Prototype-Guided Meta-Learning for Skeleton-Based One-Shot Action Recognition

Jing Zhou
School of Computer Science
Shenyang Aerospace University
Shenyang, China
zhoujing@stu.sau.edu.cn

Cuiwei Liu
School of Computer Science
Shenyang Aerospace University
Shenyang, China
liucuiwei@sau.edu.cn

Huaijun Qiu
School of Computer Science
Shenyang Aerospace University
Shenyang, China
20220071@email.sau.edu.cn

***Abstract*—This paper addresses Skeleton-based One-shot Action Recognition (SOAR), aiming to recognize novel actions from only a single reference sample. The core challenge lies in learning a generalizable feature space by leveraging data-rich base actions. Meta-learning has emerged as the prevalent paradigm by constructing numerous meta-tasks over base actions to simulate one-shot recognition scenarios. However, the classifiers built from one-shot support samples often fail to capture the complete action distribution in meta-tasks, leading to unstable optimization and limited generalization. To overcome these issues, we propose a Prototype-Guided SOAR (PGSOAR) framework featuring a teacher-student structure. The teacher model is a pre-trained base-action classifier, and the student model learns to solve the SOAR task. During episode training, the teacher simultaneously extracts action prototypes for each meta-task, using them as support data to construct a robust target classifier. The student model, implemented via a Multi-stream Adaptive Spatio-Temporal Graph Convolutional Network (MAST-GCN), is then supervised by the target classifier through knowledge distillation, thereby achieving more reliable and stable optimization. At test time, an Adaptive Prototype Refinement (APR) strategy alleviates single-reference bias by rectifying novel action prototypes with a minimal number of unlabeled query samples. Extensive experiments on two large-scale benchmarks demonstrate that the proposed PGSOAR framework consistently outperforms previous methods, validating its strong generalization and robustness in one-shot action recognition.**

***Index Terms*—Skeleton-based one-shot action recognition, teacher-student structure, prototype**

I. Introduction

Action recognition is one of the fundamental research directions in computer vision and is widely applied in video surveillance and medical rehabilitation [1], [2]. In particular, skeleton-based action recognition has gained prominence owing to its distinctness and privacy-preserving attributes. Most of existing approaches [3], [4] require extensive annotated data to train a high-performance action recognition model. However, a common challenge in real-world scenarios arises when action classes continually expand, making it impractical to assemble large-scale training data for these new actions in a short period. Therefore, few-shot learning [5], [6] has emerged as a promising solution, allowing action recognition models to adapt quickly to new actions by leveraging a minimal number of annotated samples.

This paper focuses on Skeleton-based One-shot Action Recognition (SOAR) [7], an challenging task where a new action is recognized based on a single reference sample. The core challenge of SOAR lies in developing generalizable representations that leverage prior knowledge from well-learned base actions to recognize novel actions with scarce samples. Early studies [8], [9] employ deep metric learning [10] to optimize a feature space with high separability for base actions. In this feature space, a nearest-neighbor classifier is built using reference samples for classification of novel actions. However, these approaches highly rely on base actions, leading to a limitation in generalization toward novel actions that differ substantially from previously learned ones.

Advanced SOAR methods [11], [12] typically follow the meta-learning paradigm. By enabling models to "learn how to learn", these approaches aim to achieve fast adaptation to new tasks, rather than enhancing discrimination on base actions. Specifically, they leverage episode training to construct a lot of meta-tasks on base actions. Each meta-task simulates a one-shot action recognition scenario by randomly sampling a support set and a query set. The support set comprises one annotated example per action for constructing a few-shot classifier (i.e., nearest-neighbor classifier), while the generalization performance of this classifier is evaluated on the query set. Although episode training ensures diversity in meta-tasks, the few-shot classifier constructed from one support sample struggles to capture the complete data distribution of an action. This leads to unstable optimization objectives and constrains the generalization capability of the learned feature space.

This paper proposes a Prototype-Guided Skeleton-based One-shot Action Recognition framework (PGSOAR), where base action prototypes provide strong training signals for meta-tasks. Concretely, the proposed PGSOAR adopts a teacher-student structure [13], where the teacher model is a pre-trained classifier of base actions and the student model learns to achieve the SOAR task. During episode training, the teacher model simultaneously extracts action prototypes for each meta-task and uses them as support data to construct

This work was supported in part by the Natural Science Foundation of Liaoning Province, China under Grant No. 2025-MSLH-561, and in part by the Foundation of Liaoning Educational Committee under Grant JYTMS20230271, and in part by the High Level Talent Research Start-up Fund of Shenyang Aerospace University under Grant No.23YB03.

Corresponding author: Cuiwei Liu

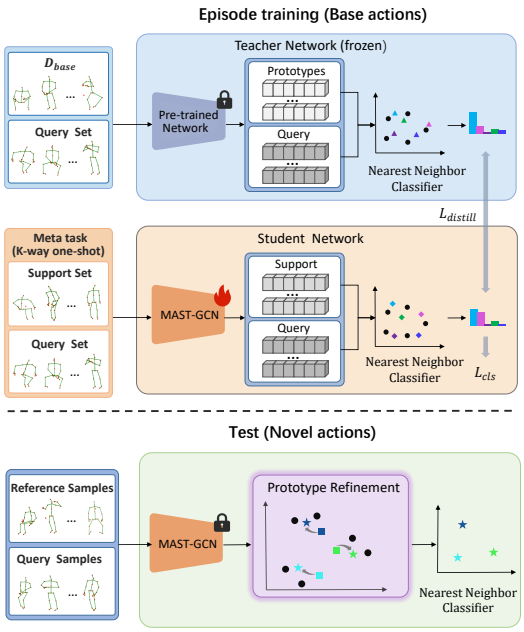

Fig. 1: Overview of the proposed PGSOAR framework.

a robust target classifier. This classifier then serves as a strong guidance through knowledge distillation, delivering a reliable supervision for the optimization of student model. The student model is implemented by a Multi-stream Adaptive Spatio-Temporal Graph Convolutional Network (MAST-GCN), which adaptively transfers crucial cues from the motion stream to the joint stream with a spatio-temporal channel attention mechanism [14]. At test time, an Adaptive Prototype Refinement (APR) strategy is introduced by integrating individual reference samples with a limited number of unlabeled query samples into more stable class representations, thereby enhancing the robustness of novel action classifier. Our main contributions are summarized as follows.

- We propose the PGSOAR framework, where base action prototypes offer stable learning targets for meta-tasks via knowledge distillation. This framework not only capitalizes on the diversity of meta-tasks but also incorporates reliable prior knowledge, thus achieving a feature space with strong generalization and discrimination capabilities.
- We develop MAST-GCN as the student model, which dynamically attends to key joints and frames to enhance fine-grained action discrimination.
- We introduce a prototype refinement strategy that fuses reference samples with a few unlabeled queries to construct enriched prototypes of novel actions, alleviating single-reference bias without extra annotation.

II. Related Work

A. Skeleton-based one-shot action recognition

SOAR has gained significant attention as an important research direction in computer vision. Liu et al. [7] formally established the evaluation protocol on the large-scale NTU RGB+D 120 dataset. Subsequent works [8], [9] first convert skeleton sequences into pseudo-images and then apply CNN-based encoders, thereby constructing a discriminative feature space through metric learning. Also building upon pseudo-images, Peng et al. [15] adopt a Transformer architecture to learn the spatiotemporal correlations in skeleton sequences.

Recently, Graph Convolutional Networks (GCNs) [3] have emerged as the mainstream approaches due to their capability to model the natural skeleton topology. To extract maximal information from the scarce single reference, researchers aim to construct highly discriminative feature extractors with meta-learning. while some works [16] aim to aggregates joint features into a single global vector, alternative strategies seek to preserve distinctive local representations from body parts [11], [12] or semantic parts [17]. Building upon these features, intricate alignment [18], metric mechanisms [19] and multi-scale strategy [20] are employed to overcome the spatio-temporal misalignment between query and support samples.

Although these methods have made significant progress in feature representation and matching mechanisms, they exhibit an excessive reliance on a single reference sample to construct the action classifier. This single-reference bias directly weakens both the learning of the feature space and the generalization capability of the few-shot classifier toward novel actions. The proposed PGSOAR framework incorporates reliable prior knowledge from base action prototypes during the training phase to build a feature space with strong discriminative power. During the testing phase, it constructs enriched prototypes for novel actions in an unsupervised manner to mitigate the single-reference bias.

B. Knowledge distillation

Knowledge distillation [13] aims to transfer knowledge from a complex teacher to a light-weight student by aligning their output distributions. Typically, soft labels derived from posterior probabilities of the teacher are utilized in this process. Compared to hard labels, soft labels [21] provide dense supervision by revealing inter-class semantic correlations, thereby enhancing the generalization capability of the student.

In skeleton-based action recognition task, knowledge distillation strategies have been widely adopted to enhance the model robustness and efficiency. A two-stage distillation framework [22] guides the student to capture temporal evolution patterns to achieve early action recognition. Furthermore, auxiliary modalities like textual semantics can be leveraged to enhance the discriminability of skeleton representations [23]. To strike a balance between performance and efficiency, a spiking graph convolutional network [24] integrates multimodal fusion with knowledge distillation in a unified framework.

This paper focuses on SOAR and transfers knowledge from pre-trained teacher to mitigate sample scarcity in one-shot meta-tasks. While learning from diverse meta-tasks, our proposed PGSOAR framework introduces strong supervision signals derived from action prototypes via knowledge distillation, thereby achieving stable feature space optimization.

III. METHOD

A. Problem definition

The goal of SOAR is to accurately classify unlabeled samples when only a single reference sample is available for each action. The core of this task lies in leveraging the rich data from base actions to learn a discriminative and generalizable feature space, where samples from the same action are tightly clustered, while those from different actions are dispersed. Specifically, the SOAR model is trained on a base dataset D_{base} with abundant labeled samples from M base actions and then evaluated on a novel dataset D_{novel} consisting of N previously unseen actions. In this evaluation, each action in D_{novel} is represented by a single reference sample.

This study follows the fundamental meta-learning paradigm and employs episode training. In each episode, a K-way one-shot meta-task is constructed by randomly sampling a support set S and a query set Q from D_{base}. The support set $S = \{(x_{\text{s}}^i, y_{\text{s}}^i)\}_{i=1}^K$ contains K base actions, each represented by one labeled sample. Here, x_s^i denotes the skeleton sequence of the i-th action and y_{s}^i is its action label. Correspondingly, the query set $Q = \{x_{\text{q}}^j, y_{\text{q}}^j\}_{j=1}^K$ contains K samples from these actions. In each meta-task, a nearest-neighbor classifier is built on the support samples to predict the action class of query samples. The feature space is progressively optimized by minimizing the discrepancy between the predicted and ground-truth labels, making it specialized for one-shot action recognition.

B. PGSOAR

Episode training with a single support sample often fails to capture data distributions, leading to unstable optimization. To address this, we propose PGSOAR framework, enhanced by auxiliary supervision derived from base action prototypes. At test time, an N-way nearest-neighbor classifier is built on the reference samples of novel actions in D_{novel}. Our PGSOAR framework mitigates the single-reference bias by incorporating an adaptive prototype refinement strategy, which fuses reference samples with a limited number of unlabeled queries into enhanced class prototypes.

1) Teacher-student structure: Inspired by [25], the PGSOAR framework adopts a teacher-student architecture, as shown in Figure 1. The teacher model is pre-trained on D_{base} to generate high-quality base action prototypes. During episode training, the student model learns a discriminative and generalizable feature space under guidance of these prototypes. Once the student model constructs a meta-task including a support set S and a query set Q, the teacher model simultaneously retrieves the action prototypes. Specifically, in each K-way one-shot meta-task, the teacher model extracts the prototypes corresponding to these K actions from a pre-constructed action prototype bank $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_M\}$, forming a prototype support set S^A. Each action prototype $\mathcal{A}_a$ is computed by

$$\mathcal{A}_a = \frac{1}{|D_{\text{base}}^a|} \sum_{x_i \in D_{\text{base}}^a} f(x_i), \tag{1}$$

where D_{base}^a denotes the set of all samples of action a in D_{base}, and $f(\cdot)$ represents the feature extractor of the teacher model. This calculation effectively integrates the feature distribution within the action through mean aggregation, filtering out noise from individual samples. Then a target nearest-neighbor classifier is built on the prototype support set S^A and makes predictions for query samples in Q, providing stable and reliable supervisory signals to the student model.

2) Overall optimization objective: In each K-way one-shot meta-task, the student model extracts features from the query set Q and makes prediction using the nearest-neighbor classifier constructed from the support set S. For a query sample x_{q}^j, the logits $\mathbf{h}_{\text{q}}^j = [h_{\text{q}}^{j,1}, ..., h_{\text{q}}^{j,K}]$ output from the classifier are passed through a softmax function to obtain a K-dimensional probability distribution $\mathbf{p}_{\text{q}}^j = [p_{\text{q}}^{j,1}, ..., p_{\text{q}}^{j,K}]$, representing the posterior probabilities of a query sample belonging to each action class. The one-shot classification loss is computed as the Cross-Entropy (CE) between the predicted probabilities and the ground-truth labels by

$$L_{\text{cls}} = -\frac{1}{K} \sum_{j=1}^{K} \sum_{a=1}^{K} y_{\text{q}}^{j,a} \log(p_{\text{q}}^{j,a}), \tag{2}$$

where K is the number of actions, and each action has only one query sample, so K is also the number of query samples. The ground-truth label $y_{\text{q}}^{j,a}$ is defined as 1 if the query sample x_{q}^j belongs to action class a, and 0 otherwise. $p_{\text{q}}^{j,a}$ indicates the predicted probability that sample x_{q}^j belongs to action a. By optimizing L_{cls}, the student model learns to classify query samples into their corresponding actions.

To achieve effective guidance from the teacher to the student, we introduce the knowledge distillation loss L_{distill} implemented by the Kullback-Leibler (KL) divergence between the logits distributions of the student model and the teacher model on query set Q.

$$L_{\text{distill}} = \sum_{j=1:K} \text{KL}(\mathbf{h}_{\text{q}}^j \,\|\, \hat{\mathbf{h}}_{\text{q}}^j), \tag{3}$$

where $\mathbf{h}_{\text{q}}^j$ and $\hat{\mathbf{h}}_{\text{q}}^j$ indicate logits of x_{q}^j output from the student and teacher, respectively. The student model makes predictions using a nearest-neighbor classifier constructed on the support set S, whereas the teacher model relies on a target classifier built on the prototype support set S^A.

Finally, we design an adaptive multi-objective optimization strategy to integrate the one-shot classification loss L_{cls} and the knowledge distillation loss L_{distill}. The former guides the model to acquire one-shot classification capability with hard labels, and the latter employs soft targets to enable the student model to mimic the teacher's output distribution. Given the evolving nature of the learning characteristics, the two losses are dynamically balanced at different training phases. The overall optimization objective is by

$$L = \left(1 - \alpha \cdot \frac{\bar{e}}{e}\right) L_{\text{cls}} + \left(\alpha \cdot \frac{\bar{e}}{e}\right) L_{\text{distill}}, \tag{4}$$

where e denotes the total training epochs and $\bar{e}$ indicates the current epoch. The hyperparameter α controls the weights of L_{cls} and L_{distill}. In the early stage, the overall objective favors

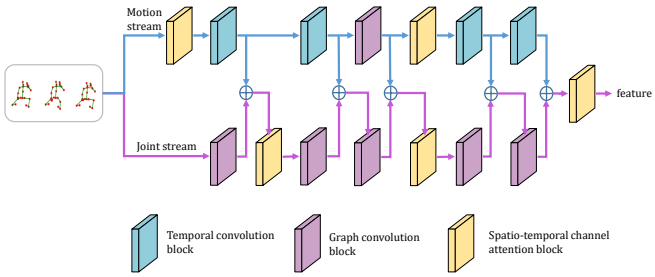

Fig. 2: Overall architecture of the MAST-GCN.

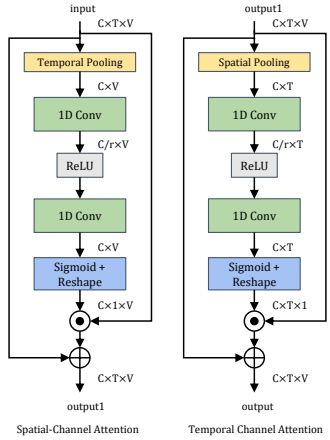

Fig. 3: Details of the spatio-temporal channel attention module in MAST-GCN.

L_{cls}, enforcing the student model to acquire one-shot classification capabilities and build stable feature representations. Subsequently, the weight of L_{distill} is gradually increased, enabling the model to transfer knowledge from action prototypes.

C. MAST-GCN

The student model utilizes MAST-GCN, a parallel architecture that adaptively fuses joint and motion streams. As shown in in Figure 2, to capture fine-grained action patterns, spatio-temporal channel attention modules [14] are embedded within both branches, enabling the network to dynamically focus on salient joints and temporal segments. This mechanism hierarchically refines local spatio-temporal features before an average pooling operation integrates them into the final global sequence representation.

Figure 3 illustrats the architecture of the spatio-temporal channel attention module [14], which employs a dual-branch architecture to independently learn channel attention weights along the spatial and temporal dimensions. Specifically, given an input feature $X \in \mathbb{R}^{(C\times T\times V)}$, where C, T, and V denote the channel count, temporal length, and number of joints respectively, the spatial channel attention branch compresses X into $X_{cs} \in \mathbb{R}^{(C\times V)}$ through temporal pooling, eliminating the temporal dimension to focus on spatial patterns. The channel dimension is reduced to C/r via a 1D convolutional layer, where r denotes the reduction ratio. Subsequently, a ReLU activation function is applied to introduce non-linearity. A 1D convolution is applied to restore the original channel dimension, generating the spatial attention weights $A_{cs} \in \mathbb{R}^{(C\times 1\times V)}$ through a Sigmoid function. The temporal channel attention branch follows a similar yet complementary procedure. It first compresses the input into $X_{ct} \in \mathbb{R}^{(C\times T)}$ through spatial pooling, which preserves temporal information while eliminating the spatial dimension. Subsequently, a 1D convolution is applied, followed by subsequent processing layers, to generate the temporal attention weights $A_{ct} \in \mathbb{R}^{(C\times T\times 1)}$. Finally, the attention weights from both branches are fused with the original input features through element-wise multiplication.

$$Y = (X + X \odot A_{cs}) \odot A_{ct}, \tag{5}$$

where $\odot$ denotes element-wise multiplication.

D. Adaptive prototype refinement strategy

Typically, reference samples of novel actions are mapped into the feature space optimized by MAST-GCN to construct an N-way nearest-neighbor classifier. This classifier uses the single reference sample r_n as the prototype of action n and is therefore inherently susceptible to single-reference bias. To this end, we randomly sample a subset $\mathcal{B}$ comprising B unlabeled queries from D_{novel} and merge them with the reference samples $\{r_1, r_2, \ldots, r_N\}$ to form more representative prototypes of novel actions.

Specifically, we compute the Euclidean distance $d(x_b, r_n)$ between a query sample $x_b \in \mathcal{B}$ and the reference sample r_n of action n in the optimized feature space. The nearest-neighbor classifier assigns a pseudo-label $y_b = \arg\min_n d(x_b, r_n)$ to each query sample according to these distance statistics. To further quantify classification reliability, we derive the confidence score $\omega_{b,n}$ for sample x_b belonging to action n using a Softmax function:

$$\omega_{b,n} = \frac{\exp(-d(x_b, r_n))}{\sum_{j=1}^{N} \exp(-d(x_b, r_j))}. \tag{6}$$

Finally, the prototypes of novel actions are updated by

$$r'_n = \frac{r_n + \sum_{x_b\in\mathcal{B}} \mathbb{I}(y_b = n) \cdot \omega_{b,n} \cdot x_b}{1 + \sum_{x_b\in\mathcal{B}} \mathbb{I}(y_b = n) \cdot \omega_{b,n}}, \tag{7}$$

where the indicator function $\mathbb{I}(\cdot)$ equals 1 if the condition is met and 0 otherwise. Through this process, the prototypes of novel actions are endowed with prior knowledge from a minimal number of queries, effectively rectifying the inherent bias in single-reference scenarios without extra annotations.

IV. Experiments

A. Datasets

Our proposed PGSOAR is evaluated on two widely used SOAR datasets. The NTU RGB+D 60 dataset [26] is a large-scale action recognition dataset containing 56,880 action samples. It was constructed by capturing 60 types of daily interactive actions performed by 40 subjects. Each action sequence is represented by skeleton motion data in the form of 3D coordinate sequences for 25 body joints. The NTU RGB+D

TABLE I: One-shot action recognition accuracy (%) on NTU RGB+D 60 and NTU RGB+D 120 datasets.

Method	NTU RGB+D 60	NTU RGB+D 120				
	Base Classes	Base Classes				
	50	20	40	60	80	100
APSR [7]	-	29.10	34.80	39.20	42.80	45.30
SL-DML [8]	54.82	36.70	42.40	49.00	46.40	50.90
Skeleton-DML [9]	55.54	28.60	37.50	48.60	48.00	54.20
SMAM-Net [16]	73.60	35.80	46.20	51.70	52.20	56.40
JEANIE [18]	80.00	38.50	44.10	50.30	51.20	57.00
Trans4SOAR [15]	74.19	-	-	-	-	57.05
ALCA-GCN [12]	-	38.70	46.60	51.00	53.70	57.60
STA-MLN [19]	-	42.40	48.00	53.10	54.30	59.90
Part-Aware [11]	-	43.00	50.30	55.70	56.50	65.60
M&C-scale [20]	82.70	44.10	55.30	60.30	64.20	68.70
APLE-GCN [17]	81.65	43.64	53.88	61.12	66.70	69.06
PGSOAR (Ours)	**83.31±0.24**	**44.29±0.62**	**49.54±0.74**	**61.74±1.29**	**69.30±1.1**	**72.09±0.51**

120 dataset [7] extends the NTU RGB+D 60 dataset and is one of the largest publicly available benchmarks for SOAR. It contains a total of 114,480 action samples, expanding the original set of 60 actions with 60 additional activities, resulting in 120 distinct actions performed by 106 participants.

B. Experimental settings

For fair comparison with existing methods, we conduct experiments following standard SOAR benchmarks [7]. On the NTU RGB+D 60 dataset, the first 50 action categories are taken as base actions, with all corresponding samples used for training. The remaining 10 actions are treated as novel classes, forming the validation set. Similarly, NTU RGB+D 120 is split into 100 base and 20 novel actions. For each novel action, one sequence is used as the reference sample, and the rest serve as queries.

To ensure data format consistency, all skeleton sequences are normalized to 60 frames through uniform sampling and zero-padding. During the pre-training phase, we adopt ST-GCN [3] as the teacher model, which is trained on the full base action data using standard cross-entropy loss. In the meta-training phase, our model is trained for 100 epochs with 1,000 episodes per epoch.

The Adam optimizer is employed with a learning rate of 0.001 and a weight decay of 10^{-6}. We report the results with parameter α in Equation (4) fixed at 0.4. The influence of this key hyperparameter is evaluated through an ablation study. During the testing phase, a few unlabeled query samples are employed to refine the prototypes of novel actions. The impact of the number of these samples is evaluated in Section IV-D.

C. Experimental results

Table I presents the performance of the proposed PGSOAR on both datasets, comparing it against existing SOAR methods. To ensure statistical reliability and mitigate the variance from random sampling in the adaptive prototype refinement strategy, we report the average performance of PGSOAR across ten independent experiments. Trained on 50 and 100 base classes for NTU RGB+D 60 and NTU RGB+D 120 respectively, our proposed PGSOAR consistently achieves superior performance compared to all existing methods on both benchmarks.

On the larger-scale NTU RGB+D 120 dataset, we further evaluate the 20-way One-shot action classification performance when employing different numbers of base classes (20, 40, 60, 80). As the number of base classes increases, the action recognition performance of all methods consistently improves, which demonstrates the importance of base action diversity for enhancing the generalization ability. Our PGSOAR outperforms previous methods based on RNNs [7], CNNs [8], [9], Transformer [15], and GCNs [12], [16], [18], [19] under all configurations of base classes. In comparison with part-based [11], [17] and multi-scale [20] representations, our method aggregates joint features into a compact global representation, achieving superior performance under settings with a large number of base classes, while maintaining competitive results even with fewer base classes. This highlights its robustness across varying base class scales by transferring reliable knowledge from the teacher model.

TABLE II: Ablation study results (%).

Teacher	APR	NTU RGB+D 60	NTU RGB+D 120
		78.70	65.31
✓		80.06	66.88
✓	✓	83.31±0.24	72.09±0.51

TABLE III: Evaluation of α on the NTU RGB+D 60 dataset.

parameter α	α=0.3	α=0.4	α=0.5	α=0.6
Accuracy	79.68	80.06	79.74	79.76

D. Ablation study

We perform an ablation study to analyze the contribution of the two key components in PGSOAR. As shown in Table II, the teacher guidance optimizes a more generalizable feature space. The unsupervised APR strategy increases the precision of the novel-class by 3.25% and 4.85% on NTU RGB+D 60 and NTU RGB+D 120, respectively. These results confirm that PGSOAR effectively mitigates severe single-reference bias.

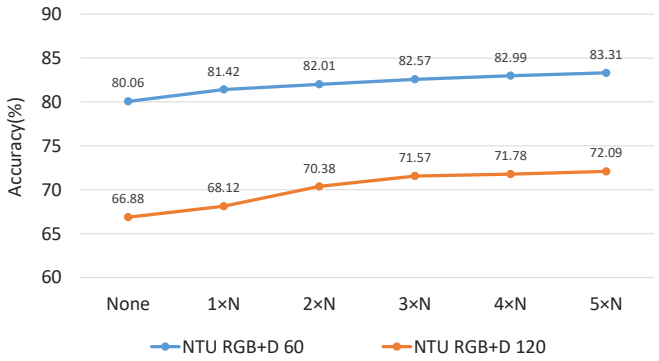

Fig. 4: Evaluation on the query sample number in APR.

Next, we also evaluate the distillation weight parameter α in Equation (4) and the unlabeled query subset size $|\mathcal{B}|$. Table III reveals that performance peaks at $\alpha = 0.4$. This empirical finding indicates that the student model strikes an optimal balance between learning discriminative features from the classification loss and absorbing structured knowledge via distillation. For the APR strategy, figure 4 summarizes the model performance on both datasets when the subset size ranges from $1 \times N$ to $5 \times N$. Results show that significant performance improvement can be achieved with an exceptionally small set of unlabeled queries, highlighting the data efficiency of the proposed APR strategy.

V. Conclusion

This paper have presented the PGSOAR framework to address the single-reference bias in skeleton-based one-shot action recognition. During episode training, PGSOAR employs a teacher-student architecture where the teacher provides prototype-based supervision, while the student utilizes MASTGCN for adaptive multi-stream feature fusion. During testing, the APR strategy refines novel action prototypes using 5N unlabeled queries. Extensive experiments and ablation studies on NTU RGB+D 60 and NTU RGB+D 120 benchmarks confirm PGSOAR's strong generalization and the effectiveness of its core components.

References

[1] C. Liu, J. Lv, X. Zhao, Z. Li, Z. Yan, and X. Shi, "A novel key point trajectory model for fall detection from rgb-d videos," in *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design*, pp. 1021–1026, IEEE, 2021.

[2] Y. Kong and Y. Fu, "Human action recognition and prediction: a survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[3] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, pp. 7444–7452, 2018.

[4] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1474–1488, 2023.

[5] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen, "Temporal-relational crosstransformers for few-shot action recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 475–484, 2021.

[6] S. Zheng, S. Chen, and Q. Jin, "Few-shot action recognition with hierarchical matching and contrastive learning," in *European conference on computer vision*, pp. 297–313, Springer, 2022.

[7] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Mattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[8] R. Remmesheimer, N. Theisen, and D. Paulus, "Sld-ml: Signal level deep metric learning for multimodal one-shot action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pp. 4573–4580, 2021.

[9] R. Remmesheimer, S. Haring, N. Theisen, and D. Paulus, "Skeleton-ml: Deep metric learning for skeleton-based one-shot action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3709–3710, 2020.

[10] K. Musgrave, S. Belongie, and S.-N. Lim, "A metric learning reality check," in *Proceedings of the European Conference on Computer Vision*, pp. 681–699, Springer, 2020.

[11] T. Chen, D. Zhou, J. Wang, S. Wang, Q. He, C. Hu, E. Ding, Y. Guan, and X. He, "Part-aware prototypical graph network for one-shot skeleton-based action recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–8, IEEE, 2023.

[12] A. Zhu, Q. Ke, M. Gong, and J. Bailey, "Adaptive local-component-aware graph convolutional network for one-shot skeleton-based action recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6038–6047, 2023.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[14] S. Wang, J. Pan, B. Huang, P. Liu, Z. Li, and C. Zhou, "Ice-gcn: An interactional channel excitation-enhanced graph convolutional network for skeleton-based action recognition," *Machine Vision and Applications*, vol. 34, no. 3, p. 40, 2023.

[15] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Delving deep into one-shot skeleton-based action recognition with diverse occlusions," *IEEE Transactions on Multimedia*, vol. 25, pp. 1489–1504, 2023.

[16] Z. Li, X. Gong, R. Song, P. Duan, J. Liu, and W. Zhang, "Smam: Self and mutual adaptive matching for skeleton-based few-shot action recognition," *IEEE Transactions on Image Processing*, vol. 32, pp. 392–402, 2022.

[17] C. Liu, S. Liu, H. Qiu, and Z. Li, "Adaptive part-level embedding gcn: Towards robust skeleton-based one-shot action recognition," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[18] L. Wang, J. Liu, L. Zheng, T. Gedeon, and P. Koniusz, "Meet jeanie: a similarity measure for 3d skeleton sequences via temporal-viewpoint alignment," vol. 132, pp. 4091–4122, Springer, 2024.

[19] X. Li, J. Lu, X. Chen, and X. Zhang, "Spatial-temporal adaptive metric learning network for one-shot skeleton-based action recognition," *IEEE Signal Processing Letters*, vol. 31, pp. 321–325, 2024.

[20] S. Yang, J. Liu, S. Lu, E. M. Hwa, and A. C. Kot, "One-shot action recognition via multi-scale spatial-temporal skeleton matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 5149–5156, 2024.

[21] B. B. Sau and V. N. Balasubramanian, "Deep model compression: Distilling knowledge from noisy teachers," *arXiv preprint arXiv:1610.09650*, 2016.

[22] C. Liu, X. Zhao, Z. Li, Z. Yan, and C. Du, "A novel two-stage knowledge distillation framework for skeleton-based action prediction," *IEEE Signal Processing Letters*, vol. 29, pp. 1918–1922, 2022.

[23] H. Xu, Y. Gao, Z. Hui, J. Li, and X. Gao, "Language knowledge-assisted representation learning for skeleton-based action recognition," *IEEE Transactions on Multimedia*, 2025.

[24] N. Zheng, H. Xia, Z. Liang, and Y. Du, "Mk-sgn: A spiking graph convolutional network with multimodal fusion and knowledge distillation for skeleton-based action recognition," *Neurocomputing*, p. 131796, 2025.

[25] H.-J. Ye, L. Ming, D.-C. Zhan, and W.-L. Chao, "Few-shot learning with a strong teacher," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 46, no. 03, pp. 1425–1440, 2024.

[26] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1010–1019, 2016.