,3*], Shaojing Fan[4], Zhigang Wang[5], Zhenguang Liu[5,6,7*], Yingying Jiao[8*]

[1]College of Computer Science and Technology, Jilin University,
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University,
[3]Public Computer Education and Research Center, Jilin University,
[4]Department of Electrical and Computer Engineering, National University of Singapore,
[5]The State Key Laboratory of Blockchain and Data Security, Zhejiang University,
[6]Shandong Rendui Network Co., Ltd.,
[7]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security,
[8]College of Computer Science and Technology, Zhejiang University of Technology,
wusifan2021@gmail.com, chenhp@jlu.edu.cn, ydlv@jlu.edu.cn, dcsfs@nus.edu.sg,
wangzhigang2024@gmail.com, liuzhenguang2008@gmail.com, yingyingjiao21@gmail.com

Abstract

Video-based human pose estimation has vast applications such as *action recognition*, *sports analytics*, and *crime detection*. However, this task is challenging as it involves interpreting both spatial context and temporal dynamics to accurately localize human anatomical keypoints in video sequences. Current approaches, often based on attention mechanisms, perform well but struggle in challenging scenarios like rapid motion and pose occlusion. We attribute these failures to two fundamental limitations: *spatial uniformity*, where models indiscriminately assign attention to both joint-relevant features and background clutter, thereby introducing spatial noise; and *temporal rigidity*, an inability to adapt to large joint displacements, resulting in severe feature misalignment during rapid motion. To overcome these challenges, we introduce PSTPose, a novel progressive spatiotemporal refinement framework. Specifically, to address the *spatial uniformity* problem, we propose a Discriminative Feature Enhancement (DFE) module that emphasizes joint-relevant features and a Feature Cluster Grouping (FCG) module that forms compact, semantically meaningful regions. For the *temporal rigidity* problem, we introduce a Deformable Spatiotemporal Fusion (DSF) module that adaptively aligns features across consecutive frames via deformation-aware sampling. This design ensures robust keypoint localization, particularly in cluttered and dynamic scenes. Extensive experiments on three large-scale benchmarks, PoseTrack2017, PoseTrack2018, PoseTrack21, demonstrate that PSTPose establishes a new state-of-the-art.

Introduction

As a core component of human-centric vision understanding systems (Liu et al. 2024, 2025), human pose estimation aims to localize anatomical keypoints in images or videos. This capability is essential for diverse applications ranging from motion analysis (Tang et al. 2023; Xu et al. 2025) to motion generation (Liu et al. 2022b; Wu et al. 2024b). Consequently, the critical role of human pose estimation in practical scenarios has spurred growing interest across both academic research and industrial development.

Driven by recent breakthroughs in deep learning techniques (Tang et al. 2020, 2025; Fu et al. 2025), AI has achieved impressive performance in many real-world scenarios (Yuan et al. 2025; Li et al. 2025a,b). For human pose estimation, early works predominantly focus on image-level tasks. Andriluka et al. (Andriluka, Roth, and Schiele 2012) utilize specific handcrafted features to estimate poses in images. Significant advancements in deep learning (Krizhevsky, Sutskever, and Hinton 2012; Vaswani et al. 2017) and the availability of large-scale datasets (Lin et al. 2014; Andriluka et al. 2014) have greatly accelerated image-based human pose estimation, leading to powerful models based on convolutional networks (Xiao, Wu, and Wei 2018), Transformers (Xu et al. 2023), and diffusion models (Tan et al. 2024; Zhang et al. 2025b,a). While these approaches achieve impressive performance on static images, many real-world applications, such as action recognition, sports analytics, and human–computer interaction (Hernández et al. 2021), demand robust video-level pose estimation. However, existing image-based methods often struggle when applied to videos due to their inability to model temporal dynamics, resulting in inconsistent and inaccurate pose estimation across frames.

To this end, recent research has increasingly focused on explicit spatiotemporal modeling for video-based human pose estimation. Various approaches have emerged, from methods that integrate motion cues via heatmap residuals (Bertasius et al. 2019; Liu et al. 2021), to more recent techniques that employ information-theoretic objectives (Liu et al. 2022a; Feng et al. 2023a) or decoupled attention mechanisms (He and Yang 2024) to supervise spatiotemporal feature learning. Despite these advances, we argue that prevailing methods are hampered by two fundamental limitations (Ye et al. 2025; Weiguang, Dong, and

*Corresponding authors.

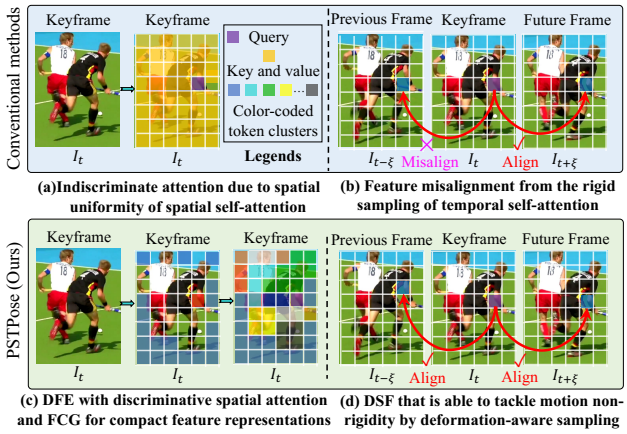

Figure 1: Comparison of conventional self-attention (a, b) with our progressive refinement framework, PSTPose (c, d). Conventional spatial attention is indiscriminate, leading to spatial noise (*spatial uniformity*) (a), and temporal attention fails to track large joint displacements due to a fixed sampling location (*temporal rigidity*) (b). In contrast, PSTPose first employs a Discriminative Feature Enhancement (DFE) module to suppress background noise, followed by a Feature Cluster Grouping (FCG) module that forms compact features (c). Then, our Deformable Spatiotemporal Fusion (DSF) adaptively samples keypoint-relevant features across frames, capturing comprehensive representations for robust pose estimation (d).

Lu 2023): 1) **Spatial uniformity.** Standard spatial attention often processes feature tokens uniformly. As shown in Fig. 1 (a), when processing a specific query, spatial attention assigns comparable attention to both critical anatomical keypoints and irrelevant background clutter. This lack of selectivity introduces spatial noise and hinders the learning of keypoint-centric representations. 2) **Temporal rigidity.** Human motion is inherently non-rigid, yet existing temporal models often operate with a fixed sampling location. This structural rigidity leads to feature misalignment across consecutive frames when joints undergo large displacements. As visualized in Fig. 1 (b), to find the elbow feature corresponding to its position in frame I_t, temporal attention samples from the same location in the previous frame $I_{t-\xi}$. Because the elbow has moved, this static sampling process results in flawed temporal aggregation.

To tackle these limitations, we propose a progressive spatiotemporal refinement framework for video-based human pose estimation. Our framework addresses the aforementioned challenges through a carefully designed, sequential pipeline. Specifically, to combat the *spatial uniformity* problem, we first present a Discriminative Feature Enhancement (DFE) module. It utilizes a differential attention mechanism to distinguish foreground from background, selectively amplifying joint-relevant features while actively suppressing spatial noise. Building upon these enhanced features, we introduce a Feature Cluster Grouping (FCG) module to model inter-joint dependencies more efficiently. It organizes feature tokens into semantically compact groups using density-peak clustering. As illustrated in Fig. 1 (c), the DFE and FCG modules effectively transform the uniform feature map into a compact keypoint-relevant representation. Finally, our Deformable Spatiotemporal Fusion (DSF) module addresses the *temporal rigidity* problem. As shown in Figure 1(d), it learns deformation-aware offsets to dynamically adjust its sampling locations (*e.g.*, elbow in frame $I_{t-\xi}$), ensuring accurate feature aggregation across frames. Our progressive refinement of spatiotemporal features enables a more robust and accurate pose estimation.

In summary, our contributions are three-fold:

- We propose a novel progressive spatiotemporal refinement framework (PSTPose) that systematically addresses two core challenges in video-based human pose estimation: *spatial uniformity* and *temporal rigidity*. PSTPose refines features across three stages, progressively enhancing joint specificity and temporal coherence. Extensive experiments on three large-scale benchmarks demonstrate the effectiveness of PSTPose, achieving new state-of-the-art performance.
- We propose a discriminative feature enhancement module that selectively amplifies joint-relevant features while suppressing spatial noise, and a feature cluster grouping module that organizes features into compact clusters to improve dependency modeling.
- We design a deformable spatiotemporal fusion that adaptively aligns and integrates features across frames using learned, deformation-aware sampling offsets, enabling robust modeling of non-rigid human motion.

Related Work

Human Pose Estimation in Images. Early research in human pose estimation focuses primarily on the static images. A significant body of work has explored CNN-based architectures for this task (Xiao, Wu, and Wei 2018; Artacho and Savakis 2020). More recently, Transformer-based models, which operate by dividing an image into patches and capturing dependencies between them, have gained prominence (Xu et al. 2023). Other emerging paradigms include generative approaches, such as diffusion models (Tan et al. 2024). While these image-level methods have achieved remarkable progress, their direct application to video often yields suboptimal results due to inherent limitations in modeling temporal information, particularly under challenging scenes such as pose occlusion or video defocus.

Human Pose Estimation in Videos. Driven by a wide range of real-world applications such as action recognition, sports analytics, and human–computer interaction, video-based human pose estimation has emerged as a pivotal research direction (Chen et al. 2025; Jiao et al. 2025; Wu et al. 2025; Wang et al. 2025) beyond static image-based methods. Early efforts combine CNNs with LSTMs to model spatial and temporal information sequentially (Artacho and Savakis 2020), or compute inter-frame heatmap residuals to generate motion-aware features (Bertasius et al. 2019; Liu et al. 2021). More recent methods have introduced sophisticated

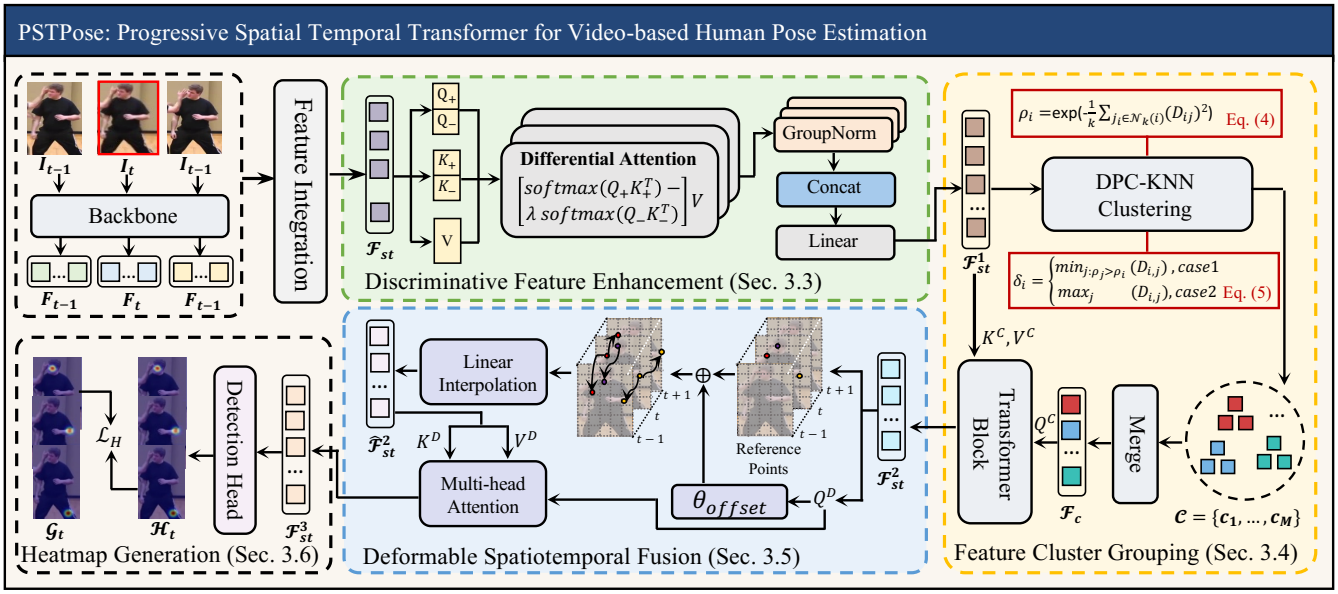

Figure 2: Overview of our PSTPose framework. Given an input video clip $\mathcal{I}_t^i = \{I_{t-\xi}^i, ..., I_t^i, ..., I_{t+\xi}^i\}$ (for simplicity, the person index i is omitted and temporal span $\xi = 1$ is used in the figure), a backbone network first extracts initial feature tokens. These features are then integrated and fed into a progressive spatiotemporal refinement pipeline consisting of three core modules. First, the Discriminative Feature Enhancement module (green block) selectively amplifies human joint regions while suppressing spatial noise from the background. Next, to facilitate more effective dependency modeling, the Feature Cluster Grouping module (yellow block) organizes the enhanced features into semantically coherent clusters. Subsequently, the Deformable Spatiotemporal Fusion module (blue block) adaptively samples keypoint-relevant information across frames using a deformation-aware sampling location to handle complex motions. Finally, the refined features are passed to a detection head to generate the keypoint heatmaps $\mathcal{H}_t$.

learning objectives grounded in information theory to better supervise temporal dependency modeling (Liu et al. 2022a; Feng et al. 2023a; Zhang et al. 2025c). He and Yang (He and Yang 2024) design a decoupled spatial and temporal self-attention mechanism to estimate human pose in videos. However, their reliance on standard self-attention for spatiotemporal feature modeling introduces two inherent limitations (Ye et al. 2025; Weiguang, Dong, and Lu 2023): spatial uniformity and temporal rigidity. To this end, we propose a progressive spatiotemporal refinement framework for video-based human pose estimation, featuring systematic and progressive refinement to resolve these problems.

Methodology

Preliminary

The objective of video-based human pose estimation is to localize a set of K anatomical keypoints for each person across video sequences. Given a keyframe I_t and its neighboring frames $\{I_{t-\xi}, ..., I_{t+\xi}\}$ (where ξ is a predefined temporal span), the task is to estimate the multi-person skeletal poses of I_t. Specifically, we first employ an object detector to identify person-specific bounding boxes in the keyframe I_t. To ensure consistent cropping of the same individual across neighboring frames $\{I_{t-\xi}, ..., I_{t+\xi}\}$, we expand each bounding box by 25%. This yields a cropped clip sequence $\mathcal{I}_t^i = \{I_{t-\xi}^i, ..., I_t^i, ..., I_{t+\xi}^i\}$ for each person i. The clip $\mathcal{I}_t^i$ serves as the input to a pose estimation network, which leverages the spatiotemporal information to predict a set of K corresponding heatmaps, $\mathcal{H}_t^i = \{H_1, H_2, ..., H_K\}$, for the keyframe I_t^i. Each heatmap H_k represents the likelihood distribution for the location of the k^{th} keypoint. The final keypoint coordinates are derived from the locations of maximum activation within these heatmaps. For clarity in the following sections, we omit the person identity superscript i and set the temporal span $\xi = 1$.

Method Overview

As illustrated in Figure 2, our PSTPose framework processes a video clip via a progressive spatiotemporal refinement pipeline composed of three novel modules. The pipeline begins with a backbone network that extracts initial feature tokens $\mathcal{F} = \{F_{t-1}, F_t, F_{t+1}\}$ from the input video clip $\mathcal{I}_t$. These frame-level features are then concatenated via a feature integration operation, yielding a raw spatiotemporal feature $\mathcal{F}_{st}$. This feature is then fed into the discriminative feature enhancement module, which selectively amplifies human joint signals while suppressing spatial noise. Subsequently, the feature cluster grouping module organizes these enhanced features into semantically coherent clusters, enabling more effective inter-joint dependency modeling. Next, the deformable spatiotemporal fusion module adaptively samples keypoint-relevant information from consecutive frames through deformation-aware sampling offsets. Fi-

nally, the refined features are passed to a detection head to generate keypoint heatmaps $\mathcal{H}_t$.

Discriminative Feature Enhancement

Prevailing attention mechanisms in video-based pose estimation often suffer from the *spatial uniformity problem*, treating feature tokens from human joints and irrelevant background with uniform importance. This uniform processing can dilute valuable feature representations with spatial noise, consequently degrading keypoint localization accuracy. To address this limitation, we propose the Discriminative Feature Enhancement (DFE) module. Unlike conventional attention mechanisms that assign uniform importance across spatial tokens, DFE introduces a novel dual-pathway architecture that disentangles foreground and background representations by learning separate feature subspaces for joint-relevant and irrelevant regions: a *positive pathway* that captures dependencies within foreground regions (*i.e.,* the person), and a *negative pathway* that models the context of the background (Ye et al. 2025). By adaptively subtracting the background signal from the foreground signal, the DFE effectively enhances joint-relevant representations and suppresses interference from irrelevant regions, thereby mitigating the effects of the spatial uniformity problem.

Specifically, given the fused features $\mathcal{F}_{st}$ from feature integration operation, we first project them into positive and negative query-key pairs (Q_+, K_+ and Q_-, K_-), alongside a standard value matrix $V \in \mathbb{R}^{N\times D_s}$:

$$[Q_+, Q_-] = \mathcal{F}_{st}W_Q, [K_+, K_-] = \mathcal{F}_{st}W_K, V = \mathcal{F}_{st}W_V, \tag{1}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{D\times D_s}$ denote learnable projection matrices, $Q_+, Q_-, K_+, K_- \in \mathbb{R}^{N\times \frac{D_s}{2}}$ are positive and negative queries/keys. The output of the DFE module is then formulated as the difference between a positive and a weighted negative pathway:

$$\mathcal{F}_{st}^1 = \underbrace{\Phi(\frac{Q_+K_+^T}{\sqrt{\frac{D_s}{2}}})V}_{\text{positive pathway}} - \lambda \odot \underbrace{\Phi(\frac{Q_-K_-^T}{\sqrt{\frac{D_s}{2}}})V}_{\text{negative pathway}}, \tag{2}$$

where Φ is the softmax operation, $\odot$ represents element-wise multiplication. For clarity, we omit the multi-head setting. The key to this mechanism is the adaptive, channel-wise balancing factor λ, which dynamically controls the degree of background suppression:

$$\lambda = \exp(\lambda_{q1}\lambda_{k1}) - \exp(\lambda_{q2}\lambda_{k2}) + \lambda_{\text{init}}, \tag{3}$$

where $\lambda_{q1}, \lambda_{q2}, \lambda_{k1}, \lambda_{k2}$ are learnable parameters and λ_{init} is a randomly initialized constant. This differential mechanism enables the DFE module to output an enhanced feature $\mathcal{F}_{st}^1$, where salient joint features are enhanced and background noise is significantly suppressed.

Feature Cluster Grouping

While the DFE module enhances joint-relevant features, modeling the intricate dependencies between joints efficiently remains a challenge. Applying global self-attention directly across all tokens treats them with uniform granularity, potentially struggling to capture the structured, hierarchical dependencies between different body parts. To address this, the Feature Cluster Grouping (FCG) module organizes the enhanced features into a small number of semantically coherent groups. Specifically, we employ the k-nearest neighbor density peak clustering (DPC-KNN) for discriminative token grouping (Du, Ding, and Jia 2016), which rests on two fundamental premises: (i) cluster centers have a higher local density than their neighbors, and (ii) distinct centers lie at relatively large distances. From these premises, two complementary measures emerge for each feature token: its local density ρ and its relative distance δ.

Formally, for each feature token f_i in the enhanced feature $\mathcal{F}_{st}^1$, we compute its local density ρ_i and its minimum distance δ_i to a higher-density token. The density ρ_i is calculated based on the average cosine similarity to its k nearest neighbors $\mathcal{N}_k(i)$:

$$\rho_i = \exp\{-\frac{1}{\mathrm{k}}\sum_{\mathrm{j_i}\in\mathcal{N}_\mathrm{k}(\mathrm{i})}(\mathrm{D_{ij}})^2\}, \tag{4}$$

where $D_{ij} = \|f_i - f_j\|_2$ represents the Euclidean distance between data points. The distance δ_i is given by:

$$\delta_i = \begin{cases}\min_{j:\rho_j>\rho_i}(D_{ij}) & \text{if } \exists\ j \text{ s.t. } \rho_j > \rho_i \\ \max_j(D_{ij}) & \text{otherwise}\end{cases}. \tag{5}$$

Tokens that are simultaneously characterized by high local density ρ_i (*i.e.*, they are surrounded by neighbors) and a large distance δ_i (*i.e.*, they are far from any denser points) are natural cluster centers. To formalize this, we compute a score $\gamma_i = \rho_i \times \delta_i$ for each token. The set of M cluster centers, $\mathcal{C} = \{c_1, \ldots, c_M\}$, is formed by selecting the tokens with the top-M highest scores γ.

Once centers are identified, all other tokens are assigned to the nearest cluster center, forming M clusters $\{C_1, \ldots, C_M\}$. Instead of simply averaging all tokens within a cluster, which could dilute the representation, we propose merging them based on learned importance scores.

For each cluster C_m, we first predict a scalar importance score s_j for every constituent token $f_j \in C_m$. These scores are then normalized across the cluster using the softmax function to produce attention-like weights $w_j = \frac{\exp(s_j)}{\sum_{f_k\in C_m}\exp(s_k)}$. The final representative feature for the cluster is computed as the weighted sum of its tokens: $\hat{f}_m = \sum_{f_j\in C_m} w_j f_j$.

This merging process adaptively aggregates information, giving more weight to important tokens and producing a compact set of M high-level features $\mathcal{F}_c = \{\hat{f}_1, \ldots, \hat{f}_M\}$.

To further enrich the semantic cluster features $\mathcal{F}_c$, we propose a semantic cluster attention. This module utilizes $\mathcal{F}_c$ to query the original feature $\mathcal{F}_{st}^1$, which serves as the key-value source. The operation is formulated as:

$$\mathcal{F}_{st}^2 = \Phi(\frac{(\mathcal{F}_cW_Q^C)(\mathcal{F}_{st}^1W_K^C)^T}{\sqrt{d_k}})(\mathcal{F}_{st}^1W_V^C), \tag{6}$$

where $Q^C = \mathcal{F}_cW_Q^C$, $K^C = \mathcal{F}_{st}^1W_K^C$, $V^C = \mathcal{F}_{st}^1W_V^C$, and Φ denotes the softmax operation. The output $\mathcal{F}_{st}^2$ thus provides a semantically compact representation.

Deformable Spatiotemporal Fusion

Another primary challenge in video-level detection is the *temporal rigidity problem*. Standard spatiotemporal attention mechanisms, which rely on fixed sampling locations, often fail to adapt to the non-rigid motion of human joints across consecutive frames. This rigidity causes erroneous feature aggregation when joints undergo large, non-rigid displacements between frames. To resolve this, we propose the Deformable Spatiotemporal Fusion (DSF) module. Instead of attending to fixed locations within the fused spatiotemporal feature space, the DSF module learns to dynamically predict sampling offsets within the fused spatiotemporal feature space. It generates these offsets for each query token, allowing it to adaptively gather and aggregate features from the most relevant spatiotemporal locations, thereby overcoming temporal deformation insensitivity and enabling robust alignment during large, non-rigid motion.

At the core of the DSF module is a novel deformation-aware offset prediction mechanism that dynamically identifies relevant features across the spatiotemporal features, enabling flexible and accurate modeling of non-rigid joint movements. Given the spatially refined features $\mathcal{F}_{st}^2$, we first project them into a query space using a learnable weight matrix W_Q^D and then an offset prediction network θ_{offset} takes this query map Q^D as input to predict a spatiotemporal offset map ΔP:

$$Q^D = \mathcal{F}_{st}^2 W_Q^D, \;\; \Delta P = \theta_{offset}(Q^D). \tag{7}$$

Using the predicted offsets, we generate a deformed feature $\hat{\mathcal{F}}_{st}^2$. This is accomplished by a linear interpolation operation φ that retrieves features from the original map $\mathcal{F}_{st}^2$ at the offset-adjusted locations $P + \Delta P$:

$$\hat{\mathcal{F}}_{st}^2 = \varphi(\mathcal{F}_{st}^2, P + \Delta P). \tag{8}$$

$\hat{\mathcal{F}}_{st}^2$ can be viewed as a dynamically-aligned representation, where features corresponding to the same body parts are dynamically aligned across the temporal dimension.

Finally, we employ a cross-attention mechanism where the initial features $\mathcal{F}_{st}^2$ act as queries Q^D to selectively aggregate important temporal information from the deformed features $\hat{\mathcal{F}}_{st}^2$ (keys K^D and values V^D):

$$\mathcal{F}_{st}^3 = \Phi(\frac{Q^D(\hat{\mathcal{F}}_{st}^2 W_K^D)^T}{\sqrt{d_k}})(\hat{\mathcal{F}}_{st}^2 W_V^D), \tag{9}$$

where $K^D = \hat{\mathcal{F}}_{st}^2 W_K^D$, $V^D = \hat{\mathcal{F}}_{st}^2 W_V^D$, and Φ denotes the softmax operation. This deformable fusion ensures robust temporal modeling that is inherently sensitive to complex motion, producing the final feature representation $\mathcal{F}_{st}^3$.

Heatmap Generation

After the progressive refinement stages, we obtain a comprehensive feature representation $\mathcal{F}_{st}^3$, which is rich in discriminative spatial detail and robust to temporal deformation. These features are then fed to a detection head to generate the keypoint heatmaps $\mathcal{H}_t = \{H_1, H_2, ..., H_K\}$, where each heatmap signifies the likelihood distribution for a specific anatomical keypoint. We apply the standard pose estimation loss to supervise the training:

$$\mathcal{L}_H = \|\mathcal{H}_t - \mathcal{G}_t\|_2^2, \tag{10}$$

where $\mathcal{H}_t$ and $\mathcal{G}_t$ are the estimated pose heatmap and the ground truth, respectively.

Experiments

Experimental Settings

Datasets. We evaluate the performance of the proposed PSTPose method across three widely-used benchmarks: PoseTrack2017 (Iqbal, Milan, and Gall 2017), PoseTrack2018 (Andriluka et al. 2018), PoseTrack21 (Doering et al. 2022). **PoseTrack2017:** This dataset includes 300 video sequences and a total of 80,144 annotated human poses. Among them, 250 sequences are allocated for training, while the remaining 50 serve as the validation set. Each annotation comprises 15 human keypoints. Notably, the training data contains dense annotations for the central 30 frames of each clip, whereas validation sequences include pose annotations sampled every four frames. **PoseTrack2018:** This dataset contains 763 video sequences and 153,615 annotated poses. Of these, 593 sequences are designated for training and 170 for validation. In addition to 15 keypoint locations, each joint annotation is accompanied by a visibility flag, offering richer supervision. **PoseTrack21:** Built upon PoseTrack2018, this dataset further increases the annotation count to 177,164 poses. It emphasizes more difficult scenarios by focusing annotations on small-scale individuals and densely crowded scenes, enhancing the dataset's utility for evaluating robust pose estimation under real-world challenges.

Evaluation metrics. To evaluate model performance, we employ standard metrics tailored to each benchmark dataset. For the PoseTrack2017, PoseTrack2018, and PoseTrack21 datasets, we use mean Average Precision (mAP), which summarizes the overall performance by averaging the Average Precision (AP) for the alignment of predicted poses with ground-truth annotations across all joints.

Implementation details. All experiments are performed using the PyTorch framework on two NVIDIA RTX 4090 GPUs for 20 epochs. We adopt ViT (Dosovitskiy et al. 2020) as the feature extractor to generate the initial features, with the input resolution fixed at 256×192. During training, four augmentation strategies are applied: (1) random rotation within the range of $[-45°, 45°]$, (2) random scaling between $[0.65, 1.35]$, (3) random truncation, and (4) horizontal flipping. The temporal sampling interval ξ is set to 1. We utilize AdamW as the optimizer, starting with a learning rate of 2e-4. The learning rate decays progressively to 2e-5, 2e-6, and 2e-7 at the 6-th, 12-th, and 16-th epochs, respectively.

Comparison with State-of-the-art Methods

Results on the PoseTrack2017 and PoseTrack2018 Datasets. We begin our evaluation on the PoseTrack2017 and PoseTrack2018 datasets. As shown in Tables 1 and 2, PSTPose achieves new state-of-the-art results with 88.0

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Simple (Xiao, Wu, and Wei 2018)	81.7	83.4	80.0	72.4	75.3	74.8	67.1	76.7
HRNet (Sun et al. 2019)	82.1	83.6	80.4	73.3	75.5	75.3	68.5	77.3
PoseWarper (Bertasius et al. 2019)	81.4	88.3	83.9	78.0	82.4	80.5	73.6	81.2
Dynamic-GNN (Yang et al. 2021)	88.4	88.4	82.0	74.5	79.1	78.3	73.1	81.1
DCPose (Liu et al. 2021)	88.0	88.7	84.1	78.4	83.0	81.4	74.2	82.8
FAMI-Pose (Liu et al. 2022a)	89.6	90.1	86.3	80.0	84.6	83.4	77.0	84.8
SLT-Pose (Gai et al. 2023)	88.9	89.7	85.6	79.5	84.2	83.1	75.8	84.2
HANet (Jin et al. 2023)	90.0	90.0	85.0	78.8	83.1	82.1	77.1	84.2
TDMI (Feng et al. 2023a)	90.0	91.1	87.1	81.4	85.2	84.5	78.5	85.7
DiffPose(Feng et al. 2023b)	89.0	91.2	87.4	83.5	85.5	87.2	80.2	86.4
DSTA (He and Yang 2024)	89.3	90.6	87.3	82.6	84.5	85.1	77.8	85.6
JMPose (Wu et al. 2024a)	90.7	91.6	87.8	82.1	85.9	85.3	79.2	86.4
FDMC (Feng et al. 2025)	89.9	90.6	86.4	80.9	84.7	83.9	76.9	85.4
TPSD-MS (Zhang et al. 2025c)	**91.1**	91.5	87.6	82.1	85.9	85.0	79.4	86.4
TPSD-ViT (Zhang et al. 2025c)	90.7	91.1	87.9	83.6	85.3	86.3	80.0	86.7
PSTPose (Ours)	90.0	**91.7**	**89.7**	**86.0**	**89.9**	**87.8**	**80.4**	**88.0**

Table 1: Quantitative results on the **PoseTrack2017** dataset. For this and the subsequent tables, *Sho., Elb., Wri., and Ank.* stand for shoulder, elbow, wrist, and ankle, respectively.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PoseWarper (Bertasius et al. 2019)	79.9	86.3	82.4	77.5	79.8	78.8	73.2	79.7
Dynamic-GNN (Yang et al. 2021)	80.6	84.5	80.6	74.4	75.0	76.7	71.8	77.9
DCPose (Liu et al. 2021)	84.0	86.6	82.7	78.0	80.4	79.3	73.8	80.9
FAMI-Pose (Liu et al. 2022a)	85.5	87.7	84.2	79.2	81.4	81.1	74.9	82.2
SLT-Pose (Gai et al. 2023)	84.3	87.5	83.5	78.5	80.9	80.2	74.4	81.5
HANet (Jin et al. 2023)	86.1	88.5	84.1	78.7	79.0	80.3	77.4	82.3
DiffPose (Feng et al. 2023b)	85.0	87.7	84.3	81.5	81.4	82.9	77.6	83.0
TDMI (Feng et al. 2023a)	86.2	88.7	85.4	80.6	82.4	82.1	77.5	83.5
DSTA (He and Yang 2024)	85.9	88.8	85.0	81.1	81.5	83.0	77.4	83.4
JMPose (Wu et al. 2024a)	86.6	88.7	**86.0**	**81.6**	83.3	83.2	78.2	84.1
FDMC (Feng et al. 2025)	86.6	88.9	84.7	79.9	82.4	82.7	77.4	83.3
TPSD-MS (Zhang et al. 2025c)	**87.0**	89.0	85.6	81.5	83.4	82.4	78.2	84.1
TPSD-ViT (Zhang et al. 2025c)	86.9	89.0	**86.0**	**81.6**	83.3	83.3	78.0	84.2
PSTPose (Ours)	86.6	**89.2**	85.8	81.3	**84.9**	**84.4**	**81.6**	**84.9**

Table 2: Quantitative comparison with SOTA methods on the PoseTrack2018 (Andriluka et al. 2018) dataset.

mAP and 84.9 mAP, respectively. Notably, its most substantial gains are on challenging distal joints. For instance, on PoseTrack2017, PSTPose outperforms DiffPose (Feng et al. 2023b) on the Elbow (↑ 2.3 AP) and Wrist (↑ 2.5 AP), while on PoseTrack2018, it surpasses the TDMI (Feng et al. 2023a) on the Ankle (↑ 4.1 AP). These joints are notoriously difficult to identify due to large inter-frame displacements—a problem we identify as temporal rigidity. Our Deformable Spatiotemporal Fusion (DSF) module directly addresses this by learning deformation-aware sampling offsets to dynamically adjust its sampling location. The significant performance increase on these high-velocity joints provides compelling evidence that DSF effectively resolves temporal rigidity, leading to more robust spatiotemporal representations.

Results on the PoseTrack21 dataset. The PoseTrack21 dataset, featuring small-scale subjects and crowded scenes, serves as a stringent test for robustness against visual noise and occlusions. On this challenging benchmark, as tabulated in Table 3, PSTPose achieves a new state-of-the-art with 84.8 AP. This strong performance can be attributed to the core components of our model. Our Discriminative Feature Enhancement (DFE) module is designed to suppress the background clutter inherent in crowded scenes, while the Deformable Spatiotemporal Fusion (DSF) module maintains alignment for limbs undergoing erratic motion. The effectiveness of this combined approach is reflected in substantial gains over TDMI (Feng et al. 2023a) on often-occluded joints, including the Shoulder (↑ 1.6 AP) and Ankle (↑ 2.5 AP), demonstrating its capability to maintain performance in complex, real-world conditions.

Qualitative Comparison. We provide a qualitative comparison against state-of-the-art methods in Fig. 3. The visualizations on the PoseTrack dataset demonstrate that while existing methods like DCPose (Liu et al. 2021) and TDMI (Feng et al. 2023a) perform well in simple cases, they often fail in challenging scenes involving occlusions, rapid motion, or video defocus. In contrast, our PSTPose consistently maintains robustness, delivering more accurate pose estimations in these complex situations.

Ablation Analysis

Ablation Study on PSTPose Components. We conduct an ablation study on the PoseTrack2017 dataset to analyze the individual contributions of our three modules: Discriminative Feature Enhancement (DFE), Feature Cluster Grouping (FCG), and Deformable Spatiotemporal Fusion (DSF). As shown in Table 4, our baseline, which employs a stan-

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
HRNet (Sun et al. 2019)	81.5	83.2	81.1	75.4	79.2	77.8	71.9	78.8
PoseWarper (Bertasius et al. 2019)	82.3	84.0	82.2	75.5	80.7	78.7	71.6	79.5
DCPose (Liu et al. 2021)	83.2	84.7	82.3	78.1	80.3	79.2	73.5	80.5
FAMI-Pose (Liu et al. 2022a)	83.3	85.4	82.9	78.6	81.3	80.5	75.3	81.2
DiffPose (Feng et al. 2023b)	84.7	85.6	83.6	80.8	81.4	83.5	80.0	82.9
TDMI (Feng et al. 2023a)	85.8	87.5	85.1	81.2	83.5	82.4	77.9	83.5
SLTPose (Gai et al. 2023)	83.3	85.1	82.7	78.5	81.3	80.8	75.6	81.3
JMPose (Wu et al. 2024a)	85.8	88.1	**85.7**	**82.5**	**84.1**	83.1	78.5	84.0
FDMC (Feng et al. 2025)	84.1	85.5	83.3	79.7	81.3	82.1	77.8	82.2
TPSD-MS (Zhang et al. 2025c)	87.0	87.6	85.3	81.5	84.0	82.6	77.9	83.9
TPSD-ViT (Zhang et al. 2025c)	**87.7**	88.0	85.0	81.7	83.4	82.8	78.3	84.1
PSTPose (Ours)	86.7	**89.1**	85.6	82.1	**84.1**	**84.7**	**80.4**	**84.8**

Table 3: Quantitative comparison with SOTA methods on the PoseTrack21 (Doering et al. 2022) dataset.

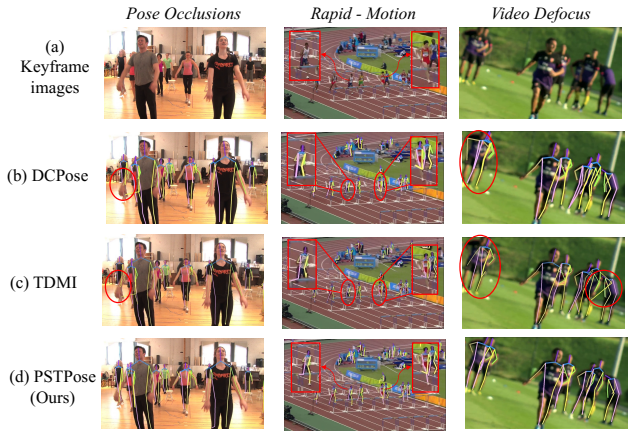

Figure 3: Qualitative results showcasing PSTPose's superior robustness against common failure modes. We compare our method (d) to baselines (b, c) (Liu et al. 2021; Feng et al. 2023a) on challenging keyframes (a), with failures circled in red. For pose occlusions (left), existing methods fail to localize heavily occluded ankles. For rapid-motion (middle), existing methods produce incorrect ankle estimations due to motion blur. For video defocus (right), existing methods miss keypoints like shoulders and ankles in degraded frames.

dard attention mechanism, achieves only 84.2 mAP. This result not only underperforms our full model but also lags behind the DSTA (He and Yang 2024), confirming the inadequacy of a naive transformer approach. For the first setting (a), removing the DFE module results in a performance drop of 1.5 AP to 86.5 AP. This decline highlights the importance of DFE in mitigating spatial indiscriminate focus. For setting (b), removing the FCG module leads to the most substantial performance degradation among the three components, with a drop of 1.8 mAP to 86.2 mAP. This underscores the critical role of FCG in organizing tokens into semantically coherent clusters. Subsequently, for setting (c), removing the DSF module causes a 1.6 mAP drop to 86.4 mAP. This validates the necessity of its motion-aware sampling, proving essential for handling non-rigid joint movements. Finally, our full PSTPose model, which integrates all three modules, achieves the best performance of 88.0 mAP. This result demonstrates that these components are not only

Method	DFE	FCG	DSF	Mean	Declines
DSTA (He and Yang 2024)	-	-	-	85.6	2.4 (↓)
Baseline	✗	✗	✗	84.2	3.8 (↓)
(a) PSTPose w/o DFE	✗	✓	✓	86.5	1.5 (↓)
(b) PSTPose w/o FCG	✓	✗	✓	86.2	1.8 (↓)
(c) PSTPose w/o DSF	✓	✓	✗	86.4	1.6 (↓)
PSTPose (Ours)	✓	✓	✓	88.0	-

Table 4: Ablation of different designs in PSTPose. "w/o X" refers to removing X module.

individually effective but also complementary, working synergistically to achieve a progressive spatiotemporal refinement for video-based human pose estimation.

Conclusion

In this paper, we aim to address two critical yet often overlooked limitations in video-based human pose estimation: *spatial uniformity* and *temporal rigidity*. We introduce PSTPose, a novel progressive spatiotemporal refinement framework designed to overcome these challenges. By first employing a discriminative feature enhancement module and a feature cluster grouping module, PSTPose mitigates spatial noise and forms compact feature representations. A deformable spatiotemporal fusion mechanism further adaptively integrates corresponding features across consecutive frames using learned, deformation-aware offsets. Extensive experiments on three challenging benchmarks validate the superiority of our approach. The success of PSTPose underscores the value of progressive refinement for spatiotemporal modeling and offers a promising direction for future research in dynamic human-centric video analysis.

Acknowledgments

This work is supported by the National Key R&D Program of China (No. 2024YFB3311605), National Natural Science Foundation of China (No. 62276112, No. 62372402) and the Key R&D Program of Zhejiang Province (No. 2025C01084).

References

Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; and Schiele, B. 2018. Posetrack: A benchmark for human pose estimation and tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5167–5176.

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 3686–3693.

Andriluka, M.; Roth, S.; and Schiele, B. 2012. Discriminative appearance models for pictorial structures. *International journal of computer vision*, 99: 259–280.

Artacho, B.; and Savakis, A. 2020. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7035–7044.

Bertasius, G.; Feichtenhofer, C.; Tran, D.; Shi, J.; and Torresani, L. 2019. Learning temporal pose estimation from sparsely-labeled videos. *Advances in neural information processing systems*, 32.

Chen, H.; Wu, S.; Wang, Z.; Yin, Y.; Jiao, Y.; Lyu, Y.; and Liu, Z. 2025. Causal-Inspired Multitask Learning for Video-Based Human Pose Estimation. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 2052–2060. AAAI Press.

Doering, A.; Chen, D.; Zhang, S.; Schiele, B.; and Gall, J. 2022. Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20963–20972.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, M.; Ding, S.; and Jia, H. 2016. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99: 135–145.

Feng, R.; Gao, Y.; Ma, X.; Tse, T. H. E.; and Chang, H. J. 2023a. Mutual information-based temporal difference learning for human pose estimation in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17131–17141.

Feng, R.; Gao, Y.; Tse, T. H. E.; Ma, X.; and Chang, H. J. 2023b. Diffpose: Spatiotemporal diffusion model for video-based human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14861–14872.

Feng, R.; Tse, T. H. E.; Chen, H.; Chang, H. J.; Zhong, H.; and Gao, Y. 2025. Video-Based Human Pose Estimation via Feature Decoupling and Multi-Hypothesis Calibration. *Pattern Recognition*, 112193.

Fu, T.; Xu, X.; Xu, W.; Chen, J.; Ren, R.; Deng, B.; Zhao, X.; Cao, J.; and Cao, X. 2025. Two Heads are Better than One: Distilling Large Language Model Features Into Small Models with Feature Decomposition and Mixture. arXiv:2511.07110.

Gai, D.; Feng, R.; Min, W.; Yang, X.; Su, P.; Wang, Q.; and Han, Q. 2023. Spatiotemporal learning transformer for video-based human pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9): 4564–4576.

He, J.; and Yang, W. 2024. Video-based human pose regression via decoupled space-time aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1022–1031.

Hernández, Ó. G.; Morell, V.; Ramon, J. L.; and Jara, C. A. 2021. Human pose detection for robotic-assisted and rehabilitation environments. *Applied Sciences*, 11(9): 4183.

Iqbal, U.; Milan, A.; and Gall, J. 2017. Posetrack: Joint multi-person pose estimation and tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011–2020.

Jiao, Y.; Wang, Z.; Liu, Z.; Fan, S.; Wu, S.; Wu, Z.; and Xu, Z. 2025. Optimizing Human Pose Estimation Through Focused Human and Joint Regions. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 4102–4110. AAAI Press.

Jin, K.-M.; Lim, B.-S.; Lee, G.-H.; Kang, T.-K.; and Lee, S.-W. 2023. Kinematic-aware hierarchical attention network for human pose estimation in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5725–5734.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, B.; Dong, H.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2025a. Exploring Efficient Open-Vocabulary Segmentation in the Remote Sensing. *arXiv preprint arXiv:2509.12040*.

Li, B.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2025b. Stitchfusion: Weaving any visual modalities to enhance multimodal semantic segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1308–1317.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, J.; Huang, X.; Hao, Z.; Zhu, R.; Zheng, Q.; Wang, S.; Chen, S.; Liu, C.; Huang, L.; Tao, J.; and Fan, Y. 2025. MAS-ISP: A Proxy-Free Online Hyperparameter Optimization Framework for ISP Hardware System. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, 1–7.

Liu, J.; Liu, Z.; Huang, X.; Zhu, R.; Zheng, Q.; Hao, Z.; Liu, T.; Tao, J.; and Fan, Y. 2024. Auto-ISP: An Efficient Real-Time Automatic Hyperparameter Optimization Framework for ISP Hardware System. In *Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC)*, DAC '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706011.

Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; and Wang, X. 2021. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 525–534.

Liu, Z.; Feng, R.; Chen, H.; Wu, S.; Gao, Y.; Gao, Y.; and Wang, X. 2022a. Temporal feature alignment and mutual information maximization for video-based human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11006–11016.

Liu, Z.; Wu, S.; Xu, C.; Wang, X.; Zhu, L.; Wu, S.; and Feng, F. 2022b. Copy Motion From One to Another: Fake Motion Video Generation. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 1223–1231. ijcai.org.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5693–5703.

Tan, D.; Chen, H.; Tian, W.; and Xiong, L. 2024. Diffusionregpose: Enhancing multi-person pose estimation using a diffusion-based end-to-end regression approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2230–2239.

Tang, H.; Li, Z.; Peng, Z.; and Tang, J. 2020. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *Proceedings of the 28th ACM international conference on multimedia*, 610–618.

Tang, H.; Liu, J.; Yan, S.; Yan, R.; Li, Z.; and Tang, J. 2023. M3net: multi-view encoding, matching, and fusion for few-shot fine-grained action recognition. In *Proceedings of the 31st ACM international conference on multimedia*, 1719–1728.

Tang, L.; Huang, K.; Chen, C.; Yuan, Y.; Li, C.; Tu, X.; Ding, X.; and Huang, Y. 2025. Dissecting generalized category discovery: Multiplex consensus under self-deconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 297–307.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Z.; Fan, S.; Liu, Z.; Wu, Z.; Wu, S.; and Jiao, Y. 2025. Multi-Grained Feature Pruning for Video-Based Human Pose Estimation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, 1–5. IEEE.

Weiguang, L.; Dong, L.; and Lu, W. 2023. Survey of Deformable Convolutional Networks. *Journal of Frontiers of Computer Science & Technology*, 17(7).

Wu, S.; Chen, H.; Yin, Y.; Hu, S.; Feng, R.; Jiao, Y.; Yang, Z.; and Liu, Z. 2024a. Joint-Motion Mutual Learning for Pose Estimation in Video. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8962–8971.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024b. Do as I Do: Pose Guided Human Motion Copy. *IEEE Trans. Dependable Secur. Comput.*, 21(6): 5293–5307.

Wu, S.; Zhang, H.; Liu, Z.; Chen, H.; and Jiao, Y. 2025. Enhancing Human Pose Estimation in Internet of Things via Diffusion Generative Models. *IEEE Internet Things J.*, 12(10): 13556–13567.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, 466–481.

Xu, H.; Ke, X.; Wu, H.; Xu, R.; Li, Y.; Xu, P.; and Guo, W. 2025. Dancefix: An exploration in group dance neatness assessment through fixing abnormal challenges of human pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8869–8877.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2023. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2): 1212–1230.

Yang, Y.; Ren, Z.; Li, H.; Zhou, C.; Wang, X.; and Hua, G. 2021. Learning dynamics via graph neural networks for human pose estimation and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8074–8084.

Ye, T.; Dong, L.; Xia, Y.; Sun, Y.; Zhu, Y.; Huang, G.; and Wei, F. 2025. Differential Transformer. In *The Thirteenth International Conference on Learning Representations*.

Yuan, Y.; Tang, L.; Chen, Y.; Chen, C.; Huang, Y.; and Ding, X. 2025. ASGS: Single-Domain Generalizable Open-Set Object Detection via Adaptive Subgraph Searching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20911–20921.

Zhang, R.; Huang, Y.; Cao, Y.; and Wang, H. 2025a. MoleBridge: Synthetic Space Projecting with Discrete Markov Bridges. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhang, R.; Huang, Y.; Lou, Y.; Xin, Y.; Chen, H.; Cao, Y.; and Wang, H. 2025b. Exploit Your Latents: Coarse-Grained Protein Backmapping with Latent Diffusion Models. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 1111–1119. AAAI Press.

Zhang, R.; Lin, D.; Wang, X.; Liu, R.; Sheng, B.; Baciu, G.; Chen, C. P.; and Li, P. 2025c. Temporal-Interim Pose Synthesis and Distillation for Dynamic Human Pose Estimation. *IEEE Transactions on Neural Networks and Learning Systems*.