# DanceFix: An Exploration in Group Dance Neatness Assessment Through Fixing Abnormal Challenges of Human Pose

**Huangbiao Xu[1,2], Xiao Ke[1,2*], Huanqi Wu[1,2], Rui Xu[1,2], Yuezhou Li[1,2], Peirong Xu[1,2], Wenzhong Guo[1,2]**

[1]Fujian Provincial Key Laboratory of Networking Computing and Intelligent Information Processing, College of Computer and Data Science, Fuzhou University, Fuzhou 350116, China
[2]Engineering Research Center of Big Data Intelligence, Ministry of Education, Fuzhou 350116, China
{kex, guowenzhong}@fzu.edu.cn,{huangbiaoxu.chn, wuhuangi135, xurui.ryan.chn, liyuezhou.cm}@gmail.com

## Abstract

The fair and objective assessment of performances and competitions is a common pursuit and challenge in human society. The application of computer vision technology offers hope for this purpose, but it still faces obstacles such as occlusion and motion blur. To address these hindrances, our DanceFix proposes a bidirectional spatial-temporal context optical flow correction (BOFC) method. This approach leverages the consistency and complementarity of motion information between two modalities: optical flow, which excels at pixel capture, and lightweight skeleton data. It enables the extraction of pixel-level motion changes and the correction of abnormal skeleton data. Furthermore, we propose a part-level dance dataset (Dancer Parts) and part-level motion feature extraction based on task decoupling (PETD). This aims to decouple complex whole-body parts tracking into fine-grained limb-level motion extraction, enhancing the confidence of temporal information and the accuracy of correction for abnormal data. Finally, we present the DNV dataset, which simulates fully neat group dance scenes and provides reliable labels and validation methods for the newly introduced group dance neatness assessment (GDNA). To the best of our knowledge, this is the first work to develop quantitative criteria for assessing limb and joint neatness in group dance. We conduct experiments on DNV and video-based public JHMDB datasets. Our method effectively corrects abnormal skeleton points, flexibly embeds, and improves the accuracy of existing pose estimation algorithms.

## Introduction

Human action recognition and assessment is a significant visual task that has found widespread use in real-world scenarios. In recent years, the application of action recognition in sports and dance entertainment (Guo, Zhao, and Li 2021; Lei et al. 2023) has gained increasing attention, and virtual games (Wu et al. 2021; Hu et al. 2021) based on human interaction (*e.g.*, dance, sports) have also extended the research in action assessment. *In this paper, we focus on the challenges in action understanding, taking dance action assessment as an example and proposing targeted solutions.* Specifically, we introduce the Group Dance Neatness Assessment (GDNA) task, which aims to assess how neatly and consistently dancers move with each other in a group dance—a critical factor in group dance performances.

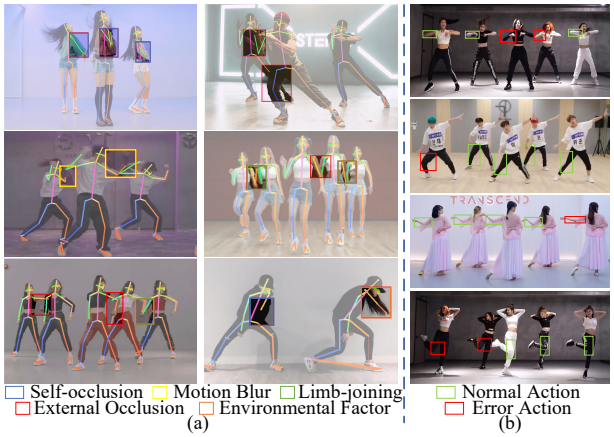



Figure 1: (a) Common abnormal challenges in dance scenes. (b) Group dancing in real life is difficult, even impossible, to achieve complete neatness. (Best viewed in color.)

Various modalities are available to study human action, such as RGB video (Li et al. 2022; Wen et al. 2023), optical flow (Wang et al. 2023; Radevski et al. 2023), and skeleton data (Zhou, Liu, and Wang 2023; Lin, Zhang, and Liu 2023), but these works study only one modality, ignoring the information consistency among modalities. The current commonly used skeleton data is lightweight structural data with high computational efficiency and robustness, which is not easily affected by the background. However, the acquisition of skeleton data in practice depends on the accuracy of pose estimation algorithms. The human dance actions studied in this paper are more variable with more complex action information, which are more prone to abnormalities such as self-occlusion, external occlusion, and motion blur leading to inaccurate skeleton data (Fig. 1 (a)). This low-credibility skeleton data reduces the accuracy of the dance neatness assessment. In addition, we also find inconsistent motion properties of body parts during actions. The human body is a flexible articulated structure, with interconnected limbs that often exhibit different motion properties. It is difficult to achieve uniform changes in all body parts during movement. Such differences are even more obvious in dance (*e.g.*, hand

*Corresponding author.

motions are more flexible and strenuous than legs).

Besides, dance action assessment is not only entertaining and interesting but also has important practical value. Using machines to assess human actions in performances and competitions can effectively ensure fairness and eliminate individual subjective effects. However, the lack of reliable labeling data and unified quantitative standards for many scenes has made it difficult to achieve accurate quality assessment of action sequences, hindering the effective dissemination of research results and concepts. Moreover, most of the dance scenes that look neat at a glance are not neat after a closer look, and the dance scenes that are 100% neat in reality are difficult to find and do not even exist (Fig. 1 (b)), making it difficult to define the exact neatness of dance.

Based on the above analysis, we summarize the main challenges for dance action assessment as follows: (1) How to solve the shortcomings of a single modality and enhance the accuracy of action information, particularly under abnormal cases like occlusion and motion blur? (2) How to eliminate the effects of the non-uniform motion of body parts and adaptively extract motion information from each part with various degrees of intensity? (3) How to obtain reliable labeling data and quantify criteria for group dance?

Although existing keypoint detection or correction methods (Moon, Chang, and Lee 2019; Zeng et al. 2022) have achieved great progress, they are not adapted to the complexity and variability of dance scenes. Therefore, we aim to explore methods that adapt to complex actions like dance. To address these challenges, we propose a novel abnormal skeleton data correction method called bidirectional spatial-temporal context optical flow correction (BOFC). We exploit the fact that the motion information expressed across modalities is consistent with each other. Among them, optical flow can calculate the pixel changes before and after the appearance of anomalies, which has the potential to solve anomalies such as occlusion (Poux et al. 2021; Feng et al. 2023). BOFC first extracts the reliable pre- and post-temporal context optical flow information of abnormal frames. Using this, it calculates the pixel motion of skeleton points and compensates for the defects of skeleton data, ultimately combining fine-grained optical flow and lightweight skeleton data to maximize benefits. By combining the motion consistency of modalities, skeleton data, and optical flow, we can correct abnormal skeleton data effectively.

To extract motion properties of body parts, we decouple the complex human body into limb parts inspired by task decoupling and focus on the "instance-level" information (Yang et al. 2020) of each part. We construct a part-level dance dataset (Dancer Parts) to address the motion difference of body parts and propose a part-level motion feature extraction based on task decoupling (PETD). PETD extracts the motion properties of each part, including candidate skeleton point regions and motion changes, and uses this information to detect the before and after frames of abnormal frames. This improves the reliability of optical flow information of BOFC, correction, and dance neatness assessment.

To improve the reusability and feasibility of DanceFix, we construct a dataset that simulates fully neat virtual dance scenes, called Dancing-Neatly-in-Virtual (DNV). We first validate the feasibility of automatic quantitative group dance neatness assessment and the validity of DanceFix on the DNV. To this end, we propose five new metrics to comprehensively assess GDNA, including three static: neatness score (NS) and relative mean per joint angle/position error (R-MPJAE/R-MPJPE) and two dynamic: relative mean per joint angle/position velocity error (R-MPJAVE/R-MPJPVE). We further validate the effectiveness of our method in improving pose accuracy on the publicly JHMDB dataset. The main contributions are summarized as:

- We propose a bidirectional spatial-temporal context optical flow correction (BOFC) to combat anomalies such as occlusions and blurs for correcting the skeleton data.
- We propose a part-level motion feature extraction based on task decoupling (PETD) and a corresponding part-level dataset (Dancer Parts) to extract the candidate regions of skeleton points and motion variations of body parts to obtain accurate spatial-temporal information.
- To explore and tackle GDNA, we construct a simulated fully neat dance dataset (DNV), which develops a unified criterion for automatically quantifying dance neatness assessment. To the best of our knowledge, this is the first work to develop assessing the dance's neatness.
- We propose five new metrics (three static and two dynamic) for group dance. Extensive experiments illustrate that our DanceFix can effectively embed and improve the accuracy of existing pose estimation algorithms.

## Related Work

**Skeleton-Based Action Recognition.** The data-driven methods for feature extraction using deep learning gradually replace the traditional manual feature approach (Vemulapalli, Arrate, and Chellappa 2014; Dalal and Triggs 2005), and the main neural networks for skeleton-based action recognition include RNN (Zhu et al. 2016; Li et al. 2021), CNN (Ke et al. 2017; Xu et al. 2022), GCN (Chen et al. 2021; Xie et al. 2024), and Transformer (Wu et al. 2024).

However, pose estimation methods often perform poorly in scenes such as occlusion and motion blur, thus the inaccurate skeleton data limits these methods in applications. To solve this, our BOFC captures motion information before and after anomalies using optical flow estimation. We then correct abnormal skeleton data by using the motion consistency between multiple modalities to obtain high-quality skeleton data for skeleton-based action understanding.

**Human Keypoint Detection.** Most pose estimation works (Yu et al. 2021; Yang et al. 2023b; An et al. 2024) are performed mainly on single-frame images, which can only estimate frame-by-frame for video and often perform poorly in the face of occlusion and motion blur scenes common to video. Recently, more promising results have been obtained by modeling temporal features (Xu et al. 2021; Zhang et al. 2022a; Feng et al. 2023) based on video-based pose estimation to learn and exploit the temporal information of videos.

However, these methods only use one frame before and after to directly localize the current frame. We find that abnormal cases like occlusion often lead to inaccurate pose estimation in multiple continuous frames in dance and sports

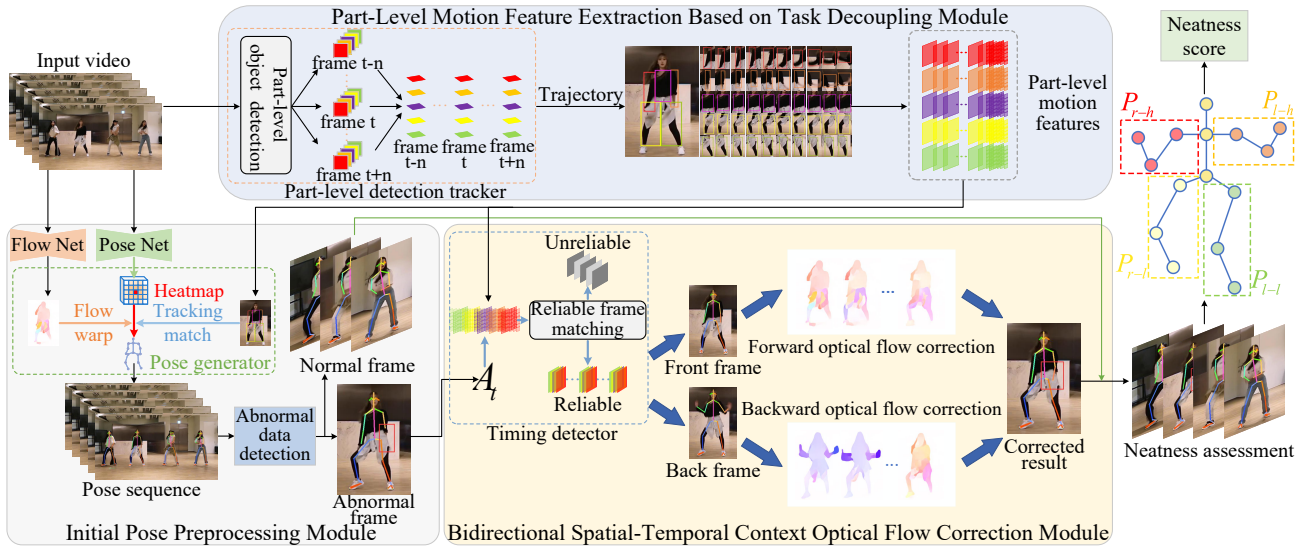



Figure 2: Block diagram of DanceFix. The input video sequence is pre-processed to estimate the initial pose and perform abnormal skeleton data detection, and the PETD module extracts part-level motion information to assist in the search for reliable spatial-temporal context information. Then, the BOFC module is used to correct the abnormal skeleton data based on the spatial-temporal context optical flow information, and finally to complete the group dance neatness assessment.

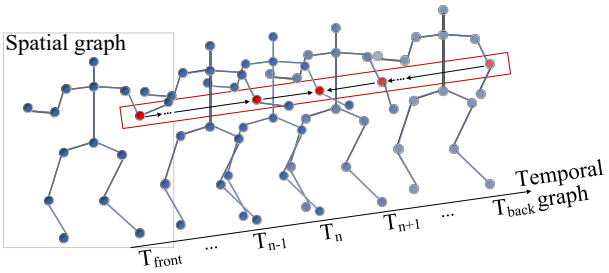



Figure 3: Block diagram of bidirectional spatial-temporal context optical flow correction (BOFC).

competitions. For this reason, our method explores multi-frame for extracting spatial-temporal features, excluding unreliable features, and uses optical flow estimation to capture motion consistency to correct abnormal skeleton data.

**Multiple Object Tracking.** The classical MOT methods (Zhang et al. 2022b; Li et al. 2024) are based on the TBD framework, using detectors to frame-wise detect bounding boxes and data association methods to identify target identities. JDE frameworks (Zhang et al. 2021; Yu et al. 2023) use a unified model to link target detection and tracking to improve tracking efficiency. Recently, Transformer have been applied to tracking (Chu et al. 2023; Wang et al. 2024).

However, these methods only track the whole body or head, which is less suitable to scenes with complex limb motions. Therefore, we train a part-level tracker to track body parts more finely and capture the motion differences of parts.

## Method

In this section, we describe DanceFix in detail. DanceFix is a generic pose correction network and an exploration of the application of CV in dance (GDNA), as shown in Fig. 2.

*Remark*: Although 2D pose has some limitations compared to 3D, 2D pose is currently more widely applicable in real-life scenes and easier to obtain. Many recent works (Ye et al. 2020; Wang et al. 2021a; Lei et al. 2023; Zhong and Demiris 2024) are using 2D poses to assess human motion and performance. In this work, we explore the novel study of group dance neatness assessment (GDNA) starting in 2D pose.

### Bidirectional Spatial-Temporal Context Optical Flow Correction (BOFC)

**Problem Formulation.** For an input video $I_v$ with $N$ frames, where $I_v = \{I_t\}_{t=1}^{N}, I_t \in \mathbb{R}^{H\times W\times 3}$, we first extract the pose heatmaps and optical flows using pre-trained pose and flow networks. As in (Pfister, Charles, and Zisserman 2015; Song et al. 2017), the confidence estimates of neighboring frames are aligned by shifting confidence values along the track directions through a flow-warping layer. The initial pose is obtained based on the warped heatmaps. We then apply the simplest abnormal data detection to the initial pose. Specifically, we consider three cases as unreliable abnormal skeleton data that require correction: low confidence ($c_{t,k}$), large variance in the confidence sequence ($C_t$) of a body in initial poses, and a high instantaneous motion rate of the skeleton position ($p_t^k$) between two continuous frames. The above process is written as:

$$D\left(\{p_t^k\}\right) = c_t^k < \tau_1 \; or \; Var\left(C_t\right) > \tau_2 \; or \; \left\|p_t^k - p_{t+1}^k\right\|_2 > \tau_3, \quad (1)$$

$$A_t = \bigcup\nolimits_k p_t^k = D(E(I_t)), \quad (2)$$

where $A_t$ is the set of abnormal skeleton data in the abnormal frame $I_t$, $p_t^k$ is $k$-$th$ abnormal skeleton point in $I_t$, and $E(\cdot)$, $D(\cdot)$ are pose estimation and abnormal data detection.

**Abnormal Skeleton Data Correction.** In practice, continuous anomalous frames are more common due to intense motions and complex changes in dance and sports. In this case, skeleton information from just the two frames before and after is often not credible. Therefore, we extend vision to detect anomalies frame by frame within a certain number of frames. We exclude abnormal skeleton data until we find the nearest reliable pre-sequence frame and post-sequence frame, and extract the motion's past and future information.

Next, for the skeleton points $p_i$ in the pre-sequence $i$-$th$ frame, we introduce forward optical flow information $\vec{f}_i$ between $i$-$th$ and $(i{+}1)$-$th$ frame, and calculate the skeleton points $p_{i+1}$ in $(i{+}1)$-$th$ frame based on motion consistency. We use $p_{i+1}$ as input for calculating $p_{i+2}$. The operation cycles until the abnormal skeleton data is corrected. We denote the start of the forward correction as $p_{front}$. In the other direction, we start from $p_{back}$ to perform backward correction on the motion information of post-sequence frames. We finally fuse the forward and backward correction results. In the case of extreme anomalies, which can only obtain a single reliable pre-sequence or post-sequence information, we make corrections based on this single information. The overall process is shown in Fig. 3 and can be written as:

$$\vec{c_t} = \sum_{i=front}^{t-1} \vec{BOFC}\left(p_i, \vec{f}_i, p_{i+1}\right), \tag{3}$$

$$\overleftarrow{c_t} = \sum_{i=back}^{t+1} \overleftarrow{BOFC}\left(p_i, \overleftarrow{f}_i, p_{i-1}\right), \tag{4}$$

$$c_t = F\left(\vec{c_t}, \overleftarrow{c_t}\right), \tag{5}$$

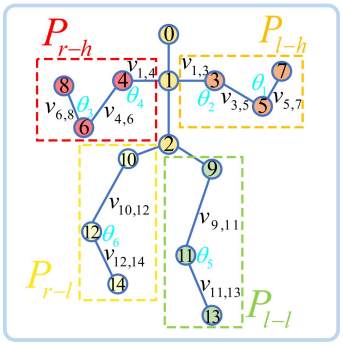
where,$\vec{c_t}$ and $\overleftarrow{c_t}$ are the order and reverse order optical flow corrections for abnormal skeleton data $p_t$, $F(\cdot)$ is the linear layer used for fusion operation to obtain the final results $c_t$.

**Loss Function.** To correct exact skeleton points, we adopt the mean squared error loss between ground truth $p$ and corrected skeleton point positions $\hat{p}$, which is defined as:

$$\mathcal{L}_{MSE}(\hat{p}, p) = \|\hat{p} - p\|^2. \tag{6}$$

## Part-Level Motion Feature Extraction Based on Task Decoupling (PETD)

**Part-Level Tracking.** To address non-uniform part motions, we propose a part-level motion feature extraction based on task decoupling (PETD). PETD detects and tracks each body part of dancers in the input sequence. Trained on DP, PETD's tracking sequence preserves the spatial-temporal motion of parts, providing candidate regions and motion details for associated skeleton points. When extracting initial poses, these regions are encoded as masks for temporal tracking matching on heatmaps. When fixing abnormal frames, the regions initially screen limb skeleton points. We further screen them to find points that are consistent with the overall limb motion trends and match the relative movement of articulated limbs as *reliable* data, and the rest are regarded as *unreliable*. The experiments prove that PETD can eliminate misjudgments caused by unreliable skeleton data and sufficiently improve the correction effect of BOFC.




0) head
1) neck
2) belly
3) left shoulder
4) right shoulder
5) left elbow
6) right elbow
7) left wrist
8) right wrist
9) left hip
10) right hip
11) left knee
12) right knee
13) left ankle
14) right ankle

$\theta_1$)∠3−5−7
$\theta_2$)∠1−3−5
$\theta_3$)∠4−6−8
$\theta_4$)∠1−4−6
$\theta_5$)∠9−11−13
$\theta_6$)∠10−12−14

Figure 4: An example of numbering 2D pose. We denote the human skeleton points as $\{p^k\}_{k=0}^{14}$, each limb as vector $v_{p^{i_1},p^{i_2}}$, each joint angle as $\{\theta_i\}_{i=1}^{6}$, and split the body part into the left hand, right hand, left leg and right leg subsets, i.e. $P_{l-h}$,$P_{r-h}$,$P_{l-l}$ and $P_{r-l}$.

**Loss Function.** We adopt the intersection over union loss for part-level tracking, which is defined as:

$$\mathcal{L}_{IoU}(\hat{x}, x) = 1 - \frac{area(\hat{x}) \cap area(x)}{area(\hat{x}) \cup area(x)}, \tag{7}$$

where $x$ and $\hat{x}$ are the ground truth and detection box respectively, $area(\cdot)$ denotes the area of the box.

## Group Dance Neatness Assessment (GDNA)

Our method aims to further explore the application of CV technologies in dance, an interesting and valuable scene. While existing action quality assessment (AQA) (Zhou et al. 2023, 2024a,b; Ke et al. 2024; Xu et al. 2024) task focuses on action completion and performance quality, relying on ratings by experts. In many dance scenes, especially group dance, there is still a lack of sufficient, authoritative, and unified assessment data. Creating reliable datasets akin to AQA is difficult. To this end, we use the simple and applied method to assess group dance by dance neatness.

**Assessment Algorithm.** We propose dance neatness assessment algorithms based on the extracted skeleton data. It is worth noting that dancers in groups often communicate with each other or audiences using eye contact and look conveys. The demand for head neatness is not high, so we exclude head in group dance assessment. As shown in Fig. 4, we number the 2D poses of the human body to describe the semantic features of each layer of the target. We first extract the limb-level features and split the limbs into eight parts: left forearm, left upper arm, right forearm, right upper arm, left thigh, left calf, right thigh, and right calf. After that, we link the two skeleton points involved in each limb part to form the limb feature vector $v_{p_{i_1},p_{i_2}}$. The cosine similarity of the same limb feature vector among the dancers is calculated, and the higher the similarity, the closer the limbs are to parallel, indicating neater movements. We further extract more fine-grained joint angle features, numbering the body joints commonly used in sports, and denote angles by $\theta = \{1, 2, \cdots, 6\}$, representing left elbow angle, left shoulder angle, right elbow angle, right shoulder angle, left knee

angle, and right knee angle, respectively. We normalize the joint angle values and use the L1 distance to calculate the degree of difference in the same joint angle among dancers. The smaller the distance, the neater the movement. With these two algorithms, the two grainy feature information of the human body is fused, and the correlation and neatness of limb and joint motion are analyzed comprehensively. For a group dance of $M$ dancers, the neatness assessment is:

$$S_{limbs}\left(i_1, i_2\right) = \frac{1}{L}\sum_{l=1}^{L}\frac{v_l^{i_1}\cdot v_l^{i_2}}{\left|v_l^{i_1}\right|\times\left|v_l^{i_2}\right|}, \tag{8}$$

$$S_{joints}\left(i_1, i_2\right) = \frac{1}{J}\sqrt{\sum_{j=1}^{J}\left(\frac{1}{1+\theta_j^{i_1}}-\frac{1}{1+\theta_j^{i_2}}\right)^2}, \tag{9}$$

$$NS = \frac{1}{M}\sum_{i_1,i_2\in M}\left(\lambda_1 S_{limbs}\left(i_1, i_2\right)+\lambda_2 S_{joints}\left(i_1, i_2\right)\right), \tag{10}$$

where $L$ and $J$ are the number of limbs and joints, we set $L=8, J=6$. $i_1$,$i_2$ are two dancers, $v_l^{i_1}$ and $v_l^{i_2}$ are the $l$-$th$ limb vectors, $\theta_j^{i_1}$ and $\theta_j^{i_2}$ are the $j$-$th$ joint angles, $\lambda_1$,$\lambda_2$ are the weights of limb neatness $S_{\text{limbs}}$ and joint neatness $S_{\text{joints}}$. $NS$ is the overall neatness score of the group.

Based on the traditional joint metrics (MPJAE and MPJPE), We propose four new strict metrics for assessing group dance, called relative mean per joint angle/position error (R-MPJAE/R-MPJPE) and their first-order velocity difference forms of relative mean per joint angle/position velocity error (R-MPJAVE/R-MPJPVE). We calculate the error in joint angle and position between pairs of dancers. Smaller errors indicate a neater dance. To eliminate the effect of dancers' different spatial positions, we take the neck point that can be stably detected as the base and normalize the dancer poses. Each joint position is subtracted by the base point (Toshev and Szegedy 2014), and divided by the torso diameter (the distance from left shoulder to right hip) as in PCK (Yang and Ramanan 2012). For dancers $i_1$ and $i_2$:

$$R\text{-}MPJAE(i_1, i_2) = \frac{1}{J}\sum_{j=1}^{J}\left|\theta_j^{i_1}-\theta_j^{i_2}\right|, \tag{11}$$

$$R\text{-}MPJPE(i_1, i_2) = \frac{1}{K}\sum_{k=1}^{K}\left\|\frac{p_k^{i_1}-p_{neck}^{i_1}}{\left\|p_{ls}^{i_1}-p_{rh}^{i_1}\right\|_2}-\frac{p_k^{i_2}-p_{neck}^{i_2}}{\left\|p_{ls}^{i_2}-p_{rh}^{i_2}\right\|_2}\right\|^2, \tag{12}$$

$$R\text{-}MPJAVE(i_1, i_2) = \frac{1}{J}\sum_{j=1}^{J}\left|\frac{R\text{-}MPJAE_{t+\Delta t}(i_1,i_2)-R\text{-}MPJAE_t(i_1,i_2)}{\Delta t}\right|, \tag{13}$$

$$R\text{-}MPJPVE(i_1, i_2) = \frac{1}{K}\sum_{k=1}^{K}\left\|\frac{R\text{-}MPJPE_{t+\Delta t}(i_1,i_2)-R\text{-}MPJPE_t(i_1,i_2)}{\Delta t}\right\|^2, \tag{14}$$

where $p_{neck}$, $p_{ls}$, and $p_{rh}$ are joint positions of the neck, left and right shoulder, respectively, with $K=15$ joints in total.

## Datasets

### Dancer Parts (DP)

**Motivation.** To address non-uniform part motions, we propose a PETD module to track part-level motion. Although (Yang et al. 2020) has proposed an "instance-level" body part dataset, it only focuses on the head, palms, and feet, and is not suitable for dances with complex part motions. To implement PETD, we consider the common motion properties of the dancers' limbs and construct the DP dataset.

We divide the body into six parts: left and right hands, left and right legs, head, and torso. Although the GDNA algorithms do not directly consider the torso, its labeling is essential for reconnecting and re-identifying parts during occlusions, improving tracking sequence length and credibility. So we build the DP (part-level dance dataset) based on these six parts, encompassing various dance types, scenes, and variations (Fig. 5 (a)). In addition, we label 15 common human keypoints as in (Jhuang et al. 2013) to learn accurate pose correction. The DP dataset, following the MOT20 (Dendorfer et al. 2020) standard, includes 65 group dance videos (50 for training and 15 for testing), each labeled with the position and tracking sequence for each dancer's parts.

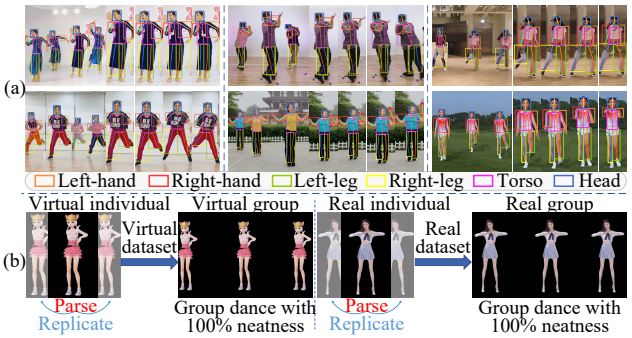



Figure 5: (a) Examples of the Dancer Parts dataset. We label the 15 common human keypoints, and six body parts: the left and right hands, the left and right legs, head, and torso. (b) Fully neat virtual dance dataset (DNV) generation method. Generate virtual and real-world datasets.

### Dancing-Neatly-in-Virtual (DNV)

**Motivation.** In reality, perfectly neat group dances hardly exist, and there is no professional quantitative label to define neatness. To tackle the GDNA task and verify the effectiveness of our assessment method on reliable data with quantitative metrics, we propose a DNV dataset based on the idea of motion synthesis (Tevet et al. 2022; Lee, Moon, and Lee 2023) to simulate the fully neat dances.

As shown in Fig. 5 (b), we generate virtual group dances by parsing and replicating virtual individual dances, created through methods such as dance generation algorithms, games, or LLMs. Then, real group dances are generated using real individual dances sourced from video websites. The resulting dataset is 100% neat, providing quantitative criteria for GDNA, facilitating research, and validating our method. We remove the background to avoid noise interfering with the 100% neatness, but also hire professional dancers to check that DNV's dance moves match real-life dances. To our knowledge, this is the first work to develop quantitative criteria for dance neatness assessment. The DNV dataset consists of 50 group dance videos (25 virtual, 25 real) at 30 FPS. We only use DNV for testing to verify GDNA.

**JHMDB.** We also validate our pose correction effects in

| Name | FPS | Resolution | Clips | Length | Box |
|---|---|---|---|---|---|
| DNV | 30 | 1280×720 | 50 | 5,300 | 238 |
| Dancer Parts | 30 | 1920×1080 | 65 | 13,666 | 1,190 |
| JHMDB | 15-40 | 320×240 | 928 | 31,838 | 928 |

Table 1: Parameters for DNV, Dancer Parts, and JHMDB.

| Methods | NS↑ | R-MPJAE↓ | R-MPJPE↓ | R-MPJAVE↓ | R-MPJPVE↓ |
|---|---|---|---|---|---|
| AlphaPose (ICCV 17) | 85.74 | 15.64 | 7.91 | 13.73 | 3.63 |
| **+DanceFix (Ours)** | **94.25** | **7.94** | **6.39** | **6.28** | **3.15** |
| Hrnet (CVPR 19) | 89.56 | 9.79 | 7.82 | 8.98 | 5.15 |
| **+DanceFix (Ours)** | **91.77** | **6.49** | **6.38** | **6.14** | **4.59** |
| RSN (ECCV 20) | 87.36 | 7.90 | 4.88 | 7.54 | 2.17 |
| **+DanceFix (Ours)** | **91.64** | **5.20** | **4.37** | **4.83** | **1.88** |
| LiteHrnet (CVPR 21) | 92.90 | 6.99 | 4.33 | 6.57 | 1.90 |
| **+DanceFix (Ours)** | **96.26** | **4.11** | **4.07** | **2.85** | **1.53** |
| PVT (ICCV 21) | 87.41 | 10.82 | 8.50 | 8.38 | 4.28 |
| **+DanceFix (Ours)** | **94.50** | **6.44** | **6.82** | **4.83** | **3.79** |
| PVT2 (CVME 22) | 92.58 | 5.08 | 4.91 | 3.77 | 2.76 |
| **+DanceFix (Ours)** | **94.37** | **4.06** | **4.34** | **2.75** | **2.04** |
| ED-Pose (ICLR 23) | 94.71 | 3.87 | 2.12 | 3.49 | 1.69 |
| **+DanceFix (Ours)** | **96.08** | **2.36** | **1.85** | **1.96** | **1.23** |
| DGN (IJCV 24) | 94.67 | 4.24 | 3.18 | 3.73 | 2.34 |
| **+DanceFix (Ours)** | **96.30** | **2.52** | **2.05** | **2.09** | **1.37** |
| SHaRPose (AAAI 24) | 94.40 | 3.71 | 2.97 | 3.36 | 1.96 |
| **+DanceFix (Ours)** | **96.02** | **2.22** | **1.63** | **1.75** | **1.18** |

Table 2: Comparison of the correction effects of applying DanceFix to the latest pose estimation algorithm on whole-body NS (%)↑, R-MPJAE↓, R-MPJPE (×100)↓, R-MPJAVE↓, and R-MPJPVE (×100)↓.

the widely-used video-based JHMDB (Jhuang et al. 2013) dataset for other action scenes. To facilitate comparison with previous work (Wei et al. 2016; Xiao, Wu, and Wei 2018), we evaluate our method in three subset divisions of JHMDB and report the average results. We show the content parameters of the three datasets in Table 1.

# Experiments

## Implementation Details

We conduct experiments on a Tesla V100 with PyTorch 1.8.0, using the open-source code and pre-trained models of previous works to obtain initial poses and dance neatness. We adopt the ByteTrack (Zhang et al. 2022b) and FlowFormer (Huang et al. 2022) as the backbone for part-level tracking and optical flow estimation, with learning rates of 1e-3 and 1e-4. The input size for training is 800 x 1440. We use the Adam optimizer and set the weight decay to 1e-5. We train our part-level tracker and fine-tune the pose and optical flow estimation models on the Dancer Parts. The DNV is just to verify the effect of our pose correction and dance neatness assessment. The hyperparameters $\tau_1$, $\tau_2$, $\tau_3$, $\lambda_1$ and $\lambda_2$ in Eq. (1) and Eq. (10) are 0.8, 0.02, 10, 0.5 and 0.5.

| Methods | Sub1 | Sub2 | Sub3 | Avg |
|---|---|---|---|---|
| CPM (CVPR 16) | 59.0 / 32.8 | 52.7 / 28.0 | 61.0 / 34.9 | 57.6 / 31.9 |
| **+DanceFix (Ours)** | **67.8 / 43.1** | **61.5 / 36.9** | **68.7 / 43.5** | **66.0 / 41.2** |
| SBL (ECCV 18) | 74.8 / 49.0 | 65.6 / 40.4 | 74.5 / 48.3 | 71.7 / 45.9 |
| **+DanceFix (Ours)** | **79.8 / 57.7** | **73.2 / 49.7** | **79.8 / 57.1** | **77.6 / 54.8** |
| PoseWarper (NIPS 19) | 69.2 / 43.9 | 64.8 / 38.5 | 72.9 / 44.4 | 68.9 / 42.2 |
| **+DanceFix (Ours)** | **75.5 / 50.9** | **71.1 / 45.8** | **78.2 / 52.0** | **74.9 / 49.5** |
| DCPose (CVPR 21) | 61.8 / 36.9 | 56.5 / 32.2 | 64.1 / 37.3 | 60.8 / 35.5 |
| **+DanceFix (Ours)** | **70.7 / 45.6** | **66.2 / 40.8** | **73.3 / 46.2** | **70.1 / 44.2** |
| DeciWatch (ECCV 22) | 86.6 / 62.6 | 82.3 / 56.1 | 86.2 / 61.7 | 85.0 / 60.1 |
| **+DanceFix (Ours)** | **88.0 / 67.6** | **83.8 / 60.6** | **87.4 / 66.2** | **86.4 / 64.8** |
| HANet (WACV 23) | 87.3 / 63.4 | 83.9 / 57.9 | 87.2 / 62.4 | 86.1 / 61.2 |
| **+DanceFix (Ours)** | **89.6 / 69.9** | **86.6 / 63.8** | **89.7 / 69.1** | **88.6 / 67.6** |

Table 3: Comparison of the correction effect of applying DanceFix to existing methods on the public JHMDB dataset on the PCK@0.2 / PCK@0.1↑ metrics.

**Evaluation Metrics.** We validate the correction effect on the human body by calculating the neatness score (NS) of the left hand, right hand, left leg, and right leg, and the NS, R-MPJAE, R-MPJPE, R-MPJAVE, and R-MPJPVE of the whole body (all are introduced in Method). In addition, we validate our transferability for abnormal skeleton data correction on the JHMDB using the PCK metric.

## Experiment Results

**Correction Effects on State-of-the-Art Methods.** We report the effect of pose correction on prior works (Fang et al. 2017; Sun et al. 2019; Cai et al. 2020; Yu et al. 2021; Wang et al. 2021b, 2022; Yang et al. 2023a; Tu, Wu, and Wang 2024; An et al. 2024) on DNV in Table 2. The first row lists the initial dance neatness, while the second row shows the improvement achieved by our method. We see that our method has a good correction effect on all methods with an average increase of 3.54%, 2.97, 0.97, 2.90, and 0.57 in whole-body NS, R-MPJAE, R-MPJPE (×100), R-MPJAVE, and R-MPJPVE (×100) respectively, showing the strong effectiveness of the DanceFix.

**Results on JHMDB Dataset.** We report the average correction effect on the public JHMDB in Table 3. Our method achieves good corrections with existing methods (video-based pose estimation (Wei et al. 2016; Zeng et al. 2022; Jin et al. 2023) and pose tracking (Xiao, Wu, and Wei 2018; Bertasius et al. 2019; Liu et al. 2021)). The PCK@0.2 / 0.1 are improved by 8.4% / 9.3%, 5.9% / 8.9%, 6.0% / 7.3%, 9.3% / 8.7%, 1.4% / 4.7%, and 2.5% / 6.4% on six methods, respectively. *Although DanceFix aims to be studied for dance scenes, there are also great effects in general scenes, showing the strong effectiveness of our method.*

## Analysis

**Ablation Studies.** We conduct experiments using AlphaPose (Fang et al. 2017) as the pose estimation backbone, which is a competitive and widely used algorithm. Correcting a non-optimal initial pose sequence helps DanceFix to

| Methods | Real-world Dataset | | | | | Virtual Dataset | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | LH | RH | LL | RL | All | LH | RH | LL | RL | All |
| AlphaPose | 95.34 | 84.41 | 89.59 | 96.50 | 85.74 | 94.94 | 93.82 | 99.36 | 98.33 | 94.39 |
| Replace | 95.04 | 90.39 | 95.36 | 98.81 | 91.48 | 95.43 | 95.07 | 99.47 | 98.41 | 94.97 |
| Replace+PETD | **97.74** | 89.17 | 95.36 | 98.82 | 92.57 | 95.53 | **95.11** | **99.48** | 98.41 | 95.00 |
| Line Speed | 95.15 | 90.09 | 95.36 | 98.82 | 91.88 | 95.44 | 94.90 | 99.44 | 98.41 | 94.89 |
| Line Speed+PETD | 97.40 | 89.53 | **95.37** | 98.82 | 92.22 | 95.51 | 94.95 | 99.45 | 98.41 | 94.92 |
| Ours (BOFC) | 95.03 | 92.22 | 95.36 | **98.83** | 93.19 | 95.66 | 94.99 | **99.48** | 98.40 | **95.02** |
| Ours (BOFC+PETD) | 97.71 | **92.80** | 95.36 | **98.83** | **94.25** | **95.69** | 95.00 | **99.48** | **98.42** | **95.02** |

Table 4: Ablation experiments with BOFC and PETD modules for the NS (%)↑ (including LH (Left-hand), RH (Right-hand), LL (Left-leg), RL (Right-leg), ALL (Whole-body)) on DNV's Real-world Dataset and Virtual Dataset.

| Dancers | $S_{limbs}$ (s) | $S_{joints}$ (s) | R-MPJAE (s) | R-MPJPE (s) |
|---|---|---|---|---|
| 2 | 0.002 | 0.008 | 0.008 | 0.018 |
| 7 | 0.038 | 0.035 | 0.037 | 0.076 |
| 15 | 0.056 | 0.082 | 0.082 | 0.118 |
| 30 | 0.107 | 0.161 | 0.168 | 0.252 |

Table 5: Computational cost of our proposed dance neatness assessment metric in 60 frames, with time units in seconds.

better face the challenges in dance scenes. We randomly select 20 videos from DNV, 10 real and 10 virtual scenes, and average the results. As shown in the left part of Table 4, compared with directly replacing abnormal skeleton data with reliable temporal frames (denoted as *Replace*) and calculating based on the movement speed of the skeleton points in the temporal frames (denoted as *Line Speed*), our BOFC is more effective in correcting the skeleton data of dancers, achieving an NS increasing by 8.38% of the whole body than AlphaPose. With the PETD added, the whole-body NS of the three methods improves by 1.09%, 0.34%, and 1.06%. Moreover, DanceFix also achieves good results on virtual datasets, which shows the feasibility of applying this method to scenes such as VR and games. The above results indicate the effectiveness of the components of DanceFix.

**Computational Cost.** The group dance assessment metrics we proposed require pairwise dancer comparison, which result in an exponential increase of comparison numbers as the number of dancers increases. To this end, we study the computational cost for different numbers of dancers, as shown in Table 5. For a 60-frame video (2 seconds at 30 FPS), all calculations can be completed within 1 second, even with up to 30 dancers. This suggests that our proposed assessment metrics can be applied in real time for industrial applications.

**Case Studies.** We compare our results with the Avg. expert neatness of ten professional dancers. As shown in Fig. 6 (a), DanceFix corrects and gives more accurate neatness.

**Visualization.** We show the correction effect of DanceFix on abnormal cases under realistic complex dance scenes in Fig. 6 (b), such as occlusion and motion blur. Our DanceFix can effectively correct the abnormal skeleton data.

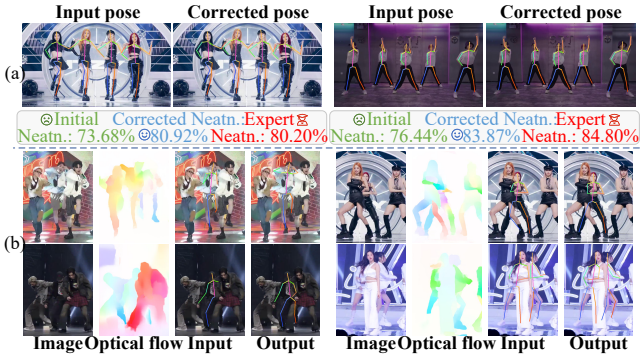



Figure 6: (a) Case studies for comparing expert neatness scores. (b) The feasibility of optical flow and correction effects under realistic complex, flashy dances.

## Conclusion and Discussion

In this paper, we have explored a new Group Dance Neatness Assessment (GDNA) task. We have proposed a bidirectional optical flow correction (BOFC), aiming to correct abnormal skeleton data. Meanwhile, we have collected the Dancer Parts dataset and trained a part-level tracker (PETD) to obtain reliable temporal frames. We have also proposed the DNV dataset and dance assessment algorithms. Experiments on DNV and JHMDB have shown that our method can be flexibly embedded into existing works to correct abnormal skeleton data and improve accuracy. To the best of our knowledge, we are the first to define quantitative criteria for group dance assessment. We hope that our work will provide a meaningful and fun idea for dance assessment.

**Potential Impact.** DanceFix is applicable to group dance assessment, motion instruction, medical rehabilitation, group sports, military training, and many other scenes.

**Limitations and Future Research.** The proposed method examines the neatness of group dance, which assumes that dancers follow the same actions. This work contributes two new datasets, including real-world dance sourced from video websites, and we are actively contacting the creators to obtain appropriate consent. We further plan to expand our research on 3D pose perspective with more specialized abnormal pose detection and build dance datasets with expert labels to explore further dance research as in the AQA task.

## Acknowledgments

This work was supported in part by the National Key Research and Development Plan of China under Grant 2021YFB3600503, in part by the National Natural Science Foundation of China under Grant 61972097 and U21A20472, in part by the Major Scientific Research Project for Technology Promotes Police under Grant 2024YZ040001, in part by the Natural Science Foundation of Fujian Province under Grant 2021J01612, in part by the Major Science and Technology Project of Fujian Province under Grant 2021HZ022007.

## References

An, X.; Zhao, L.; Gong, C.; Wang, N.; Wang, D.; and Yang, J. 2024. SHaRPose: Sparse High-Resolution Representation for Human Pose Estimation. In *AAAI*, 691–699.

Bertasius, G.; Feichtenhofer, C.; Tran, D.; Shi, J.; and Torresani, L. 2019. Learning Temporal Pose Estimation from Sparsely-Labeled Videos. In *NeurIPS*, 3021–3032.

Cai, Y.; Wang, Z.; Luo, Z.; Yin, B.; Du, A.; Wang, H.; Zhang, X.; Zhou, X.; Zhou, E.; and Sun, J. 2020. Learning delicate local representations for multi-person pose estimation. In *ECCV*, 455–472.

Chen, Z.; Li, S.; Yang, B.; Li, Q.; and Liu, H. 2021. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *AAAI*, 1113–1122.

Chu, P.; Wang, J.; You, Q.; Ling, H.; and Liu, Z. 2023. Transmot: Spatial-temporal graph transformer for multiple object tracking. In *WACV*, 4870–4880.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, 886–893.

Dendorfer, P.; Rezatofighi, H.; Milan, A.; Shi, J.; Cremers, D.; Reid, I.; Roth, S.; Schindler, K.; and Leal-Taixé, L. 2020. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2334–2343.

Feng, R.; Gao, Y.; Ma, X.; Tse, T. H. E.; and Chang, H. J. 2023. Mutual information-based temporal difference learning for human pose estimation in video. In *CVPR*, 17131–17141.

Guo, X.; Zhao, Y.; and Li, J. 2021. DanceIt: music-inspired dancing video synthesis. *TIP*, 30: 5559–5572.

Hu, L.; Zhang, B.; Zhang, P.; Qi, J.; Cao, J.; Gao, D.; Zhao, H.; Feng, X.; Wang, Q.; Zhuo, L.; Pan, P.; and Xu, Y. 2021. A Virtual Character Generation and Animation System for E-Commerce Live Streaming. In *ACM MM*, 1202–1211.

Huang, Z.; Shi, X.; Zhang, C.; Wang, Q.; Cheung, K. C.; Qin, H.; Dai, J.; and Li, H. 2022. FlowFormer: A Transformer Architecture for Optical Flow. *ECCV*, 668–685.

Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; and Black, M. J. 2013. Towards understanding action recognition. In *ICCV*, 3192–3199.

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 *WACV*, 5725–5734.

Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; and Boussaid, F. 2017. A new representation of skeleton sequences for 3d action recognition. In *CVPR*, 3288–3297.

Ke, X.; Xu, H.; Lin, X.; and Guo, W. 2024. Two-path target-aware contrastive regression for action quality assessment. *Inf. Sci.*, 664: 120347.

Lee, T.; Moon, G.; and Lee, K. M. 2023. MultiAct: Long-Term 3D Human Motion Generation from Multiple Action Labels. In *AAAI*, 1231–1239.

Lei, Q.; Li, H.; Zhang, H.; Du, J.; and Gao, S. 2023. Multi-skeleton structures graph convolutional network for action quality assessment in long videos. *Appl. Intell.*, 1–14.

Li, B.; Chen, J.; Zhang, D.; Bao, X.; and Huang, D. 2022. Representation Learning for Compressed Video Action Recognition via Attentive Cross-modal Interaction with Motion Enhancement. In *IJCAI*, 1060–1066.

Li, C.; Xie, C.; Zhang, B.; Han, J.; Zhen, X.; and Chen, J. 2021. Memory attention networks for skeleton-based action recognition. *TNNLS*, 33(9): 4800–4814.

Li, Z.; Zhang, D.; Wu, S.; Song, M.; and Chen, G. 2024. Sampling-Resilient Multi-Object Tracking. In *AAAI*, 3297–3305.

Lin, L.; Zhang, J.; and Liu, J. 2023. Actionlet-dependent contrastive learning for unsupervised skeleton-based action recognition. In *CVPR*, 2363–2372.

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 *CVPR*, 525–534.

Moon, G.; Chang, J. Y.; and Lee, K. M. 2019. Posefix: Model-agnostic general human pose refinement network. In *CVPR*, 7773–7781.

Pfister, T.; Charles, J.; and Zisserman, A. 2015. Flowing convnets for human pose estimation in videos. In *ICCV*, 1913–1921.

Poux, D.; Allaert, B.; Ihaddadene, N.; Bilasco, I. M.; Djeraba, C.; and Bennamoun, M. 2021. Dynamic facial expression recognition under partial occlusion with optical flow reconstruction. *TIP*, 31: 446–457.

Radevski, G.; Grujicic, D.; Blaschko, M.; Moens, M.-F.; and Tuytelaars, T. 2023. Multimodal distillation for egocentric action recognition. In *ICCV*, 5213–5224.

Song, J.; Wang, L.; Van Gool, L.; and Hilliges, O. 2017. Thin-slicing network: A deep structured model for pose estimation in videos. In *CVPR*, 4220–4229.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 5693–5703.

Tevet, G.; Gordon, B.; Hertz, A.; Bermano, A. H.; and Cohen-Or, D. 2022. Motionclip: Exposing human motion generation to clip space. In *ECCV*, 358–374.

Toshev, A.; and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 1653–1660.

Tu, J.; Wu, G.; and Wang, L. 2024. Dual Graph Networks for Pose Estimation in Crowded Scenes. *IJCV*, 132(3): 633–653.

Vemulapalli, R.; Arrate, F.; and Chellappa, R. 2014. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 588–595.

Wang, Q.; Du, J.; Yan, K.; and Ding, S. 2023. Seeing in flowing: Adapting clip for action recognition with motion prompts learning. In *ACM MM*, 5339–5347.

Wang, S.; Yang, D.; Zhai, P.; Chen, C.; and Zhang, L. 2021a. Tsa-net: Tube self-attention network for action quality assessment. In *ACM MM*, 4902–4910.

Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022. PVT v2: Improved baselines with Pyramid Vision Transformer. *CVME*, 8(3): 415–424.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021b. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 568–578.

Wang, Y.-H.; Hsieh, J.-W.; Chen, P.-Y.; Chang, M.-C.; So, H.-H.; and Li, X. 2024. Smiletrack: Similarity learning for occlusion-aware multiple object tracking. In *AAAI*, 5740–5748.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *CVPR*, 4724–4732.

Wen, Y.; Pan, H.; Yang, L.; Pan, J.; Komura, T.; and Wang, W. 2023. Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In *CVPR*, 21243–21253.

Wu, C.; Wu, X.-J.; Kittler, J.; Xu, T.; Ahmed, S.; Awais, M.; and Feng, Z. 2024. SCD-Net: Spatiotemporal Clues Disentanglement Network for Self-Supervised Skeleton-Based Action Recognition. In *AAAI*, 5949–5957.

Wu, E.; Piekenbrock, M.; Nakumura, T.; and Koike, H. 2021. Spinpong-virtual reality table tennis skill acquisition using visual, haptic and temporal cues. *TVCG*, 27(5): 2566–2576.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple baselines for human pose estimation and tracking. In *ECCV*, 466–481.

Xie, J.; Meng, Y.; Zhao, Y.; Nguyen, A.; Yang, X.; and Zheng, Y. 2024. Dynamic Semantic-Based Spatial Graph Convolution Network for Skeleton-Based Human Action Recognition. In *AAAI*, 6225–6233.

Xu, H.; Ke, X.; Li, Y.; Xu, R.; Wu, H.; Lin, X.; and Guo, W. 2024. Vision-language action knowledge learning for semantic-aware action quality assessment. In *ECCV*, 423–440.

Xu, K.; Ye, F.; Zhong, Q.; and Xie, D. 2022. Topology-aware convolutional neural network for efficient skeleton-based action recognition. In *AAAI*, 2866–2874.

Xu, L.; Guan, Y.; Jin, S.; Liu, W.; Qian, C.; Luo, P.; Ouyang, W.; and Wang, X. 2021. Vipnas: Efficient video pose estimation via neural architecture search. In *CVPR*, 16072–16081.

Yang, J.; Zeng, A.; Liu, S.; Li, F.; Zhang, R.; and Zhang, L. 2023a. Explicit Box Detection Unifies End-to-End Multi-Person Pose Estimation. In *ICLR*.

Yang, L.; Song, Q.; Wang, Z.; Hu, M.; and Liu, C. 2020. Hier R-CNN: Instance-level human parts detection and a new benchmark. *TIP*, 30: 39–54.

Yang, Y.; and Ramanan, D. 2012. Articulated human detection with flexible mixtures of parts. *TPAMI*, 35(12): 2878–2890.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023b. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 4210–4220.

Ye, Z.; Wu, H.; Jia, J.; Bu, Y.; Chen, W.; Meng, F.; and Wang, Y. 2020. Choreonet: Towards music to dance synthesis with choreographic action unit. In *ACM MM*, 744–752.

Yu, C.; Xiao, B.; Gao, C.; Yuan, L.; Zhang, L.; Sang, N.; and Wang, J. 2021. Lite-hrnet: A lightweight high-resolution network. In *CVPR*, 10440–10450.

Yu, E.; Li, Z.; Han, S.; and Wang, H. 2023. RelationTrack: Relation-Aware Multiple Object Tracking With Decoupled Representation. *TMM*, 25: 2686–2697.

Zeng, A.; Ju, X.; Yang, L.; Gao, R.; Zhu, X.; Dai, B.; and Xu, Q. 2022. DeciWatch: A Simple Baseline for 10x Efficient 2D and 3D Pose Estimation. In *ECCV*, 607–624.

Zhang, J.; Tu, Z.; Yang, J.; Chen, Y.; and Yuan, J. 2022a. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *CVPR*, 13232–13242.

Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; and Wang, X. 2022b. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, 1–21.

Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; and Liu, W. 2021. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *IJCV*, 129(11): 3069–3087.

Zhong, Y.; and Demiris, Y. 2024. DanceMVP: Self-Supervised Learning for Multi-Task Primitive-Based Dance Performance Assessment via Transformer Text Prompting. In *AAAI*, 10270–10278.

Zhou, H.; Liu, Q.; and Wang, Y. 2023. Learning discriminative representations for skeleton based action recognition. In *CVPR*, 10608–10617.

Zhou, K.; Li, J.; Cai, R.; Wang, L.; Zhang, X.; and Liang, X. 2024a. CoFInAl: Enhancing Action Quality Assessment with Coarse-to-Fine Instruction Alignment. In *IJCAI*, 1771–1779.

Zhou, K.; Ma, Y.; Shum, H. P.; and Liang, X. 2023. Hierarchical Graph Convolutional Networks for Action Quality Assessment. *TCSVT*, 33(12): 7749–7763.

Zhou, K.; Wang, L.; Zhang, X.; Shum, H. P. H.; Li, F. W. B.; Li, J.; and Liang, X. 2024b. MAGR: Manifold-Aligned Graph Regularization for Continual Action Quality Assessment. In *ECCV*, 375–392.

Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; and Xie, X. 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In *AAAI*, 3697–3704.