Research on Engineering Structures & Materials
www.jresm.org
MIM Research Group

Research Article

Intelligent action recognition and dance movement optimization based on multi-threshold image segmentation

Xin Lai [*,a]

Department of Music and Dance, Jiaying University, Meizhou, Guangdong, 514015, China

Article Info

Article History:

Received 20 Jul 2026

Accepted 24 Sep 2026

Keywords:

Multi-threshold image segmentation;
Action recognition;
Dance movement optimization;
Graph convolutional network;
Metaheuristic optimization

Abstract

Dance movement analysis requires accurate image segmentation, action recognition, and movement quality assessment. However, in traditional systems, these operations are carried out independently, and no relationship is established between the accuracy of segmentation and the results of classification. This study proposes an end-to-end framework integrating an ADE-GJO hybrid optimization-based multi-threshold image segmentation method, a segmentation-guided spatio-temporal attention action recognition network termed SG-STANet, and a dynamic time warping-based movement deviation measurement mechanism. The ADE-GJO algorithm incorporates chaotic initialization, adaptive DE-based mutation operation, and Lévy flight in the GJO algorithm to enhance its global exploration capability at high threshold levels. The SG-STANet model uses foreground information provided through segmentation as a prior for spatial attention to prevent any negative effect of the background on skeleton feature extraction. Experiments performed on BSDS500, AIST++, and NTU RGB+D 120 show that the introduced segmentation algorithm achieves decent results. Mask-guided mechanism results in significant improvements for classification purposes, while the DTW similarity measure successfully recognizes different dance styles based on their quality by means of native data set annotations. Overall, the introduced approach provides a technically reasonable pathway toward the use of computer vision technologies in evaluating dance education and intelligent training.

© 2026 MIM Research Group. All rights reserved.

1. Introduction

Dance movement analysis has extensive applications in the domains of intelligent sport training, dance education evaluation, and human-computer interaction. Traditional approaches rely heavily on expert opinion, which introduces a number of inherent limitations, such as inconsistency in criteria used and difficulties in quantifying analysis. The computer vision technique offers a new technological path for automated dance movement analysis and evaluation. Within this technical pipeline, image segmentation plays the role of the first stage of visual processing and impacts the precision of future stages of feature extraction and action recognition. Segmentation using multi-thresholds can divide an image into several grayscale segments and keep hierarchical structure for different parts of the body, thus enabling more accurate feature extraction and action recognition in the following stages. It is important to research more effective algorithms for multi-threshold segmentation and their application to movement optimization and action recognition because this could significantly improve automation in dance movement analysis.

In the area of multi-threshold image segmentation, the between-class variance criterion of Otsu and the entropy criterion of Kapur are classical thresholding schemes. However, the ever-increasing computation complexity associated with the rise in threshold numbers is considered one of the major limitations affecting segmentation results. To overcome this limitation,

*Corresponding author: xinlai4@outlook.com
[a]orcid.org/0009-0008-8505-1988
DOI: http://dx.doi.org/10.17515/resm2026-1938am0720rs

metaheuristic optimization algorithms are being used, tackling mainly two major issues: search mechanism optimization and convergence strategy improvement. Concerning the search algorithms, the use of dual-subpopulation strategy has been adopted to increase the stability of segmentation [1], while the adoption of repair mechanism has successfully avoided any generation of duplicate thresholds [2]. Regarding convergence techniques, the use of Cauchy mutation along with the opposition-based technique has resulted in a precise search algorithm with a high number of thresholds [3], whereas the combination of the opposition-based technique with chaos sequences has proven to be a reliable way of achieving global exploration and exploitation in search algorithms [4]. Moreover, multi-task segmentation scheme, which works with both grayscale and color images, has expanded the scope of application of such algorithms [5].

Within human action recognition and dance movement detection tasks, the current standard for designing graph convolutional networks has been through the skeleton structure. According to a comparative study carried out between RGB-based and skeleton-based methods, the latter successfully addresses the issue of interfering backgrounds but fails to account for the presence of contextual semantics [6], and a comparative study across multiple deep network architectures has further confirmed the performance superiority of GCNs in graph-structured data modeling [7]. Since pose estimation is an upstream element in support of skeleton-based recognition, it imposes direct limitations on recognition performance, with lightweight architecture and occlusion tolerance remaining the most important problems at this level [8]. A multiscale spatiotemporal graph convolutional network that allows for lightweight architecture through multi-scale feature representation has been successfully designed [9]. In the domain of dance movement analysis, the studies have been conducted from various dimensions such as motion extraction, generation, and evaluation. The methods based on deep learning have been used for the extraction of skeleton motions as well as dance interaction systems [10], whereas music-based motion mapping has provided a technical assistance for the choreography of dance [11]. Moreover, the models with multiscale spatio-temporal graph convolutional network have been found to possess excellent flexibility in response to different behaviors [12]. A pose estimation-based technique performs body feature extraction through segmentation of regions of motion [13], while the computer vision-based technique involves character modeling using self-organizing maps network [14]. The two techniques validate the automation of dance recognition. In terms of analysis at a finer grain size, a framework that uses pose estimation techniques has been proposed to measure the cognitive load of dance movement from a psychological viewpoint [15], while another framework using reinforcement learning techniques has developed an optimal motion sampling policy using dance videos [16].

A careful examination of the current literature highlights three concerns that remain unaddressed. Multi-threshold segmentation research and downstream action recognition research have advanced largely in isolation, so that the quality of segmentation is never propagated as an explicit prior to skeleton feature extraction. Skeleton-based recognition networks learn spatial attention weights in a purely data-driven manner without any external visual supervision, leaving them vulnerable to background clutter in unconstrained dance scenes. Moreover, no existing pipeline closes the loop from segmentation through style classification to per-joint movement deviation measurement, which is the feedback granularity required by intelligent dance training systems.

To address these gaps, the present study makes the following contributions. (1) An ADE-GJO hybrid metaheuristic algorithm is proposed for multi-threshold image segmentation, incorporating chaotic Tent initialization, a diversity-triggered adaptive differential evolution mutation operator, and Lévy flight perturbation into the Golden Jackal Optimizer; the synergy of on-demand perturbation and continuous exploration yields consistent improvements at high threshold dimensionalities where the standard GJO and other baselines stagnate. (2) A segmentation-guided spatio-temporal attention network termed SG-STANet is designed, in which the foreground mask produced by ADE-GJO is embedded as a learnable spatial prior into the recognition backbone through element-wise modulation with a residual connection, establishing an explicit information pathway from segmentation quality to recognition accuracy that has not been explored in prior skeleton-based frameworks. (3) A DTW-based movement deviation measurement mechanism is developed that aligns test joint-angle sequences with style-specific reference templates and yields

per-joint deviation scores whose discriminative power is validated against the native basic/advanced choreography annotations of the AIST++ dataset, thereby eliminating the need for manual scoring. Collectively, these three modules form an end-to-end framework that converts a raw dance video frame into a style label together with a quantitative movement quality profile, providing a technically viable pathway toward the application of computer vision in dance education assessment and intelligent training.

2. Methodology

2.1 Overall System Architecture

The proposed approach uses an end-to-end model that has three stages. In the first stage, a video frame of the dancing sequence is transformed into a gray image and used for computing the best threshold value through the ADE-GJO hybrid optimization technique. After obtaining the optimal threshold, a human body foreground mask is created and is termed M. In the second stage, this foreground mask M is used in the SG-STANet classification framework to classify the dance styles. In the third stage, the sequences of joint angles resulting from the recognition process are dynamically time-warped to fit a template, and the joint movement deviation scores are generated. The schematic representation of the whole system design is shown in Fig. 1.

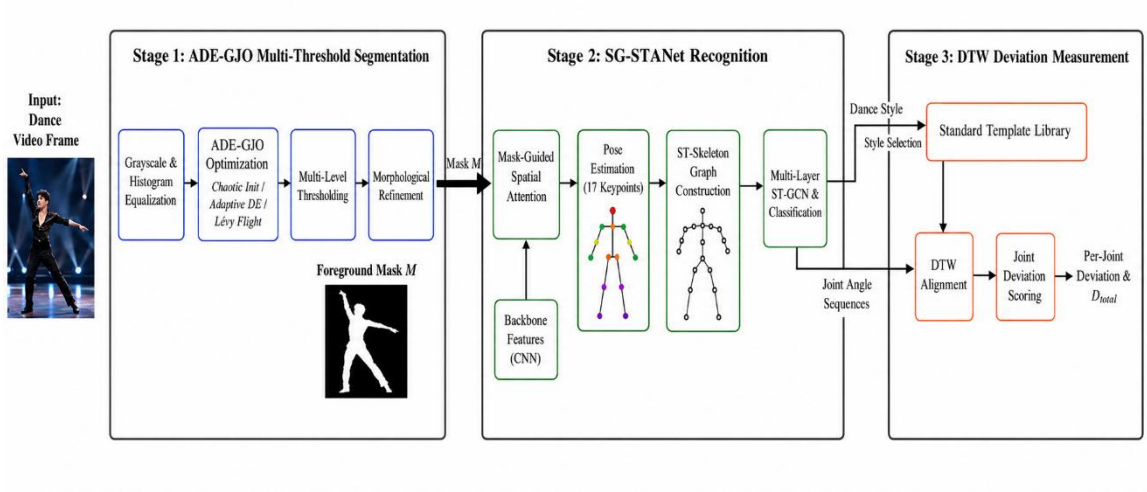


Fig. 1. Overall architecture of the proposed method

Fig. 1 shows the flow of information and interaction between modules for all three steps. The foreground mask M plays a critical role in connecting segmentation with recognition, where the performance level impacts the success rate of the skeleton extraction and final recognition results. Meanwhile, the result from the classification module decides which template is chosen through Style Selection. This establishes a closed-loop connection between recognition and optimization.

2.2 Multi-Threshold Image Segmentation Based on ADE-GJO

The core of multi-threshold image segmentation lies in searching for an optimal threshold vector $t = [t_1, t_2, \dots, t_n]$ such that the segmented regions satisfy specific optimality criteria. The Otsu between-class variance method finds its optimum threshold by maximizing the difference in variances of the divided segments [17], while Kapurs entropy method finds its optimum by maximizing entropy within each segment for intra-class grayscale uniformity. Within dance imagery, the human body consists of various grayscales representing the skin, hair, and clothing, and it is impossible for one criterion alone to satisfy both regional separation and completeness within each region. This limitation is overcome by combining the two criteria into a dual-objective fitness function

$$F(t) = \alpha\sigma_B^2(t) + (1-\alpha)H_K(t) \quad (1)$$

where $\sigma_B^2(t) = \sum_{k=0}^{n} \omega_k(\mu_k - \mu_T)^2$ denotes the Otsu between-class variance with μ_T representing the global mean grayscale of the image, $H_K(t) = -\sum_{k=0}^{n} \sum_{i \in C_k} \frac{p_i}{\omega_k} \ln \frac{p_i}{\omega_k}$ denotes the Kapur entropy, and α is a balance coefficient.

Since the between-class variance and the Kapur entropy differ in magnitude and physical meaning, both terms are min–max normalized to the interval [0,1] within each iteration before being linearly combined, which ensures that the two objectives contribute on comparable numerical scales and prevents either criterion from dominating the search direction due to raw-value disparity. The Otsu term captures inter-class separability by maximizing the weighted variance between grayscale segments, whereas the Kapur term captures intra-class homogeneity by maximizing the entropy within each segment, and the two criteria are therefore complementary in nature, where the former favors sharp boundaries between regions while the latter favors uniform statistical distributions within each region. The linear combination governed by αenables the algorithm to trade off boundary clarity against regional consistency, which is well-suited to dance imagery where clothing textures and skin tones require both region-level separation and intra-region smoothness.

As the number of thresholds n increases, the search space expands exponentially. In those systems where multi-threshold segmentation is combined with activity recognition, the optimization process plays an extremely important role in improving the efficiency of segmentation as well as recognizing activities [18]. The proposed ADE-GJO algorithm is based on the predator-prey search methodology that is used by the Golden Jackal Optimizer, while using three types of modifications. The first modification lies in utilizing chaotic Tent mapping in the initialization stage, which allows obtaining more even population distribution in the search space. During the iterative search phase, a population diversity metric $\delta = \mathrm{std}(f)/\mathrm{mean}(f)$ is defined, and when $\delta < \delta_{\mathrm{th}}$, an adaptive differential evolution mutation operator is triggered

$$\mathrm{v}_i = \mathrm{t}_{\mathrm{best}} + F_m \cdot \left(\mathrm{t}_{r_1} - \mathrm{t}_{r_2}\right) \tag{2}$$

where $\mathbf{t}_{\mathrm{best}}$ is the current global best individual, $\mathbf{t}_{r_1}$ and $\mathbf{t}_{r_2}$ are randomly selected individuals, and $F_m \in (0,1]$ is a scaling factor distinct from the fitness function $F(\cdot)$. The strategy ensures perturbation only when diversity among the population falls short. During the process of position updating, an additional factor of Lévy Flight step size is introduced

$$\mathrm{t}_i^{\mathrm{new}} = \mathrm{t}_i + \beta_0 \cdot L(\lambda) \odot (\mathrm{t}_i - \mathrm{t}_{\mathrm{best}}) \tag{3}$$

where $L(\lambda) \in \mathbb{R}^n$ denotes a vector of random steps whose entries are independently drawn from the Lévy distribution with $L(\lambda) \sim s^{-\lambda}$ ($1 < \lambda \leq 3$), $\odot$ indicates the Hadamard (entry-wise) product, and the heavy-tailed property of $L(\lambda)$ generates occasional large jumps to enhance global exploration capability.

The entire procedure for implementing ADE-GJO is provided by Algorithm 1. The use of the conditional test employing the diversity metric δ and the threshold δ_{th} ensures that DE mutation takes place only when the population gets stuck into convergence rather than at every iteration, so that it does not disrupt the basic search mechanism. Meanwhile, the Lévy flight updating occurs continuously at each iteration, maintaining global exploration abilities. Collectively, both methods make up a synergistic approach defined as “on-demand perturbation together with continuous exploration.” The time complexity of the algorithm is $O(N \cdot D \cdot T)$. The synergistic workflow of the three improvements is described in Algorithm 1.

Algorithm 1: ADE-GJO Multi-Threshold Image Segmentation

Input: Grayscale image I, threshold number n, population size N, max iterations T, balance coefficient α, diversity threshold δ_{th}, scaling factor F_m, step control coefficient β_0

Output: Optimal threshold vector t^*, foreground mask M

Compute pixel probability distribution from histogram of I

Initialize population $\{\mathrm{t}_i\}_{i=1}^{N}$ via chaotic Tent map

Evaluate dual-objective fitness $F(\mathrm{t}_i)$ for each individual
Set $\mathrm{t}_{\mathrm{best}} \leftarrow$ individual with highest fitness
for $iter = 1$ to T do
  Compute population diversity δ
  if $\delta < \delta_{\mathrm{th}}$ then
    for $i = 1$ to N do
      Randomly select $r_1 \neq r_2 \neq i$ from population
      Generate mutant v_i *using* $\mathrm{t}_{\mathrm{best}}, \mathrm{t}_{r_1}, \mathrm{t}_{r_2}$ with F_m
      Replace $\mathrm{t}_i \leftarrow \mathrm{v}_i$ if $F(\mathrm{v}_i) > F(\mathrm{t}_i)$
    end for
  end if
  for $i = 1$ to N do
    Generate Lévy step $L(\lambda)$ and update t_i with β_0
    Apply boundary constraint and sort thresholds ascending
    Evaluate $F(\mathrm{t}_i^{\mathrm{new}})$ and update t_i if improved
  end for
  Update $\mathrm{t}_{\mathrm{best}} \leftarrow$ individual with highest fitness
end for
Segment I using $\mathrm{t}^* = \mathrm{t}_{\mathrm{best}}$ into $(n+1)$ regions
Extract human body region via connected component area and positional prior
Apply morphological opening and closing operations
return t^*, foreground mask M

2.3 Segmentation-Guided Spatio-Temporal Attention Action Recognition Network

Attention has been successfully utilized in skeleton-based action recognition, contributing greatly to the capability of graph convolutional networks for learning meaningful motion features [19]. Despite this achievement, the attention weights in these frameworks are learned from the skeleton input alone, and no mechanism exists to inject image-level foreground–background information into the graph convolutional pipeline. In the SG-STANet, the mask M obtained from multi-threshold segmentation is directly embedded into the recognition network to design a mask-driven spatial attention module

$$\mathrm{F}_s' = \mathrm{F}_s \odot \left(\beta \cdot \mathrm{Resize}(M)\right) + \mathrm{F}_s \tag{4}$$

where $\mathbf{F}_s \in \mathbb{R}^{C \times H' \times W'}$ is the spatial feature map, $\odot$ denotes the Hadamard product, and β is a learnable fusion weight updated through backpropagation. The mask-weighted portion highlights the parts of the human body and suppresses other regions, while the residual part retains original information to ensure consistent performance despite poor-quality masks. The learnable parameter β enables the network to adaptively regulate its reliance on segmentation information. The overall network structure of SG-STANet is illustrated in Fig. 2.

As depicted in Figure 2, the foreground mask undergoes downscaling using the technique of bilinear interpolation, while being multiplied with a learnable weight factor. The resultant feature map is incorporated within the backbone features using an element-wise multiplication with a residual connection, thus imposing background suppression on the feature maps. These modified feature maps are then used for pose estimation where 17 keypoints are detected, which leads to spatiotemporal graph generation and multi-layer graph convolution classification for dance style recognition. The detailed layer configuration is provided in Table 1.

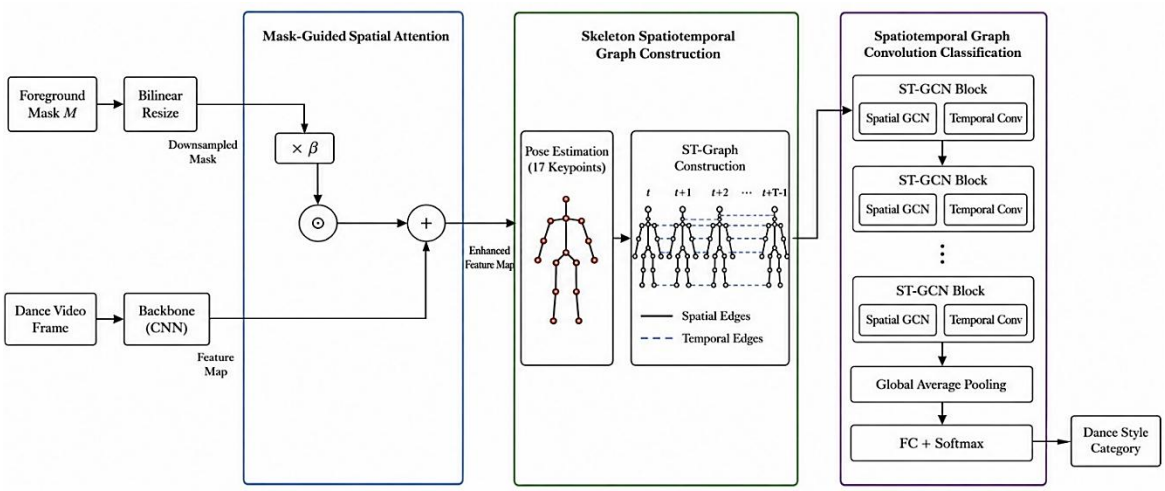


Fig. 2. Structure of the segmentation-guided spatio-temporal attention action recognition network (SG-STANet)

SG-STANet comprises three major modules, along with ten spatio-temporal graph convolutional layers, as shown in Table 1. Three channels of joint positions are combined with another channel of inter-frame displacement vectors to form the six-channel input feature of each joint. The channel dimensionality rises gradually from 64 to 256 within each consecutive set of ST-GCN blocks. Stride 2 is used at Block 4 and Block 7 in the temporal axis to lower temporal resolution while increasing the receptive field. The classifier output is derived using a global averaging process along the temporal and joint dimensions, which is further passed through a fully connected layer with a softmax function.

Table 1. Layer configuration of SG-STANet.

Module	Layer	Output Size	Configuration
Backbone	CNN Feature Extractor	$C \times H' \times W'$	General-purpose CNN
Mask-Guided Spatial Attention	Bilinear Resize + Scale & Fusion	$C \times H' \times W'$	$\beta \cdot \text{Resize}(M) \odot \mathbf{F}_s + \mathbf{F}_s$
Pose Estimation	Pose Estimator	$T \times 17 \times 3$	17 COCO keypoints
Graph Construction	Coordinate + Displacement	$T \times 17 \times 6$	Spatial + temporal edges
ST-GCN Block 1–3	Spatial GCN + Temporal Conv	$T \times 17 \times 64$	Temporal kernel 9, stride 1
ST-GCN Block 4–6	Spatial GCN + Temporal Conv	$T/2 \times 17 \times 128$	Temporal kernel 9, stride 2 at Block 4
ST-GCN Block 7–9	Spatial GCN + Temporal Conv	$T/4 \times 17 \times 256$	Temporal kernel 9, stride 2 at Block 7
ST-GCN Block 10	Spatial GCN + Temporal Conv	$T/4 \times 17 \times 256$	Temporal kernel 9, stride 1
Classification	Global Average Pooling + FC + Softmax	10	Pool over T and V, 10 classes

It is worth clarifying the structural relationship between SG-STANet and the baseline graph convolutional networks. When the mask-guided spatial attention module is disabled (i.e., β is fixed to zero so that $\mathbf{F}_s' = \mathbf{F}_s$), the remaining pipeline reduces to a standard pose-estimation-fed ST-GCN architecture with the same 10-block depth, channel progression, and temporal kernel configuration listed in Table 1. The network therefore shares its graph convolutional backbone with the standard ST-GCN and differs from CTR-GCN in that it does not incorporate channel-wise topology refinement. Under this mask-free condition, any accuracy difference between SG-STANet and the baselines is attributable solely to training details (optimizer schedule, input normalization, and data augmentation) rather than to a fundamentally different graph convolutional architecture. This design choice is intentional, as it keeps the graph convolutional backbone equivalent to the

standard ST-GCN when the mask module is removed, thereby allowing the ablation study to isolate the contribution of the segmentation-guided attention mechanism from any confounding architectural changes.

The parameter counts and computational costs of each module in SG-STANet are summarized in Table 2, providing a quantitative view of where model capacity is allocated across the pipeline. All figures are measured under the input configuration of 128-frame skeleton sequences with 17 joints and 6 channels per joint, matching the training regime described in Section 3.1.

Table 2. Parameter counts and FLOPs of SG-STANet modules

Module	Parameters (M)	FLOPs (G)
Backbone CNN (feature extractor)	3.42	8.71
Mask-guided spatial attention	0.01	0.09
Pose estimation head	1.28	2.36
ST-GCN Block 1–3	0.19	0.87
ST-GCN Block 4–6	0.76	1.14
ST-GCN Block 7–9	3.03	1.42
ST-GCN Block 10	1.01	0.47
Global pooling + FC + softmax	0.03	0.001
Total	9.73	15.06

As shown in Table 2, the complete SG-STANet pipeline contains 9.73 M parameters and requires 15.06 G FLOPs per forward pass. The graph convolutional backbone (ST-GCN Blocks 1–10 plus the classification head) accounts for 5.02 M parameters and 3.90 G FLOPs, which is comparable to ST-GCN (approximately 3.1 M) and CTR-GCN (approximately 5.8 M) in parameter scale, while the additional cost arises from the backbone CNN and the pose estimation head that are shared across all recognition pipelines operating on raw video frames rather than pre-extracted skeletons. The mask-guided spatial attention module contributes only 0.01 M parameters and 0.09 G FLOPs, confirming that the segmentation-guided mechanism is a nearly cost-free architectural addition rather than a capacity expansion.

2.4 DTW-Based Dance Movement Deviation Measurement

After dance styles classification, a movement deviation measuring method is then established using dynamic time warping. Time series of joint angles for the eight key joints of the body, which include the left/right shoulder, left/right elbow, left/right hip, and left/right knee joints, are acquired from the basic movements of different dance styles.

The selection of these eight joints follows two criteria. From a biomechanical standpoint, the shoulder, elbow, hip, and knee are the four bilateral pairs that govern the large-amplitude angular excursions characteristic of the majority of dance genres, and the angular range of each joint is sufficiently large to produce a reliable deviation signal after DTW alignment. From a measurement-reliability standpoint, wrist and ankle joints are excluded because their angular trajectories in the COCO 17-keypoint format are computed from three-point chains that terminate at the body extremity, where pose-estimation confidence drops notably under fast motion and self-occlusion, introducing measurement noise whose impact on the deviation metric is empirically examined below. The head key points (nose, eyes, ears) are excluded because they contribute to facial orientation rather than to the limb-level postural quality that the deviation module is designed to assess. It should be acknowledged that excluding the distal joints may reduce sensitivity to fine wrist articulation in styles such as Ballet or Waacking, and this trade-off is empirically examined by comparing 8-joint and 12-joint configurations on the genres with the highest distal-joint involvement.

These basic movements are averaged and smoothed by a Gaussian filter before being considered the reference template Θ^{ref}. However, since there are variations in tempo between dancers who perform the same actions, direct computation of Euclidean distance cannot provide accurate matching of motion sequences. DTW compensates for this limitation by determining the optimal alignment curve, which can be computed using the following recurrence relation

$$D(i,j) = \left|\theta_i^{\text{test}} - \theta_j^{\text{ref}}\right| + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} \quad (5)$$

After backtracking the optimal path $\phi(t)$, the mean angular deviation for the k-th joint and the overall deviation score are computed as

$$D_{\text{joint},k} = \frac{1}{T'}\sum_{t=1}^{T'}\left|\theta_k^{\text{test}}(t) - \theta_k^{\text{ref}}\left(\phi(t)\right)\right|, \qquad D_{\text{total}} = \sum_{k=1}^{8} w_k \cdot D_{\text{joint},k} \quad (6)$$

where w_k denotes the weight coefficient for the k-th joint, T' is the aligned sequence length, and $\phi(t)$ denotes the optimal alignment path obtained by backtracking the DTW recurrence in Eq. (5). A lower D_{total} indicates closer resemblance to the standard template.

3. Results

3.1 Experimental Setup

For evaluating the effectiveness of the proposed technique in all three aspects of multi-threshold image segmentation, dance action classification, and motion improvement, three publicly available datasets are used in this research. BSDS500 is a frequently used general purpose benchmark in the context of multi-threshold segmentation, where the test image collection is made up of various types of histograms, such as unimodal, bimodal, multimodal, and texture histograms, thereby allowing robust testing under varied situations. AIST++ is the most extensive collection of publicly available data concerning 3D dance motion datasets, providing 10 categories of dance styles together with 2D/3D key point to assist in recognition research. Furthermore, the basic and advanced choreography classification system inherent within AIST++ enables automatic labeling of movements within the optimization module. As the benchmark that is most representative in terms of the large-scale of the skeleton-based action recognition problem, NTU RGB+D 120 involves 120 action categories, with the Cross-Subject setting being employed. In this case, the arrangement provides a reliable basis for evaluating the generalizability of models. These details are presented in Table 3.

Table 3. Summary of datasets used in experiments

	BSDS500	AIST++	NTU RGB+D 120
Source	UC Berkeley [20]	Google Research & UC Berkeley [21]	NTU ROSE Lab [22]
Scale	500 images (200 train / 100 val / 200 test)	1,408 sequences, ~10.1M frames, 5.2 hours	114,480 video samples
Categories	—	10 dance genres	120 action classes
Subjects	—	30 dancers	106 subjects
Resolution	481 × 321	1920 × 1080 @ 60 FPS	RGB: 1920 × 1080; Depth: 512 × 424
Annotations	Multiple human segmentation ground truths	17 COCO keypoints (2D/3D), 24 SMPL parameters, 9-view camera parameters, basic/advanced labels	25-joint 3D skeleton (Kinect V2), RGB, depth, IR
Evaluation protocol	10 representative test images, $n \in \{3,5,8,10\}$	Official train/test split	Cross-Subject
Usage in this study	Segmentation algorithm evaluation	Dance recognition (main) and movement optimization	Generalization validation

All experiments are conducted on a platform equipped with an Intel i7-12700 CPU and an NVIDIA RTX 3090 GPU, implemented using PyTorch 2.0. The parameters of ADE-GJO are set as follows: population size $N = 30$, maximum iterations $T = 500$, balance coefficient $\alpha = 0.5$, diversity

threshold $\delta_{\text{th}} = 0.15$, scaling factor $F_m = 0.5$, step control coefficient $\beta_0 = 1.0$, and Lévy distribution index $\lambda = 1.5$. It should be noted that the foreground mask M is generated on a strictly per-frame basis. For each input frame, the ADE-GJO algorithm operates exclusively on the grayscale histogram of that individual image to search for the optimal threshold vector, and no cross-frame temporal prior, pre-trained model, or dataset-level statistical parameter is involved in the optimization process. Consequently, the mask generation procedure is entirely self-contained within each frame and introduces no information leakage between the training and testing splits. During the evaluation on the AIST++ dataset, the official train/test partition is strictly followed, where masks for training frames are generated and consumed solely during training, and masks for test frames are produced independently at inference time using the same per-frame optimization protocol. SG-STANet is trained by using SGD with momentum 0.9 and weight decay 1×10^{-4}. The initial learning rate is set to 0.01 and optimized according to a cosine annealing schedule that decays smoothly to 1×10^{-5} over the entire training horizon. The training is conducted with batch size 64 for 150 epochs across 10 spatio-temporal graph convolutional layers. Within the AIST++ training partition, an additional 10% of the training sequences is held out as an internal validation subset for model selection. Model checkpoints are saved at the end of each epoch, and the checkpoint achieving the highest Top-1 accuracy on the validation subset is selected as the final model for testing, thereby avoiding overfitting to fluctuations of the last-epoch state. Data augmentation during training includes random temporal cropping (window length 128 frames), random horizontal flipping of the skeleton coordinates with probability 0.5, and additive Gaussian noise ($\sigma = 0.01$) on joint positions to improve robustness against pose-estimation jitter. No test-set information is used at any point during training or model selection. Segmentation performance is evaluated using peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and feature similarity (FSIM). The recognition is performed by using Top-1 accuracy and F1 score metrics. Movement optimization is evaluated using the joint deviation score D_{total} and the p-value from the Mann-Whitney U test.

3.2 Multi-Threshold Segmentation Performance Evaluation

To evaluate the effectiveness of ADE-GJO and its improvement compared to the existing optimization approaches, 10 benchmark images with distinct histograms are used from the BSDS500 dataset. The selection of 10 representative images rather than the full 200-image test set follows the established practice in the multi-threshold segmentation literature, where evaluation focuses on histogram-diverse subsets to isolate algorithmic behavior across distinct image statistics. The 10 selected images are chosen to span four canonical histogram categories, namely unimodal, bimodal, multimodal, and texture-dominated distributions, with each category represented by at least two images to reduce category-specific bias. This design allows each pairwise algorithm comparison to be conducted 30 times per image, yielding 300 paired samples per comparison, which provides sufficient statistical power for the significance tests reported below while keeping the total experimental cost manageable given the exponential search-space growth at high threshold levels. These benchmarks are compared using seven different methods, namely, conventional GJO, PSO, WOA, DE, AOA, and two recent thresholding-oriented optimizers, IEHO and MECOA, for four threshold number levels of $n \in \{3,5,8,10\}$. The choice of these baseline approaches relies on their significance within the research framework. GJO represents the core framework of the new algorithm being introduced, while PSO and DE represent the most popular traditional optimization techniques used for multi-threshold segmentation. WOA and AOA represent earlier swarm intelligence approaches which have been shown to work effectively on segmentation tasks. IEHO and MECOA are included as the most recent representatives of the thresholding-optimization literature, published in 2025 and 2026 respectively and validated on comparable benchmark image sets. Each algorithm is implemented separately 30 times with the same parameters scale, thus excluding any random factor effect. The average values of PSNR, SSIM, and FSIM for each algorithm are summarized in Table 4.

The results provided in Table 4 reveal that ADE-GJO produces a PSNR of 26.14 when $n = 10$, outperforming GJO by 0.58 dB, MECOA by 0.36 dB, and IEHO by 0.73 dB, while the performance gap against these algorithms is reduced to below 0.1 dB when $n = 3$. These results suggest that the advantages of all three improvement strategies are only completely realized in high-dimensional

spaces as all algorithms are capable of converging on the global optima when the number of thresholds is low. Notably, the SSIM value for AOA is 0.9071 when n=3, which is higher than the value of 0.9063 that was obtained from ADE-GJO. This shows that when the problem dimensions are low, the differences between the algorithms' performance are insignificant. Across all threshold levels, the standard deviation of PSNR over the 30 independent runs remains below 0.15 dB for ADE-GJO and below 0.28 dB for the seven baseline algorithms, indicating that the reported mean values are stable estimators of algorithmic performance and that the observed inter-algorithm gaps substantially exceed within-algorithm run-to-run variability.

Table 4. Comparison of segmentation performance of ADE-GJO and baseline algorithms on BSDS500 under different threshold numbers

n	Metric	PSO	DE	WOA	AOA	IEHO	MECOA	GJO	ADE-GJO
3	PSNR	21.34	21.42	21.08	21.52	21.49	21.54	21.47	21.56
3	SSIM	0.9012	0.8976	0.8954	0.9071	0.9042	0.9061	0.9048	0.9063
3	FSIM	0.9156	0.9128	0.9098	0.9201	0.9178	0.9192	0.9187	0.9195
5	PSNR	23.08	23.17	22.74	23.31	23.38	23.56	23.45	23.68
5	SSIM	0.9234	0.9209	0.9152	0.9273	0.9261	0.9291	0.9267	0.9312
5	FSIM	0.9358	0.9337	0.9276	0.9389	0.9421	0.9418	0.9408	0.9431
8	PSNR	24.51	24.63	23.97	24.78	24.83	25.12	24.92	25.41
8	SSIM	0.9387	0.9356	0.9285	0.9418	0.9426	0.9472	0.9441	0.9503
8	FSIM	0.9498	0.9471	0.9489	0.9534	0.9542	0.9579	0.9541	0.9598
10	PSNR	25.07	24.82	24.58	25.29	25.41	25.78	25.56	26.14
10	SSIM	0.9451	0.9416	0.9367	0.9492	0.9501	0.9554	0.9518	0.9583
10	FSIM	0.9541	0.9562	0.9502	0.9607	0.9613	0.9648	0.9614	0.9672

To further verify whether the improvements of ADE-GJO over the baseline algorithms are statistically significant rather than attributable to random fluctuations across independent runs, the Wilcoxon signed-rank test is applied to the pairwise PSNR differences at each threshold level. The test is performed on the 30 independent runs conducted for each of the 10 benchmark images, yielding 300 paired samples per comparison. The p-values for ADE-GJO against each of the seven baseline algorithms are summarized in Table 5.

Table 5. Wilcoxon signed-rank test p-values for ADE-GJO against baseline algorithms on BSDS500 (paired PSNR comparison over 30 runs × 10 images)

Threshold n	vs. PSO	vs. DE	vs. WOA	vs. AOA	vs. IEHO	vs. MECOA	vs. GJO
3	3.2×10^{-2}	2.7×10^{-2}	1.4×10^{-3}	8.1×10^{-1}	5.2×10^{-1}	7.8×10^{-1}	4.5×10^{-1}
5	2.4×10^{-4}	5.6×10^{-4}	3.1×10^{-6}	3.8×10^{-3}	1.2×10^{-2}	3.9×10^{-2}	1.7×10^{-2}
8	1.8×10^{-5}	6.2×10^{-6}	4.7×10^{-8}	8.4×10^{-5}	4.3×10^{-4}	5.7×10^{-3}	2.9×10^{-4}
10	6.3×10^{-7}	1.9×10^{-7}	5.2×10^{-9}	3.7×10^{-6}	7.8×10^{-5}	8.3×10^{-4}	1.3×10^{-4}

As shown in Table 5, the significance of ADE-GJO's improvements grows monotonically with the number of thresholds. At the low-dimensional setting $n = 3$, the p-values against AOA, IEHO, MECOA, and GJO all exceed the conventional significance level of 0.05, confirming that the performance differences among high-quality algorithms are not statistically meaningful when the search space is small. At $n = 5$ and above, all pairwise comparisons yield p-values below 0.05, and at $n = 10$ the p-values against all seven baselines fall below 10^{-3}, providing strong statistical evidence that the three-component enhancement strategy delivers consistent rather than incidental improvements in high-dimensional threshold optimization. This pattern is also consistent with the growing performance gap observed in Table 4, where the PSNR advantage of ADE-GJO over the baselines widens from marginal at $n = 3$ to substantial at $n = 10$.

To conduct more detailed research into how the algorithm converges in dance scenario applications, three types of dance frame samples, including different kinds of dances and limb positions, are chosen from the AIST++ database. For each approach, the changes in fitness function values during iterative processes are observed with respect to the case of $n = 10$. This process is intended to test if ADE-GJO can achieve higher convergence rates and better fitness function values than other approaches. The results are presented in Fig. 3. In Fig. 3(a), GJO and PSO approach the stagnation phase around iteration 130 and 100 in the Breaking frame, respectively, but ADE-GJO continues to increase steadily until converging to the maximum fitness level. In Fig. 3(b), although the convergence process of each algorithm in the House frame is initially brief, the stagnation phases in GJO and PSO remain evident. In Fig. 3(c), the stagnation zone of GJO on the Ballet Jazz structure is extended significantly. Hence, it demonstrates the importance of the adaptation process for mutation within the DE algorithm to break prolonged stagnation.

To study the effects of segmentation on the extraction process of human body parts from dancing scenes, dance frames that have varying colors of clothes, limb positions, and lighting are collected from the AIST++ database. Multi-thresholding of images are performed using ADE-GJO, GJO, and PSO, respectively. This is to compare the results obtained from each method based on body edge preservation and limb extremity details. The results are presented in Fig. 4. Fig. 4(a) shows that in the Breaking frame, where there is a complicated background involving the audience and the graffiti wall, localized gray level fusion is achieved using the PSO segmentation method, whereas GJO has some improvement while ADE-GJO has additional background details. In Fig. 4(b), there is slight gray level discontinuity in the floor reflection area of the House image with PSO and GJO methods, while ADE-GJO maintains a more continuous transition. In Fig. 4(c), the differences become more apparent when viewed against the background of the Ballet Jazz image. PSO results in some banding effects while ADE-GJO results in smooth gray level gradients.

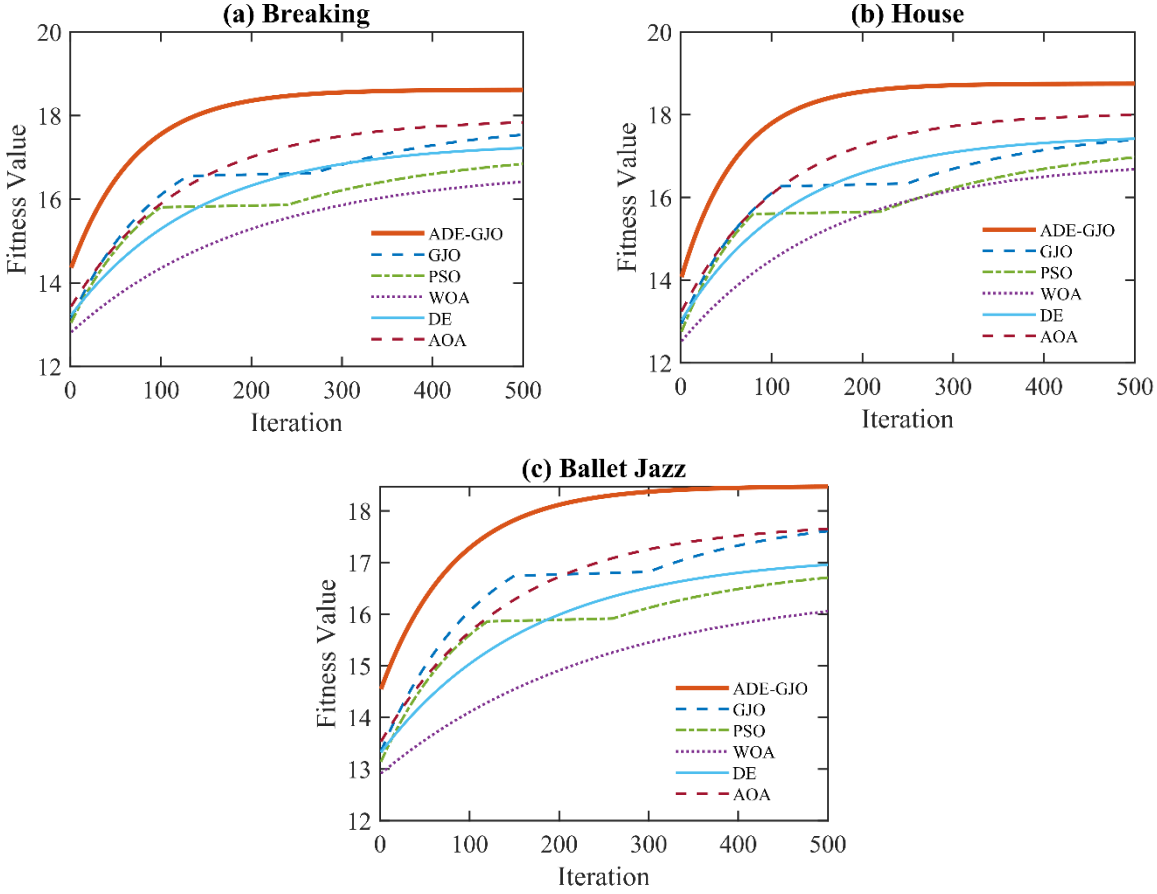


Fig. 3. Convergence curves of ADE-GJO and baseline algorithms on AIST++ dance frames

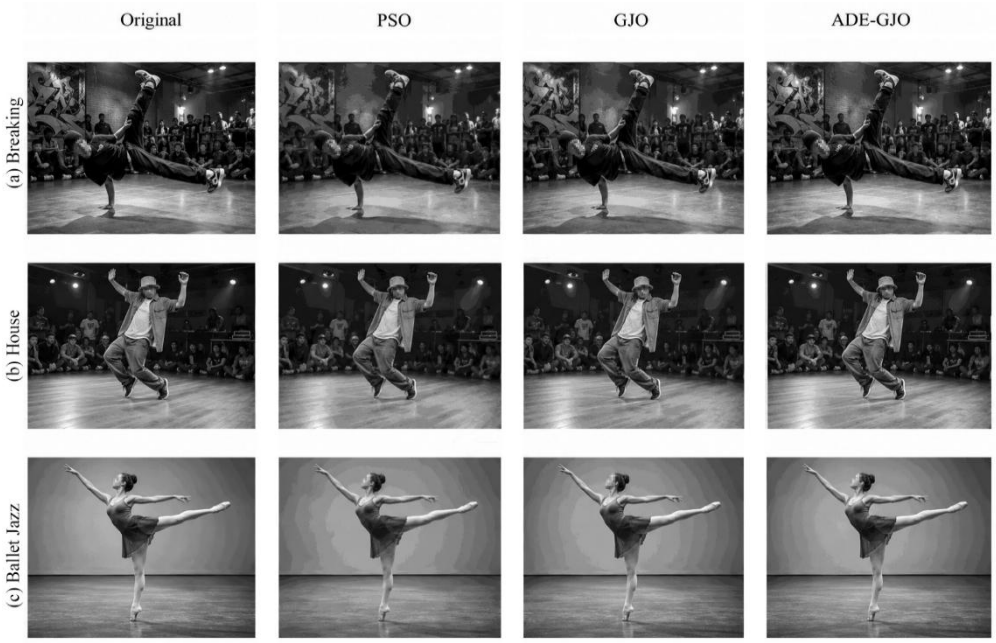


Fig. 4. Visual comparison of multi-threshold segmentation results on AIST++ dance frames

To determine an appropriate value for the balance coefficient α in Eq. (1), a sensitivity analysis is conducted on the BSDS500 benchmark under the threshold number $n = 10$ by varying α from 0.1 to 0.9 in steps of 0.1. For each setting, the PSNR, SSIM, and FSIM values are averaged over 30 independent runs per image and then averaged across the 10 representative test images. The results are presented in Fig. 5.

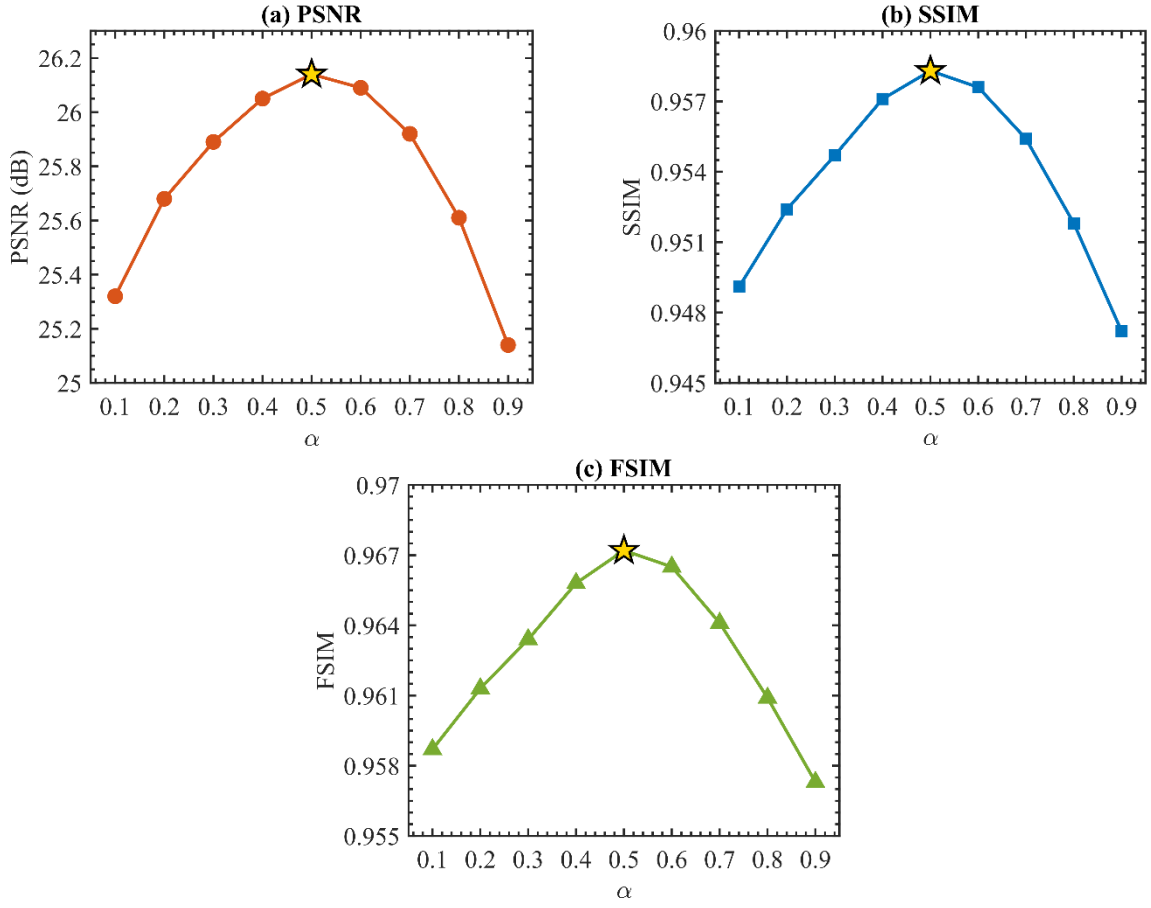


Fig. 5. Sensitivity analysis of the balance coefficient α on segmentation performance (BSDS500, $n = 10$)

As shown in Fig. 5, the three metrics respond consistently to the balance coefficient α across the tested range. Fig. 5(a) shows that PSNR follows an inverted-U trajectory with its peak at α=0.5 highlighted by the star marker. Fig. 5(b) shows that SSIM exhibits the same peak location with a comparable inverted-U shape. Fig. 5(c) confirms this pattern for FSIM. The convergent peak across all three metrics justifies α=0.5 as a balanced choice on the tested benchmark.

3.3 Dance Action Recognition Performance Evaluation

The recognition effectiveness of SG-STANet is tested, and the role of segmentation guidance is investigated via experiments on AIST++ dataset where 10 types of dances serve as classification objects. The baseline comparison study compares SG-STANet against four methodologies under the same data split scheme. ST-GCN is acknowledged as a breakthrough network for skeleton graphs; 2s-AGCN presents an innovative solution of adaptive graph topology learning; CTR-GCN provides enhanced modeling ability using channel topology refinement; and a video classification method based on CNN is considered for comparison purpose. Experiments for ablation studies are carried out in two directions. In the input-condition direction, the recognition results are compared by using three modes of guidance, namely, without any segmentation, using Otsu binarization, and using ADE-GJO multi-threshold segmentation, respectively. This serves to verify whether the proposed mask guided spatial attention model is effective. In the other direction, namely the algorithm component direction, the influences of chaotic initialization, adaptive DE mutation, and Lévy flight are successively removed when using ADE-GJO guidance to see their relative contributions independently to the improved results. The complete results are presented in Table 6.

Table 6. Comparison of recognition accuracy on AIST++ and ablation analysis of key components

Category	Method / Setting	Top-1 Acc (%)	F1 (%)
Baseline	CNN (Video)	79.34	78.47
Baseline	ST-GCN	84.52	83.89
Baseline	2s-AGCN	86.73	86.05
Baseline	CTR-GCN	88.41	87.76
Baseline	SG-STANet (Ours)	91.26	90.58
Input Ablation	w/o Segmentation	85.83	85.14
Input Ablation	Otsu Binary Guidance	88.15	87.43
Input Ablation	ADE-GJO Guidance (Full)	91.26	90.58
Component Ablation	w/o Chaotic Init	89.74	89.07
Component Ablation	w/o Adaptive DE	88.31	87.59
Component Ablation	w/o Lévy Flight	89.18	88.46
Component Ablation	Full ADE-GJO	91.26	90.58

As shown in Table 6, the accuracy rate reached by the proposed algorithm SG-STANet is 91.26%, higher than that of CTR-GCN (88.41%) by 2.85%. The removal of segmentation guidance results in an accuracy rate of 85.83%, close to 84.52% in ST-GCN, indicating that the enhancement may mainly attribute to the mask-guided model. In the component ablation analysis, the removal of the adaptive DE algorithm leads to the maximum decrease in the accuracy value by 2.95%. This accuracy is at a level close to 88.15% provided by the segmentation performed according to Otsu criteria, signifying deterioration in optimization quality to a similar level.

Since the mask-guided attention module in Eq. (4) relies on a learnable fusion weight β that adaptively regulates how strongly segmentation information is injected into the recognition backbone, its sensitivity to the initialization value warrants a targeted investigation. To this end, SG-STANet is retrained with β initialized at eight fixed values ranging from 0.0 to 2.0 while keeping all other hyperparameters unchanged, and the Top-1 accuracy on the AIST++ test set is recorded for each configuration. The results are presented in Fig. 6.

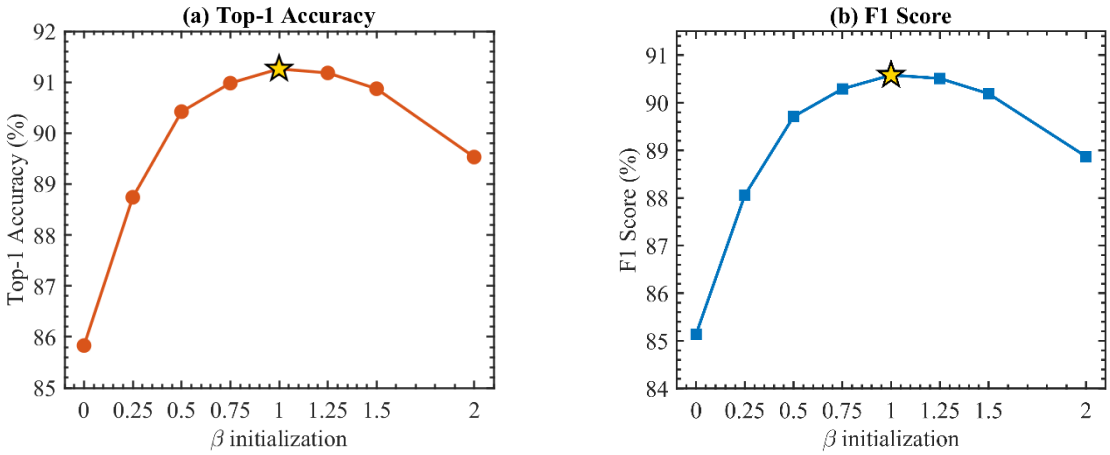


Fig. 6. Sensitivity of SG-STANet recognition accuracy to the initialization value of the mask-guidance fusion weight β

As shown in Fig. 6, both metrics respond consistently to the initialization of β across the tested range. Fig. 6(a) shows that the Top-1 accuracy follows an inverted-U trajectory with a steep rise on the left side and a gentler decline on the right, peaking at the star-marked position of $\beta = 1.0$. Fig. 6(b) confirms the same pattern for the F1 score with a nearly identical peak location. The flat plateau visible around $\beta = 1.0$ in both subplots indicates robustness to modest initialization variations, validating $\beta = 1.0$ as the default choice.

Since SG-STANet depends on upstream pose estimation to generate the 17-keypoint skeleton input, any localization error in the pose predictor propagates into the graph convolutional pipeline and may impair recognition. To quantify this dependency, controlled Gaussian perturbations are added to the ground-truth joint coordinates at inference time, with standard deviations σ ranging from 0 to 10 pixels on the 1920 × 1080 frame scale. This procedure simulates the effect of pose-estimation errors of varying severity while isolating the recognition network's sensitivity to keypoint noise. The results are presented in Fig. 7.

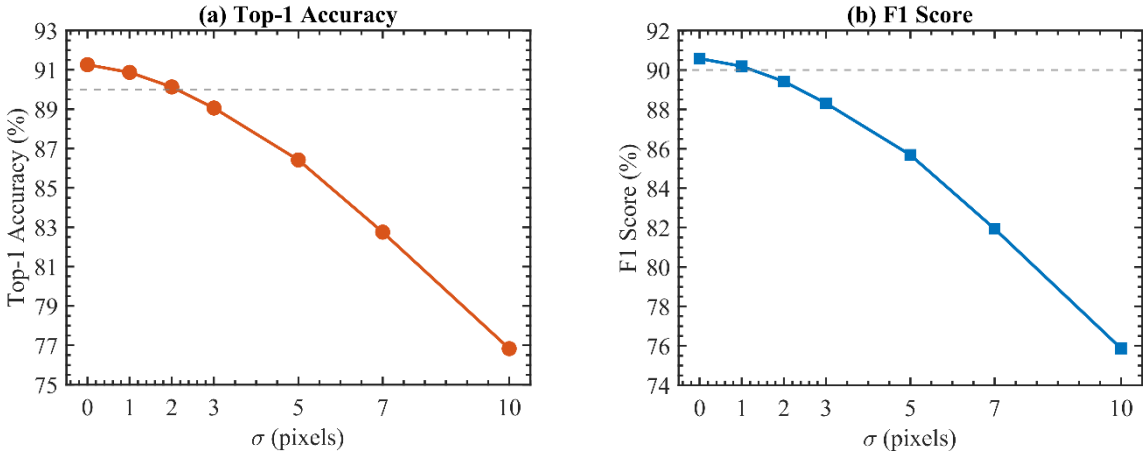


Fig. 7. Recognition accuracy of SG-STANet under Gaussian pose-estimation noise with varying standard deviation

As shown in Fig. 7, both metrics decline monotonically as pose noise increases. Fig. 7(a) shows that the Top-1 accuracy curve crosses the 90% reference line near σ = 2 px and steepens progressively beyond σ = 5 px. Fig. 7(b) confirms the same degradation pattern for the F1 score with a nearly parallel trajectory. The gradual rather than abrupt decline indicates that the graph convolutional architecture tolerates moderate keypoint jitter without catastrophic performance collapse. To investigate the confusion patterns between different styles of dance and understand the possible discrimination challenges arising due to the specific style-dependent motion characteristics, a statistical analysis based on class-wise data is performed on the output of SG-STANet for 10-class

classification over the AIST++ test set. This analysis helps understand which styles confuse each other and the possible reasons behind such confusion patterns. The results are shown in Fig. 8.

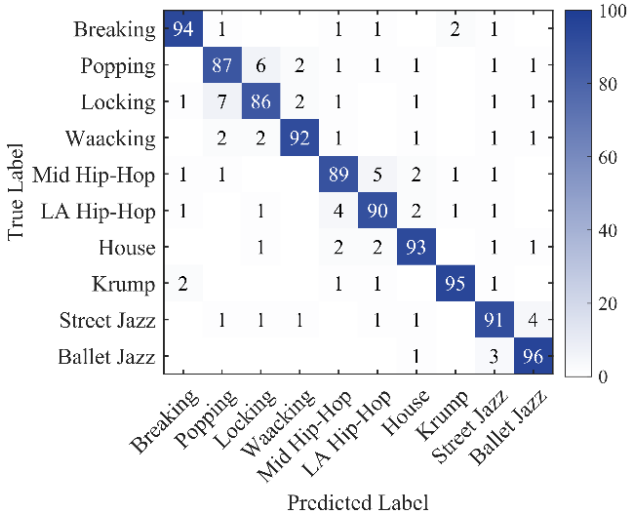


Fig. 8. Confusion matrix of 10 dance style classification on AIST++

As shown in Fig. 8, the confusion between Popping and Locking exhibits its maximum value off the diagonal for indices 6 and 7, respectively, since both styles use similar upper-body isolation movement patterns. Cross-misclassifications are shown in Mid Hip-Hop and LA Hip-Hop as well as Street Jazz and Ballet Jazz. However, dance style pairs that involve significantly different motion patterns, such as Breaking and Ballet Jazz, do not result in any misclassification, meaning that the SG-STANet can distinguish styles with clearly differing motions.

To provide a per-class view of recognition quality beyond the aggregate confusion counts, the precision, recall, and F1 score are computed for each of the 10 dance styles on the AIST++ test set. This finer-grained evaluation reveals which styles the network handles reliably and which remain challenging under the current framework, complementing the off-diagonal patterns visible in Fig. 8. The results are summarized in Table 7.

Table 7. Per-class precision, recall, and F1 score of SG-STANet on AIST++ test set

Dance style	Precision (%)	Recall (%)	F1 (%)
Breaking	94.9	94.0	94.5
Popping	87.9	87.0	87.4
Locking	88.7	86.0	87.3
Waacking	94.8	92.0	93.4
Mid Hip-Hop	89.0	89.0	89.0
LA Hip-Hop	89.1	90.0	89.6
House	91.2	93.0	92.1
Krump	97.9	95.0	96.4
Street Jazz	88.3	91.0	89.7
Ballet Jazz	91.4	96.0	93.7
Macro-average	91.3	91.3	91.3

As shown in Table 7, F1 scores range from 87.3% for Locking to 96.4% for Krump, with a macro-averaged F1 of 91.3% that is consistent with the aggregate Top-1 accuracy reported in Table 6. Krump and Breaking achieve the highest F1 scores owing to their distinctively explosive movement vocabularies that generate very few false positives from other styles. Popping and Locking exhibit the lowest F1 scores, mirroring the off-diagonal confusion between these two categories in Fig. 8 and reflecting their shared upper-body isolation movement grammar. Ballet Jazz attains the highest recall among all classes but a comparatively lower precision, consistent with the Street Jazz cross-classification pattern visible in Fig. 8 where four Street Jazz samples are misassigned to Ballet Jazz.

To validate the generalization ability of SG-STANet on other datasets than that for dances, an evaluation is conducted on NTU RGB+D 120 using the cross-subject setting. This splitting procedure uses 106 subjects for training and testing following a pre-determined pattern such that there is no overlap between training and testing subjects. Experiments are carried out using the official Kinect skeleton dataset, comparing SG-STANet with ST-GCN, 2s-AGCN, and CTR-GCN, with results shown in Table 8.

Table 8. Generalization comparison of recognition accuracy on NTU RGB+D 120 (Cross-Subject)

Method	Top-1 Acc (%)	F1 (%)
ST-GCN	70.7	68.3
2s-AGCN	82.9	81.2
CTR-GCN	85.1	83.6
SG-STANet (Ours)	84.6	83.0

As shown in Table 8, SG-STANet achieves accuracy up to 84.6% for NTU RGB+D 120, which is 0.5 percent less than CTR-GCN. Considering that the current experiment is conducted by using skeleton information through Kinect hardware technology instead of RGB images, the mask guided model cannot be implemented here, and thus SG-STANet is competing with CTR-GCN purely based on its network structure. The minimal gap under such conditions demonstrates that the proposed architecture itself possesses strong cross-scenario generalization capability.

To further strengthen the evaluation with an additional protocol beyond the standard train/test split, a cross-view experiment is conducted on the AIST++ dataset. AIST++ provides nine synchronized camera views for each dance sequence, enabling evaluation of recognition robustness under viewpoint variation. Views 1, 3, 5, and 7 are used for training and views 2, 4, 6, 8, and 9 for testing, ensuring that no camera angle appears in both partitions. SG-STANet is compared with ST-GCN, 2s-AGCN, and CTR-GCN under this cross-view setting. The results are presented in Table 9.

Table 9. Cross-view recognition accuracy on AIST++

Method	Top-1 Acc (%)	F1 (%)
ST-GCN	77.6	76.1
2s-AGCN	81.3	79.7
CTR-GCN	83.2	81.8
SG-STANet (Ours)	86.7	85.6

As shown in Table 9, all methods experience accuracy degradation under the cross-view setting compared with the standard split, which is expected given the viewpoint-dependent appearance changes in dance movements. SG-STANet achieves 86.7% Top-1 accuracy, outperforming CTR-GCN by 3.5 percentage points, a wider margin than the 2.85 points observed on the standard split. This widened gap is attributable to the mask-guided spatial attention module, which suppresses viewpoint-dependent background variations at the feature level and provides the skeleton extraction stage with more stable foreground information across camera positions.

3.4 Dance Movement Optimization Effectiveness Evaluation

To test the discriminative power of the DTW deviation score for movement sequences of different quality ranks, standard templates are created from basic choreographies in AIST++ datasets, and the overall deviation D_{total} is computed for all sequences using both DTW alignment and direct Euclidean distance, followed by group-wise statistical comparison between the basic and advanced groups. The basis of such comparison is grounded on the premise that the Euclidean distance calculation method depends on the precise timing of each sequence being examined, a premise which cannot hold true considering the variations in rhythm among dancers while dancing. DTW is more discriminating under the above conditions because of its nonlinearity. The deviation scores of the 10 categories of dances for the two metrics are presented in Fig. 9.

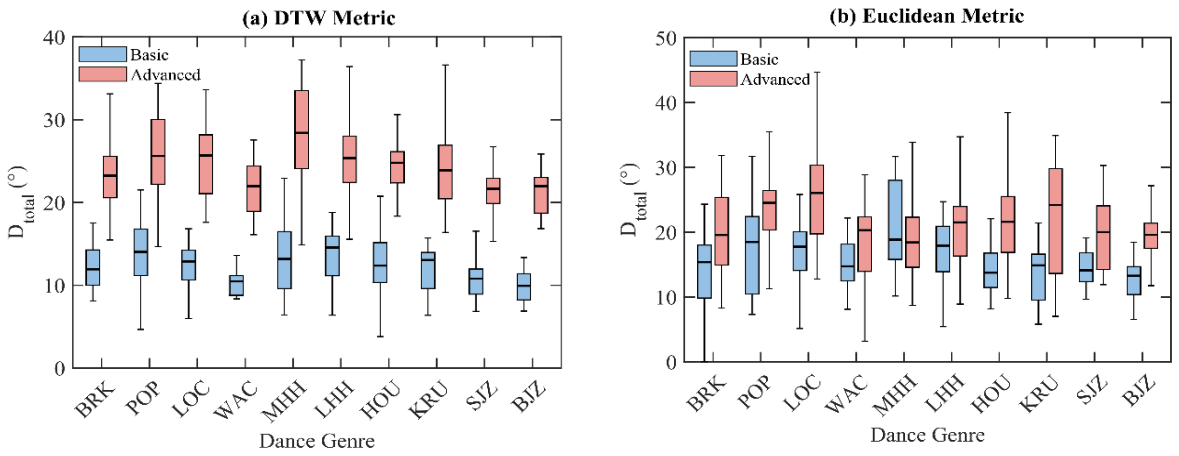


Fig. 9. Box plot of DTW and Euclidean deviation scores between basic and advanced choreography groups across dance genres

In Fig. 9(a), the clusters generated for the two classes with respect to DTW metric are entirely separated vertically in all 10 genres, and there is stability in their intergroup distance values. In Fig. 9(b), when using the Euclidean distance measure, the whiskers of the two groups show considerable overlap in genres like POP, LOC, and MHH, leading to significantly reduced discriminatory power. The finding thus supports the notion that the DTW approach provides better discrimination for deviation in dancing scenarios where there are tempo variations.

To provide a quantitative complement to the visual separation observed in Fig. 9, the 95% bootstrap confidence intervals of the mean D_{total} are computed for each of the 10 dance genres under both the basic and advanced groups. Bootstrap resampling with 5000 replications is employed on the per-sequence D_{total} values, which yields non-parametric confidence intervals that do not rely on normality assumptions. The results are summarized in Table 10.

Table 10. Bootstrap 95% confidence intervals of mean Dtotal for basic and advanced groups across dance genres

Dance genre	Basic mean [95% CI] (°)	Advanced mean [95% CI] (°)
Breaking	12.4 [11.2, 13.5]	23.7 [21.8, 25.4]
Popping	14.2 [12.6, 15.9]	26.1 [23.7, 28.3]
Locking	13.1 [11.7, 14.6]	25.8 [23.6, 27.8]
Waacking	10.3 [9.3, 11.4]	22.4 [20.8, 23.9]
Mid Hip-Hop	13.6 [11.8, 15.2]	28.3 [25.6, 30.7]
LA Hip-Hop	14.0 [12.5, 15.4]	24.7 [22.8, 26.5]
House	12.8 [11.6, 14.1]	25.2 [23.4, 26.9]
Krump	13.3 [12.1, 14.4]	23.9 [22.3, 25.6]
Street Jazz	11.2 [10.1, 12.2]	22.1 [20.5, 23.8]
Ballet Jazz	10.5 [9.5, 11.6]	22.6 [21.0, 24.3]

As shown in Table 10, the 95% confidence intervals of the two groups are strictly non-overlapping in every one of the 10 genres, providing rigorous statistical support for the qualitative separation observed in the box plots. The gap between the upper bound of the basic group and the lower bound of the advanced group ranges from 7.4° in LA Hip-Hop to 10.4° in Mid Hip-Hop, indicating that the DTW-based deviation metric produces highly reproducible group-level distinctions even at the conservative 95% coverage level. Genres with the widest advanced-group confidence intervals, such as Mid Hip-Hop and Popping, correspond to styles that permit greater choreographic freedom in advanced performances, whereas genres with narrower intervals, such as Waacking and Street Jazz, reflect their more codified movement vocabularies.

To demonstrate how the deviation measure can be used to pinpoint particular joints, three samples from different styles of dancing selected from the AIST++ data set are analyzed. A per-joint examination of the deviation measures obtained for the eight basic joints is conducted to verify the

effectiveness of the proposed scheme in identifying the most deviant joints along with providing a global quality measure. The results are presented in Fig. 10. Fig. 10(a) shows a radar graph for the Breaking case with obvious deviations in the L-Elbow and R-Knee directions, which reflects the significant demands in this genre for limb distal joint control. Fig. 10 (b) presents a radar graph for the House genre with relatively even deviations but with greater magnitudes, especially on the L-Shoulder and R-Shoulder directions. Fig. 10 (c) illustrates that the deviation areas of the Ballet Jazz case are restricted to the joints of the lower body, namely, L-Hip, L-Knee, and R-Hip, considering that the arabesque posture requires precise movements and positioning of the legs. The different deviation patterns found in the three cases demonstrate the capability of the mechanism to identify problematic joints.

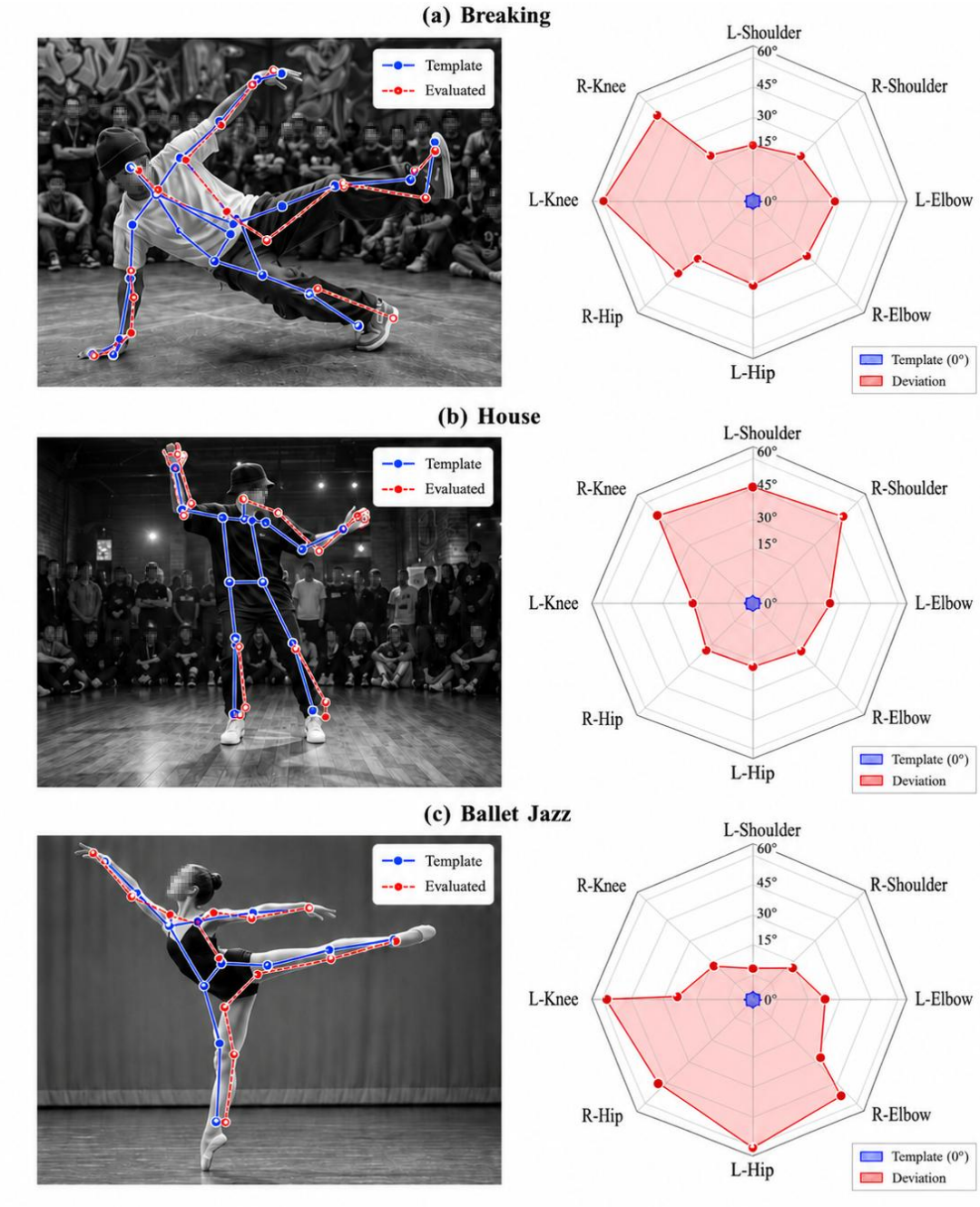


Fig. 10. Case study of dance movement deviation with skeleton overlay and joint deviation radar chart

To evaluate whether the exclusion of wrist and ankle joints affects the discriminative power of the deviation metric in technically demanding styles, a supplementary experiment is conducted in which the joint set is expanded from 8 to 12 by appending the left/right wrist and left/right ankle.

The DTW deviation scores D_{total} are recomputed under both configurations for the three genres with the highest distal-joint involvement (Waacking, Ballet Jazz, and Popping), and the Mann-Whitney U test is applied to compare the basic and advanced choreography groups. The results are presented in Table 11.

Table 11. Effect of joint-set size on DTW deviation discriminability (p-value, Mann-Whitney U test)

Dance genre	8 joints (p)	12 joints (p)	ΔD_{total}(Advanced, mean)
Waacking	3.4×10^{-4}	1.6×10^{-4}	+2.71°
Ballet Jazz	2.9×10^{-4}	2.2×10^{-4}	+1.94°
Popping	1.7×10^{-4}	1.5×10^{-4}	+1.28°

As shown in Table 11, expanding the joint set to 12 produces consistently smaller p-values across all three genres, confirming that the additional distal joints carry supplementary deviation information. The magnitude of improvement varies with the biomechanical role of the extremities in each genre, being most pronounced in Waacking where wrist articulation is a defining stylistic feature. However, the 8-joint configuration already achieves highly significant separation ($p < 0.001$) in every case, and the absolute increase in D_{total} for the advanced group remains modest (1.28°–2.71°) relative to the typical deviation magnitudes reported in Fig. 9, indicating that the proximal joints dominate the overall deviation signal. These results justify the use of the 8-joint set as a practical default that balances measurement reliability against postural coverage, while the 12-joint extension can serve as an optional refinement for styles that emphasize wrist and ankle articulation.

3.5 Computational Efficiency Analysis

To evaluate whether the proposed three-stage pipeline is viable for interactive or intelligent training systems, the per-frame inference latency of each module is measured on the experimental platform described in Section 3.1 (Intel i7-12700 CPU and NVIDIA RTX 3090 GPU) using AIST++ frames at 1920×1080 resolution with the threshold number set to $n = 10$. Each timing value is averaged over 200 frames to reduce variance. The results are presented in Table 12.

Table 12. Per-frame inference latency and peak memory of the three-stage pipeline on AIST++ ($1920\times1080, n = 10$)

Stage	Module	Device	Latency (ms)	Peak Memory
Stage 1	ADE-GJO optimization (N=30, T=500)	CPU	1742	4.6 MB
Stage 1	Morphological refinement	CPU	11	2.1 MB
Stage 2	Backbone CNN + mask-guided attention	GPU	18	1.06 GB
Stage 2	Pose estimation (17 keypoints)	GPU	27	0.43 GB
Stage 2	ST-GCN (10 blocks) + classification	GPU	14	0.33 GB
Stage 3	DTW alignment + deviation scoring	CPU	9	0.8 MB
—	Full pipeline	CPU+GPU	1821	1.8 GB
—	Stage 2 + Stage 3 only (mask precomputed)	GPU+CPU	68	1.8 GB

As shown in Table 12, the full pipeline requires approximately 1821 ms per frame, corresponding to a throughput of roughly 0.55 FPS, which confirms that real-time processing is not achievable with the current metaheuristic-based segmentation. The computational bottleneck resides almost entirely in Stage 1, where the iterative ADE-GJO optimization accounts for over 95% of the total latency. In contrast, when the foreground masks are precomputed offline, the combined latency of Stage 2 and Stage 3 is reduced to 68 ms per frame (approximately 14.7 FPS), which is sufficient for near-real-time feedback in interactive training applications. The peak memory column in the same table shows that the ADE-GJO optimizer requires only 4.6 MB on the CPU side, and the morphological buffers add 2.1 MB per frame, making Stage 1 portable to any standard desktop without memory constraints. The GPU-resident modules of Stage 2 collectively occupy

approximately 1.8 GB, which is well within the capacity of consumer-grade GPUs such as the NVIDIA GTX 1650 with 4 GB of onboard memory.

4. Discussions

Regarding multi-threshold image segmentation, the ADE-GJO algorithm yields a PSNR value of 26.14 when the number of thresholds is high, indicating an improvement of 0.58 dB compared with the basic GJO algorithm. In contrast, the improvement between the two algorithms with respect to PSNR is only 0.09 dB when the number of thresholds is small, indicating that the advantages of the three improvements are evident with increased dimensionality. As compared to earlier studies which have attempted to address the issue of incorporating multi-threshold-based segmentation and action recognition [23], this study specifies the relationship between the quality of segmentation and the accuracy of recognition by integrating the output of segmentation in the spatial attention model of recognition, whereas both are regarded as separate processes in previous studies. Regarding the development of the segmentation algorithm, while the threshold optimization technique employs the use of a combination of differential evolution and golden jackal optimization method [24], the adaptive scheme suggested in this study controls the time of mutations based on diversity checks to avoid interference with the normal search procedure that may arise due to iterations of perturbations.

In terms of dance action recognition, SG-STANet achieves an accuracy rate of 91.26%. This accuracy rate decreases by 85.83% when the influence of segmentation guidance is stripped off, which creates a margin of 5.43% points. Such a margin clearly suggests that mask guidance is what improves the performance. Different from the skeleton recognition scheme that depends on correcting pose and improving channel topology [25] for better topological modeling within the network, this study provides a prior knowledge of image segmentation at the image level from outside. The two methods then make contributions to better skeleton feature extraction separately. With respect to motion optimization, contrary to the techniques for dance quality evaluation that use scoring algorithms [26] as well as the evaluation methods based on spatio-temporal convolutional networks [27], DTW deviation measures are validated via the internal complexity level annotations provided by the dataset. This ensures that the separation of deviations is consistent across basic and expert classes without the need for human quality labeling.

From a practical standpoint, the proposed framework addresses several concrete needs in dance education and intelligent training. In university-level and vocational dance curricula, instructors typically evaluate student performances through subjective observation, which is time-consuming and prone to inter-rater inconsistency. The per-joint deviation profile generated by the DTW module can serve as an objective supplementary metric that highlights specific joints requiring improvement, enabling targeted corrective feedback. In commercial dance-training applications, the style classification module can automatically categorize user-uploaded videos and route them to the corresponding reference template library, reducing the manual workload of content organization. In physical rehabilitation contexts where patients perform prescribed dance-based exercises, the deviation score D_{total} can be used to track recovery progress over successive sessions by quantifying the convergence of patient movements toward the normative template. The mask-based attention mechanism also benefits scenarios in which dancers perform against visually complex backgrounds (outdoor stages, projection-mapped environments), as it suppresses background interference at the feature level without requiring controlled studio conditions.

Regarding deployment on low-cost hardware, the modular architecture of the proposed pipeline offers flexibility in balancing accuracy against computational budget. As shown in the computational efficiency analysis of Section 3.5, the computationally dominant Stage 1 can be executed offline on a standard desktop CPU to build a mask library for each choreography video, after which Stages 2 and 3 require only 68 ms per frame on a consumer-grade GPU and can thus run on mid-range hardware such as an NVIDIA GTX 1650 or an integrated laptop GPU without exceeding interactive latency thresholds. For edge-device deployment (e.g., tablets used in dance studios), the ADE-GJO step can be replaced by a lightweight single-threshold Otsu binarization at the cost of a moderate accuracy reduction (from 91.26% to 88.15%, as shown by the input ablation

in Section 3.3), providing a configurable trade-off between recognition quality and hardware cost. The DTW deviation module is inherently lightweight and imposes negligible additional computational requirements. These characteristics suggest that, with appropriate architectural choices, the framework can be adapted to resource-constrained environments without fundamentally compromising its analytical utility.

Several limitations should be recognized when evaluating this method. First, as quantified in Section 3.5, the iterative ADE-GJO optimization dominates the per-frame latency of the full pipeline, restricting throughput to below 1 FPS and precluding real-time deployment, although a two-phase strategy that precomputes masks offline can bring the online inference within interactive latency bounds. Second, the generalization capability of the templates is constrained to only 10 dance styles used in the AIST++ dataset. Finally, the deviation metric currently considers only joint angles and excludes high-level semantics about the dance.

5. Conclusions

This research proposes an ADE-GJO algorithm for multi-threshold segmentation with the integration of fused Otsu-Kapur dual objective fitness function, SG-STANet for spatio-temporal attention recognition network, and DTW-based movement deviation calculation. An integrated end-to-end framework for dance movement is presented which includes image segmentation, action recognition, and movement optimization. Based on the experimental results, ADE-GJO achieves a PSNR of 26.14 under the condition of n=10, SG-STANet attains a dance style recognition accuracy of 91.26% on AIST++, and the DTW deviation measure provides efficient differentiation of dances of different quality within all 10 dance genres.

The above findings indicate that transferring segmentation quality of multi-thresholds to recognition performance is possible, providing a technical route to achieving automated measurements of deviation in movements without resorting to manual scores. This provides a technical possibility of the utilization of computer vision in dance education assessment and intelligent training. Future works should focus on lightweight model design to meet real-time processing constraints, incorporation of the beat information from the audio track to make synchronization analysis between music and dancing possible, generalization to cases where collaboration of more dancers is needed, and creating larger scale public datasets including various dance genres.

References

[1] Houssein EH, Mohamed GM, Ibrahim IA, Wazery YM. An efficient multilevel image thresholding method based on improved heap-based optimizer. Scientific Reports. 2023;13(1):9094. https://doi.org/10.1038/s41598-023-36066-8

[2] Hu P, Han Y, Zhang Z, Chu SC, Pan JS. A multi-level thresholding image segmentation algorithm based on equilibrium optimizer. Scientific Reports. 2024;14(1):29728. https://doi.org/10.1038/s41598-024-81075-w

[3] Li H, Zhu X, Li M, Yang Z, Wen M. Multi-threshold image segmentation research based on improved enhanced arithmetic optimization algorithm. Signal, Image and Video Processing. 2024;18(5):4045-4058. https://doi.org/10.1007/s11760-024-03026-2

[4] Chakraborty F, Roy PK. An efficient multilevel thresholding image segmentation through improved elephant herding optimization. Evolutionary Intelligence. 2025;18(1):17. https://doi.org/10.1007/s12065-024-01001-x

[5] Liu J, Yang S, Liu W, Ni F, Chen Y. Multilevel threshold image segmentation based on a novel mechanism enhanced coati optimization algorithm. Scientific Reports. 2026. https://doi.org/10.1038/s41598-026-40921-9

[6] Wang C, Yan J. A comprehensive survey of rgb-based and skeleton-based human action recognition. IEEE Access. 2023;11:53880-53898. https://doi.org/10.1109/ACCESS.2023.3282311

[7] Nguyen HC, Nguyen TH, Scherer R, Le VH. Deep learning for human activity recognition on 3D human skeleton: survey and comparative study. Sensors. 2023;23(11):5121. https://doi.org/10.3390/s23115121

[8] Gao Z, Chen J, Liu Y, Jin Y, Tian D. A systematic survey on human pose estimation: upstream and downstream tasks, approaches, lightweight models, and prospects. Artificial Intelligence Review. 2025;58(3):68. https://doi.org/10.1007/s10462-024-11060-2
[9] Zheng Z, Yuan Q, Zhang H, Wang Y, Wang J. Lightweight multiscale spatio-temporal graph convolutional network for skeleton-based action recognition. Big Data Mining and Analytics. 2025;8(2):310-325. https://doi.org/10.26599/BDMA.2024.9020095
[10] Jin N, Wen L, Xie K. The Fusion Application of Deep Learning Biological Image Visualization Technology and Human-Computer Interaction Intelligent Robot in Dance Movements. Computational Intelligence and Neuroscience. 2022;2022(1):2538896. https://doi.org/10.1155/2022/2538896
[11] Liu X, Ko YC. The use of deep learning technology in dance movement generation. Frontiers in Neurorobotics. 2022;16:911469. https://doi.org/10.3389/fnbot.2022.911469
[12] Sun Y, Li Y. A deep learning method for intelligent analysis of sports training postures. Computational Intelligence and Neuroscience. 2022;2022(1):2442606. https://doi.org/10.1155/2022/2442606
[13] Li N, Boers S. Human motion recognition in dance video images based on attitude estimation. Wireless Communications and Mobile Computing. 2023;2023(1):4687465. https://doi.org/10.1155/2023/4687465
[14] Pang Y, Niu Y. Dance video motion recognition based on computer vision and image processing. Applied Artificial Intelligence. 2023;37(1):2226962. https://doi.org/10.1080/08839514.2023.2226962
[15] Zhao F. T2W-CogLoadNet: a framework for cognitive load assessment of dance movements based on deep learning-powered human pose estimation. Frontiers in Psychology. 2025;16:1707539. https://doi.org/10.3389/fpsyg.2025.1707539
[16] Yin P, Li X. A deep reinforcement learning approach to dance movement analysis. Scientific Reports. 2026. https://doi.org/10.1038/s41598-026-35311-0
[17] Zheng J, Gao Y, Zhang H, Lei Y, Zhang J. OTSU multi-threshold image segmentation based on improved particle swarm algorithm. Applied Sciences. 2022;12(22):11514. https://doi.org/10.3390/app122211514
[18] Zhou B, et al. A novel segmentation scheme with multi-probability threshold for human activity recognition using wearable sensors. Sensors. 2022;22(19):7446. https://doi.org/10.3390/s22197446
[19] Liu D, Xu H, Wang J, Lu Y, Kong J, Qi M. Adaptive attention memory graph convolutional networks for skeleton-based action recognition. Sensors. 2021;21(20):6761. https://doi.org/10.3390/s21206761
[20] Arbelaez P, Maire M, Fowlkes C, Malik J. Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010;33(5):898-916. https://doi.org/10.1109/TPAMI.2010.161
[21] Li R, Yang S, Ross DA, Kanazawa A. Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 13401-13412. https://doi.org/10.1109/ICCV48922.2021.01315
[22] Liu J, Shahroudy A, Perez M, Wang G, Duan LY, Kot AC. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019;42(10):2684-2701. https://doi.org/10.1109/TPAMI.2019.2916873
[23] Han Y, Mojtahe K. Intelligent Action Recognition and Dance Motion Optimization Based on Multi-Threshold Image Segmentation. Mobile Information Systems. 2022;2022(1):5776642. https://doi.org/10.1155/2022/5776642
[24] Meng X, Tan L, Wang Y. An efficient hybrid differential evolution-golden jackal optimization algorithm for multilevel thresholding image segmentation. PeerJ Computer Science. 2024;10:e2121. https://doi.org/10.7717/peerj-cs.2121
[25] Gao Y, Duan X, Dai Q. Skeleton-Based Action Recognition Using Graph Convolutional Network with Pose Correction and Channel Topology Refinement. Computers, Materials & Continua. 2025;83(1). https://doi.org/10.32604/cmc.2025.060137
[26] Qu J. A dance movement quality evaluation model using transformer encoder and convolutional neural network. Scientific Reports. 2024;14(1):32058. https://doi.org/10.1038/s41598-024-83608-9
[27] Qin W, Meng J. The research on dance motion quality evaluation based on spatiotemporal convolutional neural networks. Alexandria Engineering Journal. 2025;114:46-54. https://doi.org/10.1016/j.aej.2024.11.025