DAUfomer: A Deep Adaptive Uncertainty-Driven Transformer For Action Recognition

Yingcun Wang[1], Rong Wang[2*], Shizhong Liu[3], and Zhi Zeng[4]

[1] School of Information Engineering, Weifang Vocational College, Weifang 261031, China
[2] School of International Business, Weifang Vocational College, Weifang 262737, China
[3] Weifang Vocational College, Weifang 261031, China
[4] Shandong Qiaotong Tianxia Network Technology Co., Ltd, Weifang, 261061, China
[*] Corresponding author. E-mail: wangrongwvc@163.com

Received: May 02, 2025; Accepted: Jul. 13, 2025

Current action recognition research excessively relies on deterministic feature correlation mechanisms, struggling to address core challenges including spatiotemporal heterogeneity, action category ambiguity, and cross-frame semantic discontinuity prevalent in video stream data. To this end, this study proposes the Deep Robust Recognition Transformer (DAUfomer), to reconstruct action recognition paradigms through three synergistic modules. Multi-granularity feature extraction module employs Transformer with dual attentions to extract low-dimensional and high-information-density spatiotemporal features from high-dimensional video streams, preserving local motion details while establishing global contextual correlations. Uncertainty-driven spatial-temporal aggregation module innovatively constructs a hybrid Gaussian-Dirichlet distribution model, transforming deterministic spatiotemporal attention into a probabilistic learnable Bayesian network. This enables dynamic adaptation to data distribution shifts through latent space uncertainty quantification. Proactive semantic enhancement architecture breaks traditional causal constraints in temporal modeling by designing a bidirectional temporal distillation mechanism. It leverages latent semantic cues from future frames to construct cross-frame attention correlation graphs, enhancing current action features via gated recurrent unit-based spatiotemporal context refinement. Finally, extensive results on two real world datasets, especially on the MS-DanceAction dataset with a 4.48% ACC improvement compared to the second-best result, verify that DAUfomer conducts a new standard baseline in the action recognition task.

Keywords: Multimedia action recognition; probability-driven aggregation; proactive semantic enhancement

http://dx.doi.org/10.6180/jase.202604_29(4).0009

1. Introduction

Action recognition has become a cornerstone task in computer vision, driving advancements across diverse domains such as security, human-computer interaction, and healthcare [1, 2]. Despite its widespread applicability, existing models often grapple with critical limitations, particularly in handling the inherent complexity of spatiotemporal data. Video streams exhibit challenges like spatiotemporal heterogeneity, ambiguous action categories, and semantic discontinuities across frames. These issues are exacerbated by the reliance of current methods on deterministic feature correlations, which restrict their adaptability to dynamic environments and varying data distributions [3–6].

Traditional action recognition approaches can be categorized into CNN-based and Transformer-based methods. CNN-based architectures [7–9] are adept at extracting local spatial features through hierarchical convolutional layers, offering computational efficiency and structural simplicity. However, they struggle to model long-range temporal de-

pendencies and global contextual relationships effectively. For example, the local receptive fields of CNNs limit their ability to capture interactions between distant frames in a video sequence, which is crucial for understanding actions that evolve over time. Transformer-based frameworks [10–12], on the other hand, utilize self-attention mechanisms to capture intricate inter-frame dependencies, achieving superior performance in modeling sequential data. Yet, their computational overhead and sensitivity to noisy or redundant features pose significant challenges, especially in resource-constrained scenarios. The self-attention mechanisms in Transformers require quadratic computational complexity with respect to the sequence length, making them less efficient for long video sequences. Additionally, Transformers are more susceptible to noisy or redundant features in the input data, which can degrade their performance.

A common challenge faced by both CNN-based and Transformer-based approaches is the issue of spatiotemporal redundancy. Irrelevant features in the data can introduce noise, which not only degrades recognition accuracy but also complicates the feature selection process. This often leads to overfitting and suboptimal generalization in real-world settings where data variability and complexity are prevalent. For instance, in scenes with cluttered backgrounds or abrupt camera movements, both types of models may struggle to distinguish relevant action features from noise, resulting in decreased recognition performance.

To address these limitations, we propose the Deep Adaptive Uncertainty-Driven Transformer (DAUformer), a novel architecture that redefines action recognition through three synergistic innovations. First, the multi-granularity feature extraction module leverages dual-attention Transformers to distill low-dimensional and information-rich spatiotemporal features from high-dimensional inputs, effectively preserving both fine-grained motion details and global contextual patterns. Building on this, uncertainty-driven spatial-temporal aggregation module replaces deterministic attention mechanisms by modeling attention masks as a hybrid Gaussian-Dirichlet distribution, transforming them into a probabilistic Bayesian network. This innovation enables latent space uncertainty quantification, significantly enhancing robustness to data scarcity and distribution shifts. Meanwhile, the proactive semantic enhancement component introduces a bidirectional temporal distillation mechanism that breaks traditional causal constraints, integrating latent semantic cues from future frames to refine current action representations through spatiotemporally gated context fusion. Together, these components form a cohesive framework that addresses spatiotemporal redundancy, adapts to dynamic environments, and leverages cross-frame dependencies to achieve robust and precise action recognition. Finally, Extensive experiments on benchmark datasets validate DAUEformer's superiority, achieving state-of-the-art accuracy of 70.24% on InDanceAction and 77.97% on MSDanceAction, outperforming prior methods by 4.38% and 4.69%, respectively.

The key contributions of DAUformer include:

- A probabilistic spatiotemporal aggregation framework that replaces deterministic attention with uncertainty-aware Bayesian learning, enhancing adaptability to dynamic environments.
- A proactive semantic enhancement strategy that transcends traditional causal modeling, is proposed via leveraging cross-frame dependencies to refine action context.
- Empirical validation of a new performance benchmark in action recognition, particularly in scenarios requiring robustness to data variability and complex motion patterns.

DAUformer is organized as follows: Section 2 shows DAUformer. Section 3 provides a thorough evaluation using benchmark datasets. Section 4 summarizes the key contributions.

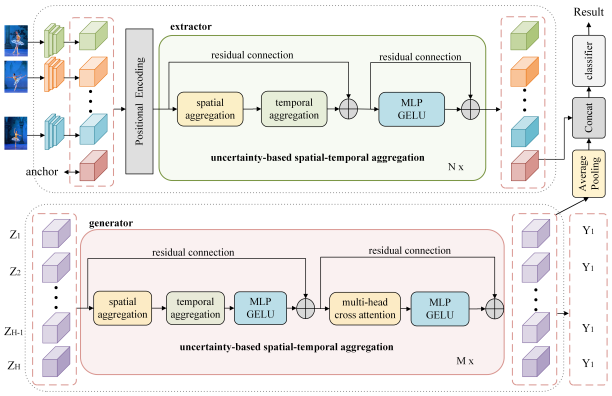

Fig. 1. The overall framework of DAUfomer, containing multi-granularity feature extraction, uncertainty-driven spatial-temporal aggregation, proactive semantic enhancement

2. Method

In this section, the problem description of action recognition is firstly defined, and then the detailed architecture of DAUfomer is introduced, containing multi-granularity feature extraction, probability-driven spatiotemporal aggregation, proactive semantics enhancement.

2.1. Problem Description

The goal of DAUEformer is to identify the currently occurring action by analyzing video frames from a historical period, where the video stream may contain multiple actions. Let $V = \{f_t\}_{t=-H}^{0}$, f_0, and $y_i$denote the video stream of the historical period H, the current action frame, the category of f_0, respectively. The problem description of real-time action recognition can be defined:

$$y_i = \text{argmax}_i(P(y_i|\phi(V))) \tag{1}$$

Here, ϕ is the mapping neural network that converts the feature vector V of historical period into scores for each action category. By taking the maximum of the output scores, the action category y_i for the current frame block can be determined.

2.2. Multi-Granularity Feature Extraction

Dual-level information extraction strategy first processes the input video stream $V = \{f_t\}_{t=-H}^{0}$ by using a video feature extractor, to generate a one-dimensional feature sequence. The purpose of this process is to transform high-dimensional raw data into lower-dimensional features with higher information density, enabling more efficient subsequent model training and action recognition. Next, an additional linear projection layer maps each vectorized frame feature into a D-dimensional representation space, generating the resulting anchor sequence expressed as

$$F = \{anchor_t\}_{t=-H}^{0} \in \mathbb{R}^{(T+1)\times D} \tag{2}$$

Meanwhile, the dual-level information extraction strategy extends a learnable category anchor $anchor_c \in \mathbb{R}^D$ to representation sequence F, resulting in the anchor representation sequence:

$$\tilde{F} = Concat(\{anchor_t\}_{t=-H}^{0}, anchor_c) \in \mathbb{R}^{(H+2)\times D} \tag{3}$$

where $Concat()$ denotes concatenation function.

$anchor_t$ aims to capture local information that refers to the details of each frame feature, which provides specific data about the instantaneous changes in movements. $anchor_t$ focuses on the features of each frame, helping the model understand the details of actions at each moment. This information includes posture, movement speed, and the positions of hands and feet, allowing the model to make precise judgments about the state of each moment. However, relying solely on local information may not fully capture the overall characteristics of the movements, leading to an insufficient understanding of the actions, thus necessitating the introduction of global information to address this shortcoming. The global information is realized through $anchor_c$, which is used to learn global distinguishing features related to the action detection task. Global information is about understanding the entire video stream, providing contextual and background information that enables the model to determine the continuity and relationships between actions. This combination allows the model to not only depend on local features during action recognition but also integrate the contextual information of the entire video stream. For example, in recognizing movements, global information helps understand the overall rhythm and style of the actions. Without $anchor_c$, the final representation obtained from other anchors inevitably lean towards the overall specified anchors, making it challenging for the model to effectively represent the learning task. In contrast, the semantic embedding of $anchor_c$ enhances the model capability for action detection through adaptive interaction with other anchors in the extractor, resulting in a more flexible and comprehensive feature representation.

Additionally, due to the lack of frame order information in the extractor, the multi-granularity feature extraction strategy also requires embedding positional encoding. Positional encoding can take the form of sinusoidal inputs or learnable embeddings to capture the relative positions of frames in the time series. The introduction of positional encoding is crucial as it allows the model to comprehend the importance and relative positions of each frame in the time series, which is vital for understanding the temporal changes in movements. Positional encoding is added to the anchor sequence through element-wise addition to retain positional information:

$$Z = \tilde{F} + E_{pos}, E_{pos} \in \mathbb{R}^{(T+2)\times D} \tag{4}$$

The inclusion of this positional encoding is essential as it enables the model to understand the temporal relationships between different frames, allowing for effective capture of the temporal variations and continuity of actions when recognizing dynamic movements. For instance, the model can recognize the start and end of movements at specific time points, ensuring the accuracy of action recognition.

To enhance the spatiotemporal information interaction among various anchors and capture long-distance semantic dependencies, the uncertainty-driven spatial-temporal aggregation is designed, whose the specific detail is presented in Section 3.3.

$$\bar{Z} = TA \times (SA \times Z) \tag{5}$$

where TA and SA denote spatial and temporal aggregation functions, respectively.

The information is further processed through the multilayer perception with a GELU activation function with the

layer normalization and residual connections to facilitate information flow and prevent gradient vanishing:

$$\hat{Z} = \bar{Z} + MLP(\text{Norm}(\bar{Z}), GELU) \tag{6}$$

Then, the spatial-temporal aggregation function and the multi-layer perception function are encapsulated into a block mechanism for iterative information extraction.

$$\begin{aligned} \bar{Z}_i &= TA \times (SA \times Z_i) \\ \hat{Z}_i &= \bar{Z}_i + MLP(\text{Norm}(\bar{Z}_i), GELU) \end{aligned} \tag{7}$$

where $i = 1, 2, ..., N$ denote the number of blocks. $\hat{Z}_N$ denotes final output of the multi-granularity feature extraction.

2.3. Uncertainty-driven spatial-temporal aggregation

The attention strategy in Transformers typically employs a deterministic information aggregation method, using fixed weighted calculations between features to capture key dependencies. However, this deterministic approach can limit the model's generalization ability when faced with data scarcity or distribution shifts (e.g., changes in styles or environmental factors), leading to performance degradation. In applications or dynamic environments, traditional attention mechanisms lack the capability to model uncertainty in the data, making it difficult to adapt to such variations. To address this issue, an uncertainty-driven spatial-temporal aggregation method is proposed, which introduces uncertainty modeling, enabling the model to adaptively adjust to different spatial-temporal feature regions, thereby enhancing its robustness and generalization in diverse scenarios.

2.3.1. Uncertainty-aware spatial attention

The goal of spatial aggregation is to identify correlation information among samples via applying a spatial attention mask to representations of each frame with a Hadamard product. Traditional attention mechanisms are deterministic and often ignore the inherent randomness in the data. Therefore, an uncertainty-aware spatial aggregation function is designed via modelling masks as the a prior distribution, e.g. the Gaussian distribution:

$$SA_{(i,j)} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}) \tag{8}$$

In the optimization, the forward pass of DAUformer adopts the reparameterization trick to capture the spatial randomness of representations and further quantify the uncertainty of predictions:

$$SA_{(i,j)} = \mu_{ij} + \epsilon\sigma_{ij}, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{9}$$

It is important to note that, apart from the spatial attention mask, all other parameters of DAUformer remain deterministic and fixed during inference. In the inference phase, multiple outputs are generated by sampling from the learned distribution of spatial attention masks. Specifically, the representation Z_i is propagated via producing different masks, i.e., $\{SA^1_{i,j}, \ldots, SA^K_{i,j}\}$. The uncertainty of the representation Z_i is then estimated from these samples:

$$\begin{aligned} I_{SA^k_{i,j}}(i,j) =& H\left[\frac{1}{K}\sum_{k=1}^{K} P\left(y \mid z, SA^k_{i,j}(i,j)\right)\right] \\ &- \frac{1}{K}\sum_{k=1}^{K} H\left[y \mid z, SA^k_{i,j}(i,j)\right] \end{aligned} \tag{10}$$

where $H[]$ denotes the entropy function. Then, the spatial attention mask is updated via weighting learned uncertainty:

$$SA_{(i,j)} = 1 + \frac{U_{SA}(i,j)}{\sum_{i,j} U_{SA}(i,j)} \tag{11}$$

where $U_{SA}(i,j)$ is defined as follows:

$$U_{SA}(i,j) = E_{p(\phi_p \mid Z_i, \phi^*_d)}\left[H\left[y \mid Z_i, \phi_p\right]\right] \tag{12}$$

where ϕ_p and ϕ_d stands for deterministic and probabilistic parameters of DAUformer. The Eq. (9) shows that regions with higher uncertainty are assigned larger weights to guide representation learning. The final representation is updated via the spatial aggregation as:

$$Z^{SA}_i = U_{SA}(i,j) \odot Z_i \tag{13}$$

This uncertainty-aware spatial aggregation method enables the model to capture the inherent spatial randomness of input features while identifying critical discriminative regions more effectively.

2.3.2. Uncertainty-aware temporal attention

The goal of temporal aggregation is to identify important frames from the video stream V, via applying a temporal attention mask to representations of various frames with a Hadamard product. Similar to the spatial aggregation function, temporal aggregation aims to model the temporal attention mask probabilistically, ensuring that all aggregated information is weighted and used for current action recognition. This modeling approach allows for dynamic adjustments of the temporal attention weights, making the model more adaptable in different situations. Specifically, the parameter ϕ is assumed to follow a Gaussian distribution with a mean μ and a covariance matrix Σ, expressed as: $\phi = \mu + \epsilon\Sigma, \quad \epsilon \sim \mathcal{N}(0, I)$. ϕ can be predicted via modelling the mean and covariance to achieve quantifying uncertainty. This enables the model to make more informed decisions based on historical information when faced with

complex data. Next, K samples are drawn from the input, and the corresponding cognitive uncertainty is estimated:

$$I_{TA^k} = H\left[\frac{1}{K}\sum_{k=1}^{K} p(y^k)\right] - \frac{1}{K}\sum_{k=1}^{K} H\left(p(y^k)\right) \quad (14)$$

By evaluating cognitive uncertainty, important frames can be effectively distinguished, as these frames are crucial for the final recognition result. The temporal attention weights are positively correlated with cognitive uncertainty, indicating that higher uncertainty reflects greater importance of the frame. Temporal attention is calculated by importance weighting of all frames as follows:

$$TA = 1 + \frac{U_{TA}}{\sum_{t=-H}^{0} U_{t;TA}} \quad (15)$$

where H is the history frame number. This mechanism not only improves the efficiency of the model but also enhances its adaptability to new scenarios. Unlike existing attention aggregation functions, the uncertainty-aware attention mechanism is based on predictive uncertainty, making the generation process more interpretable and allowing users to understand the model's decision-making basis in specific situations. This approach better addresses changes that may arise in complex environments, improving the accuracy and reliability of action recognition.

2.3.3. Spatio-temporal joint information aggregation

To jointly optimize importance weighting of spatial and temporal feature, they are combined to form a unified spatio-temporal joint information aggregation:

$$\bar{Z}_i = TA \times (SA \times Z_i) \quad (16)$$

This combination allows spatial and temporal features to be processed within the same framework, thereby enhancing the model's expressive capability. The advantage of this approach lies in its ability to fully leverage the information contained in the same input data, thereby improving the accuracy and consistency of the estimates. Thus, spatial and temporal importance weighting is obtained via using learned uncertainty in a same manner. This efficiency not only reduces the consumption of computational resources but also enhances the model's responsiveness in applications, enabling it to adapt more flexibly to changes in dynamic environments.

2.4. Proactive Semantic Enhancement

Generally, during training, the model has access to future video frames, and the semantic information contained in these frames can be combined with past observations to help the model anticipate upcoming actions. This prediction mechanism effectively guides the model in learning discriminative features, thereby improving its accuracy and efficiency in action recognition and differentiation.

Specifically, the learnable future semantic encoding is defined as $Z_t, t = 1, 2, ..., H$, and H future frames are generated in parallel using the generator, whose forward computation is as follows:

$$\bar{Z}_t = TA \times (SA \times Z_t) \quad (17)$$

$$\hat{Z}_t = \bar{Z}_t + MLP(\mathrm{Norm}(\bar{Z}_t), GELU) \quad (18)$$

$$\tilde{Z}_t = MCA(\hat{Z}_t, \hat{Z}_0) \quad (19)$$

$$\tilde{Z}_t = \tilde{Z}_t + MLP(\mathrm{Norm}(\tilde{Z}_t), GELU) \quad (20)$$

where MCA denotes multi-head cross attention. $\tilde{Z}_t$ denotes the output of the generator.

For the recognition of actions in the current frame, the representations relevant to the task from the extractor are first concatenated with the pooled representations from the generator to fully leverage information from different sources, and then the probability distribution of the current action is output through the classification head:

$$y_0 = softmax\left(MLP_c\left(Concat\left[\hat{Z}_N, Avg-pool\left(\tilde{Z}_1, \tilde{Z}_2, \ldots, \tilde{Z}_H\right)\right]\right)\right) \quad (21)$$

where MLP_c represents the parameters of the classification head, while y_0 is the estimated probability of the current action.

By performing average pooling on the relevant representations, representative features can be extracted, reducing the influence of noise and focusing on the primary information. Next, the probability distribution of the current action is output through the classification head:

In addition to estimating the current action, DAUformer also outputs predictions for the actions in the next H time frames. This design considers the temporal characteristics of actions, enabling the model to make decisions over a larger temporal window. Since future information can be utilized during training, it is essential to perform supervised training on the future predictions to ensure the model learns good representations:

$$y_t = softmax((MLP_f \tilde{Z}_t))), \quad t = 1, 2, \ldots, H \quad (22)$$

where MLP_f represents the parameters of the classification head of future frames. By supervising the future frames, DAUformer can better learn how to integrate past information with future predictions, thereby improving its generalization ability.

The final joint training loss is defined as:

$$L = -\sum_{c=1}^{C} p_0^c \log(y_0) + \lambda \sum_{t=1}^{H}\left(-\sum_{c=1}^{C} p_t^c \log(y_i^c)\right) \quad (23)$$

Table 1. Comparisons of the Top-1 and Top-5 accuracy metrics with current recognition methods on mobile devices

Method	MSDanceAction		InDanceAction	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
MnasNet	56.78	76.98	44.09	66.23
NASNet	57.12	77.12	49.09	68.56
PNASNet	60.78	80.88	55.70	72.19
Hyperformer	67.85	76.59	60.40	79.86
HD-GCN	69.92	82.90	61.00	80.10
DAR-DCNN	73.59	84.08	65.55	84.74
DAUformer	77.97	87.50	70.24	88.65

This loss function not only considers the predictions of the current action but also supervises the predictions of future actions, enhancing the model learning capability. The parameter λ is a balancing parameter to control the impact of current and future prediction losses within the overall loss. where p_0 and p_t denotes labels of the current and future frames, respectively.

3. Results and discussion

3.1. Dataset and Setup

Dataset and metric: Two action recognition datasets are used to evaluate the performance of DAUformer. MSDanceAction dataset contains 3100 action videos where each action has 50 videos. It presents a wide array of action videos, marked by its complex backgrounds and diverse characters, showcasing both its variety and intricacy. InDanceAction dataset contains 2232 action videos where each action has 62 videos. To ensure result robustness and generalization, a random split is employed, with 70% of the data allocated for training and 30% for testing. Following Zhu and Zhu [12], recognition performance is measured using Top-1 and Top-5 accuracy metrics. DAUformer is trained on the training subset, and is evaluated on the testing portion.

Implementation Details: For model training, DAUformer is implemented in PyTorch, and all experiments are conducted on Nvidia V100 GPUs to ensure efficient computation. The Adam optimizer is employed for optimization, with the batch size set to 128 to balance memory usage and convergence speed. The learning rate is fixed at 0.0001 to provide gradual updates, while weight decay is set to 0.0005 to prevent overfitting by adding regularization. Throughout the process, the network is trained in a stable environment, ensuring consistent evaluation across all experimental runs. To perform video inference on mobile devices, a Huawei Mate 40 phone is used, which is equipped with a Kirin 970 CPU and 6GB of RAM. It is important to note that the neural network chip in the Kirin is not utilized for computational acceleration. For mobile platforms, TFLite and related projects are employed. λ is set to 0.5.

3.2. Comparison analysis

Comparison method: MnasNetcite[7], NASNet[8], PNASNet[9], Hyperformer [10], HD-GCN[11], and DAR-DCNN [12].

Comparison results: Based on the results shown in Table 1, DAUformer outperforms other comparison methods on both the MSDanceAction and InDanceAction datasets. For instance, on the MSDanceAction dataset, DAUformer achieved a Top-1 accuracy of 77.97%, which is 4.38 percentage points higher than the second-ranked DAR-DCNN (73.59%); on the InDanceAction dataset, DAUformer's Top-1 accuracy is 70.24%, 4.69 percentage points higher than DAR-DCNN (65.55%). Similarly, in terms of Top-5 accuracy, DAUformer reached 87.50% on the MSDanceAction dataset and 88.65% on the InDanceAction dataset, both higher than other methods. This is due to its three core strengths. Firstly, the multi-granularity feature extraction strategy adopted by DAUformer effectively extracts low-dimensional features with high information density from high-dimensional raw data, while combining local and global information allows the model to fully understand the details and context of movements. Secondly, the uncertainty-driven spatial-temporal aggregation method models spatial and temporal attention masks using a Gaussian distribution, enabling the model to adapt automatically, enhancing robustness and generalization in situations with scarce data or distribution changes. Lastly, proactive semantic enhancement uses semantic information from future frames to predict upcoming actions, improving the recognition accuracy of current actions and enhancing the model's generalization ability through the supervision of future frames. The combination of these strategies, along with the potential use of advanced network architectures and optimization techniques, as well as targeted model training and tuning, makes DAUformer perform excellently in the task of action recognition on mobile devices.

Table 2. Ablation Analysis of DAUformer

Method	MSDanceAction		InDanceAction	
	Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
DAUformer_1	66.12	78.98	50.98	76.24
DAUformer_2	74.74	83.63	65.19	85.58
DAUformer_3	73.87	82.47	66.19	85.99
DAUformer	77.97	87.50	70.24	88.65

3.3. Ablation analysis

Following previous works [13–15], we designed three ablation experiments to evaluate the effectiveness of each module in DAUformer. DAUformer_1 only conducts multi-granularity feature extraction to mine action patterns. DAUformer_2 combines multi-granularity feature extraction with uncertainty-driven spatial-temporal aggregation. DAUformer_3 combines multi-granularity feature extraction with proactive semantic enhancement.

The experimental results are shown in Table 2. Three key observations can be made from the results: 1. DAUformer_1 achieves a Top-1 accuracy of 66.12% on MSDanceAction and 50.98% on InDanceAction. This indicates that while multi-granularity feature extraction is beneficial, it is not sufficient on its own to achieve optimal performance. 2. DAUformer_2 shows a significant improvement over DAUformer_1, with a Top-1 accuracy of 74.74% on MSDanceAction and 65.19% on InDanceAction. This demonstrates that integrating uncertainty-driven spatial-temporal aggregation enhances the model's ability to capture complex action patterns. 3. DAUformer_3 slightly outperforms DAUformer_2, achieving a Top-1 accuracy of 73.87% on MSDanceAction and 66.19% on InDanceAction. However, the complete DAUformer model surpasses DAUformer_3, indicating that the integration of all components is crucial for optimal performance. The ablation study shows that each component of DAUformer contributes to its overall performance. The multi-granularity feature extraction provides a solid foundation. The uncertainty-driven spatial-temporal aggregation significantly improves the model's ability to handle complex and dynamic actions. The proactive semantic enhancement further refines the model's predictive accuracy. The complete model achieves the highest accuracy rates, highlighting the importance of a holistic approach in developing effective action recognition systems.

4. Conclusion

In this study, we address the critical challenges of spatiotemporal redundancy, dynamic environment adaptability, and semantic discontinuity in action recognition by proposing the DAUformer, a novel uncertainty-driven Transformer framework. Our approach integrates three core innovations to redefine feature learning and spatiotemporal modeling in video analysis. Extensive empirical validation demonstrates DAUformer's superiority over existing methods, achieving state-of-the-art accuracy. The integration of uncertainty quantification further enhances its practicality for real-world scenarios characterized by data scarcity or distribution shifts, such as healthcare monitoring or security surveillance. Future work will focus on two main directions. First, we will explore the integration of lightweight neural operators to optimize computational efficiency without compromising recognition accuracy. This could involve replacing certain components of the model with more efficient alternatives, such as using approximate Bayesian methods or distilling the uncertainty-aware model into a lighter architecture. Second, we will investigate automated hyperparameter tuning methods to enhance the model's adaptability to various scenarios.

References

[1] Y. Zhong, L. Chen, C. Dan, and A. Rezaeipanah, (2022) *"A systematic survey of data mining and big data analysis in internet of things"* **The Journal of Supercomputing** ***78***(17): 18405–18453.

[2] P. Li, J. Gao, J. Zhang, S. Jin, and Z. Chen, (2022) *"Deep Reinforcement Clustering"* **IEEE Transactions on Multimedia**: DOI: 10.1109/TMM.2022.3233249.

[3] M. Korban, P. Youngs, and S. T. Acton, (2023) *"A multi-modal transformer network for action detection"* **Pattern Recognition** ***142***: 109713.

[4] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2024) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems** ***35***(9): 12965–12978. DOI: 10.1109/TNNLS.2023.3265699.

[5] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, (2023) *"Deep Incomplete Multiview Clustering via Information Bottleneck for Pattern Mining of Data in Extreme-Environment IoT"* **IEEE**

Internet of Things Journal *11*(16): 26700–26712. DOI: 10.1109/JIOT.2023.3325272.

[6] Z. Lin. "Dance movement recognition method based on convolutional neural network". In: *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*. 2023, 255–258.

[7] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le. "Mnasnet: Platform-aware neural architecture search for mobile". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, 2820–2828.

[8] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le. "Learning transferable architectures for scalable image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, 8697–8710.

[9] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy. "Progressive neural architecture search". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, 19–34.

[10] J. Lee, M. Lee, D. Lee, and S. Lee. "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 10444–10453.

[11] J. Lee, M. Lee, D. Lee, and S. Lee. "Hierarchically decomposed graph convolutional networks for skeleton-based action recognition". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, 10444–10453.

[12] F. Zhu and R. Zhu, (2021) *"Dance Action Recognition and Pose Estimation Based on Deep Convolutional Neural Network."* **Traitement Du Signal *38***(2): DOI: 10.18280/ts.380233.

[13] Z. Zhan, X. Mao, H. Liu, and S. Yu, (2025) *"STGL: Self-Supervised Spatio-Temporal Graph Learning for Traffic Forecasting"* **Journal of Artificial Intelligence Research *2***(1): 1–8. DOI: 10.70891/JAIR.2025.040001.

[14] W. Zhang and J. Wang, (2024) *"English Text Sentiment Analysis Network based on CNN and U-Net"* **Journal of Science and Engineering *1***(1): 13–18. DOI: 10.70891/JSE.2024.100009.

[15] B. Li, Y. Zhao, S. Zhelun, and L. Sheng. "Danceformer: Music conditioned 3d dance generation with parametric motion transformer". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. ***36***. 2. 2022, 1272–1279. DOI: 10.1609/aaai.v36i2.20014.