

Boosting Few-shot Action Recognition with Graph-guided Hybrid Matching

Jiazheng Xing^{1*}, Mengmeng Wang^{1*}, Yudi Ruan¹, Bofan Chen¹, Yaowei Guo¹,
Boyu Mu¹, Guang Dai^{2,3}, Jingdong Wang⁴, Yong Liu^{1†}

¹ Zhejiang University, ² SGIT AI Lab, ³ State Grid Corporation of China, ⁴ Baidu Inc.

{jiazhengxing, mengmengwang, yudiruan, bofanchen, guoyaowei, muboyu}@zju.edu.cn
yongliu@iipc.zju.edu.cn, guang.gdai@gmail.com, wangjingdong@baidu.com

Abstract

Class prototype construction and matching are core aspects of few-shot action recognition. Previous methods mainly focus on designing spatiotemporal relation modeling modules or complex temporal alignment algorithms. Despite the promising results, they ignored the value of class prototype construction and matching, leading to unsatisfactory performance in recognizing similar categories in every task. In this paper, we propose GgHM, a new framework with **Graph-guided Hybrid Matching**. Concretely, we learn task-oriented features by the guidance of a graph neural network during class prototype construction, optimizing the intra- and inter-class feature correlation explicitly. Next, we design a hybrid matching strategy, combining frame-level and tuple-level matching to classify videos with multivariate styles. We additionally propose a learnable dense temporal modeling module to enhance the video feature temporal representation to build a more solid foundation for the matching process. GgHM shows consistent improvements over other challenging baselines on several few-shot datasets, demonstrating the effectiveness of our method. The code will be publicly available at <https://github.com/jiazheng-xing/GgHM>.

1. Introduction

Compared with general action recognition, few-shot action recognition requires limited labeled samples to learn new categories quickly. It can avoid the massive, time-consuming, and labor-consuming data annotation commonly associated with supervised tasks, making it more adaptable for industrial applications. According to this advantage, increasing attention has been directed toward the field of few-shot action recognition [4, 27, 30, 35, 44, 14, 23, 38]. However, since few-shot action recognition has

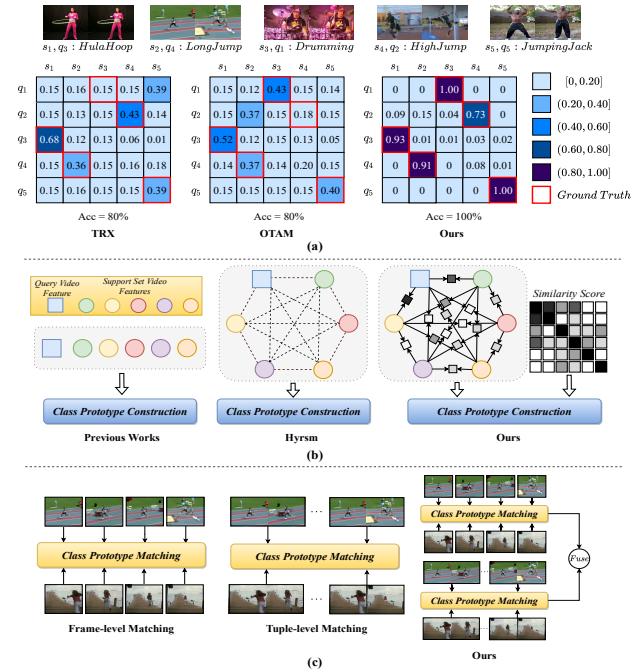


Figure 1. (a): Similarity visualization between query and support videos with different methods on the 5-way 1-shot task of UCF101 [29]. A higher score indicates a greater degree of similarity. TRX [27] misclassifies the drumming as the jumping jack, and OTAM [4] misidentifies the high jump as the long jump. Our method identifies all categories of videos accurately. (b): Different types of class prototype construction. Previous works did not do any information interaction among different videos. HyRSM [35] operates an inter-relation function without leveraging label-informed supervision. Our method utilizes the graph network with label-informed supervision to learn the correlation between different videos. (c): Different types of class prototype matching. Frame-level matching [46, 4, 35] uses single individual frames for matching, while tuple-level [30, 38, 27] matching combines several frames into a tuple as the matching unit. Our method combines both to complement each other's shortcomings.

*Equal Contribution.

†Corresponding author.

limited learning material, learning well-generalized models are challenging.

Current attempts to address the above problems [4, 46, 41, 27, 30, 38, 35] mainly adopt the metric-based framework and episode training to solve the difficulty of model migration on new categories. Empirically, we observed that previous approaches failed to effectively address the problem of misclassification of videos from similar categories. Taking the action of the *high jump* and *long jump* as an instance, some methods (e.g., OTAM [4]) is easy to confuse the two classes by assigning close prediction scores due to their similarity in scenes and sub-actions, as shown in Fig. 1(a). We have analyzed the main reasons from three folds. (i) Class prototype construction: task-oriented class features can optimize videos’ intra- and inter-class correlation. As shown in Fig. 1(b), most previous work has yet to use the whole task video features to extract relevant discriminative patterns. Although HyRSM [35] manipulates interrelationship functions on different videos to get task-specific embeddings, it does not explicitly optimize intra- and inter-class correlations. (ii) Matching mechanisms: proper matching mechanisms need to be established to solve the confusion problem of similar videos. As shown in Fig. 1(c), current work almost all use a simple class prototype matching mechanism. Some methods use the frame-level matching mechanism [46, 4, 35], which is suitable for spatial-related datasets [16, 29, 5], and the others use the tuple-level(multiple frames combined into a tuple) matching mechanism [30, 38, 27] that is appropriate for temporal-related datasets [13]. None of these previous methods can cope with video tasks of variable types well. (iii) Feature modeling: a powerful and highly discriminative feature is first needed to distinguish similar classes. Most previous works model the temporal feature through hand-designed temporal alignment algorithms [46, 4] or simple temporal attention operations [35, 38], leading to a simplistic exploration of the temporal relationship without dissecting it into more detailed patch and channel temporal relations to analyze.

Based on the above observations, we propose a novel method for few-shot action recognition, dubbed **GgHM**, a short for **G**raph-guided **H**ybrid **M**atching. Specifically, we apply a graph neural network (GNN) for constructing task-oriented features, as shown in Fig.1(b). It could interactively transfer information between video features in a task to enhance the prior knowledge of the unknown video. We utilize the ground truth of the constructed graph edges to explicitly learn the correlation of these video features to supervise the similarity score learning between the query and support videos. Second, as shown in Fig.1(c), we propose a hybrid prototype matching strategy that combines frame-level and tuple-level matching based on the bidirectional Hausdorff Distance. Although the Hausdorff metric frame-

level matching can alleviate the strictly ordered constraints of acquiring better query-support correspondences, it fails to capture temporal order. As a result, it can be confused for actions with similar action scenes strongly dependent on temporal order, e.g., *putting something in the box* and *taking something out of it*. However, the construction of tuples strictly follows a chronological order, which can compensate for the frame-level matching problem. Fig.1(a) visualizes the predicted similarities between query and support videos with different methods on the 5-way 1-shot task of UCF101 [29]. Our method achieves more discriminative results for similar videos in each task compared to OTAM [4] and TRX [27]. Additionally, we design a learnable dense temporal modeling module to consolidate the representation foundation. It includes a temporal patch and temporal channel relation modeling block, and their combination allows for dense temporal modeling in both spatial and channel domains. Finally, extensive experiments on four widely-used datasets demonstrate the effectiveness of our method.

In summary, we make the following contributions:

- We apply a graph neural network to guide the task-oriented features learning during the class prototype construction, explicitly optimizing the intra- and inter-class correlation within video features.
- We propose a hybrid class prototype matching strategy based on the frame- and tuple-level prototype matching, giving rise to effectively coping with video tasks of multivariate styles.
- We design a learnable dense temporal modeling module consisting of a temporal patch and temporal channel relation modeling block for dense temporal modeling in both spatial and channel domains.

2. Related Works

2.1. Few-shot Image Classification

Few-shot image classification uses the episodic training paradigm, using a handful of labeled training samples from similar tasks to represent a large amount of labeled training samples. Recent years, research on few-shot image classification can be mainly classified into two categories: adaptation-based and metric-based methods. The adaption-based approaches aim to find a network initialization that can be fine-tuned for unknown tasks using a small amount of labeled data, called *gradient by gradient*. The classical adaptation-based approaches are MAML [10], Reptile [25], and related deeper researches include [21, 32]. The metric-based approaches aim to learn a feature space and compare task features through different matching strategies, called *learning to compare*. The representative methods are Prototypical Networks [28], Matching Networks [31]. And there

are many methods [40, 39, 8, 18] that aim to improve upon these approaches. Our method is inspired by them and belongs to the metric-based category.

2.2. Few-shot Video Action Recognition

The core idea of few-shot action recognition is similar to that of few-shot image classification, but the former task is more complex than the latter owing to an additional temporal dimension. Due to high computational resources and long experimental time, adaptation-based methods(MetaUVFS [26]) have received little attention in few-shot action recognition. The existing research mainly applies metric-based learning, but with different focuses. Some methods focus on feature representation enhancement. For example, STRM [30] employs local and global enrichment modules for spatiotemporal modeling, HyRSM [35] uses the hybrid relation modeling to learn task-specific embeddings, and SloshNet [38] utilizes a feature fusion architecture search module to exploit the low-level spatial features and a long-term and short-term temporal modeling module to encode complementary global and local temporal representations. Other methods focus on class prototype matching strategies. For example, OTAM [4] proposes a temporal alignment module to calculate the distance value between the query video and the support set videos, TRX [27] matches each query sub-sequence with all sub-sequences in the support set, HyRSM [35] designs a bidirectional Mean Hausdorff metric to more flexibly find the correspondences between different videos. Additionally, TRPN [34], MORN [24] focus on combining visual and semantic features, and AMeFu-Net [11] centers on using depth information to assist learning. Unlike these previous methods, our method focuses on distinguishing videos from similar categories by optimizing intra- and inter-class class correlation within video features during the prototype construction and building a hybrid prototype matching strategy to effectively handle video tasks of multivariate styles.

3. Method

3.1. Problem Formulation

Few-shot learning is based on using a small number of labeled training samples from similar tasks as a proxy for many labeled training samples. For few-shot action recognition, it aims to classify an unlabeled query video into one of the N action categories in the support set with limited K samples per action class, which can be considered an N -way K -shot task. Like most previous studies, we adopt an episode training paradigm followed by [4, 35, 14, 17, 38], where episodes are randomly selected from extensive data collection. In each episode, we suppose that the set \mathcal{S} consists of $N \times K$ samples from N different action classes, and $S_k^n = \{s_{k1}^n, s_{k2}^n, \dots, s_{kT}^n\}$ represents the k -th video in class

$n \in \{1, \dots, N\}$ randomly sampled T frames. The query video denotes $Q = \{q_1, q_2, \dots, q_T\}$ sampled T frames.

3.2. Architecture Overview

Our overall architecture is illustrated in Fig.2. For the frame-selecting strategy, we follow previous work TSN [33], where the input video sequence is divided into T segments, and snippets are extracted from each segment. For simplicity and convenience, we discuss the process of the 5-way 1-shot problem and consider that the query set Q contains a single video. In this way, the query video $Q = \{q_1, q_2, \dots, q_T\}$ and the class support set videos $S^n = \{s_1^n, s_2^n, \dots, s_T^n\}$ ($S^n \in \mathcal{S} = \{S^1, S^2, \dots, S^5\}$) pass through the feature extractor to obtain the query feature \mathbf{F}_Q and the support features \mathbf{F}_{S^n} ($\mathbf{F}_{S^n} \in \mathbf{F}_{\mathcal{S}}$) in each episode. Next, we input $\mathbf{F}_{\mathcal{S}}$ and \mathbf{F}_Q to the proposed learnable dense temporal modeling module to obtain enhanced temporal features $\widetilde{\mathbf{F}}_{\mathcal{S}}$ and $\widetilde{\mathbf{F}}_Q$. We apply to mean pooling operation on $\widetilde{\mathbf{F}}_{\mathcal{S}}$ and $\widetilde{\mathbf{F}}_Q$ in the temporal dimension to obtain the relation node features $\widetilde{\mathbf{F}}_{\mathcal{S}}^{avg}$ and $\widetilde{\mathbf{F}}_Q^{avg}$ for the following graph network. Then, the relation node features are taken into the graph network with initial edge features for relation propagation. The updated edge features with enhanced temporal features generate task-oriented features $\mathbf{F}_{\mathcal{S}}^{task}$ and \mathbf{F}_Q^{task} and obtain the loss \mathcal{L}_{graph} through a graph metric. Finally, the task-oriented features are fed into the hybrid class prototype matching metric to get the class prediction \hat{y}_Q and loss \mathcal{L}_{match} .

For better clarity and consistency with the algorithm procedure, we will first introduce our learnable dense temporal modeling module, followed by the graph-guided prototype construction, and finally the hybrid prototype matching strategy. Details are shown in the subsequent subsections.

3.3. Learnable Dense Temporal Modeling Module (LDTM)

The action classification process relies heavily on temporal context information. Inspired by some temporal modeling methods based on attention mechanism [1, 43, 9, 37, 2], we design a learnable dense temporal modeling module, which consists of a temporal patch relation modeling block and a temporal channel relation modeling block, as shown in Fig.3. The two blocks are complementary, and their combination allows for dense temporal modeling in both the spatial and channel domains. Compared to PST [37], which uses a fixed patch shift strategy and a channel shift strategy, our learnable patch and channel temporal relation modeling enables the extraction of richer features.

Patch Temporal Relation Modeling (PTRM). Given a video feature map output by the feature extractor $\mathbf{F} \in \mathbb{R}^{N \times T \times C \times H \times W}$, we first reshape it to a sequence as $\mathbf{F}_{seq1} \in \mathbb{R}^{N \times HW \times C \times T}$ and then fed it into the temporal

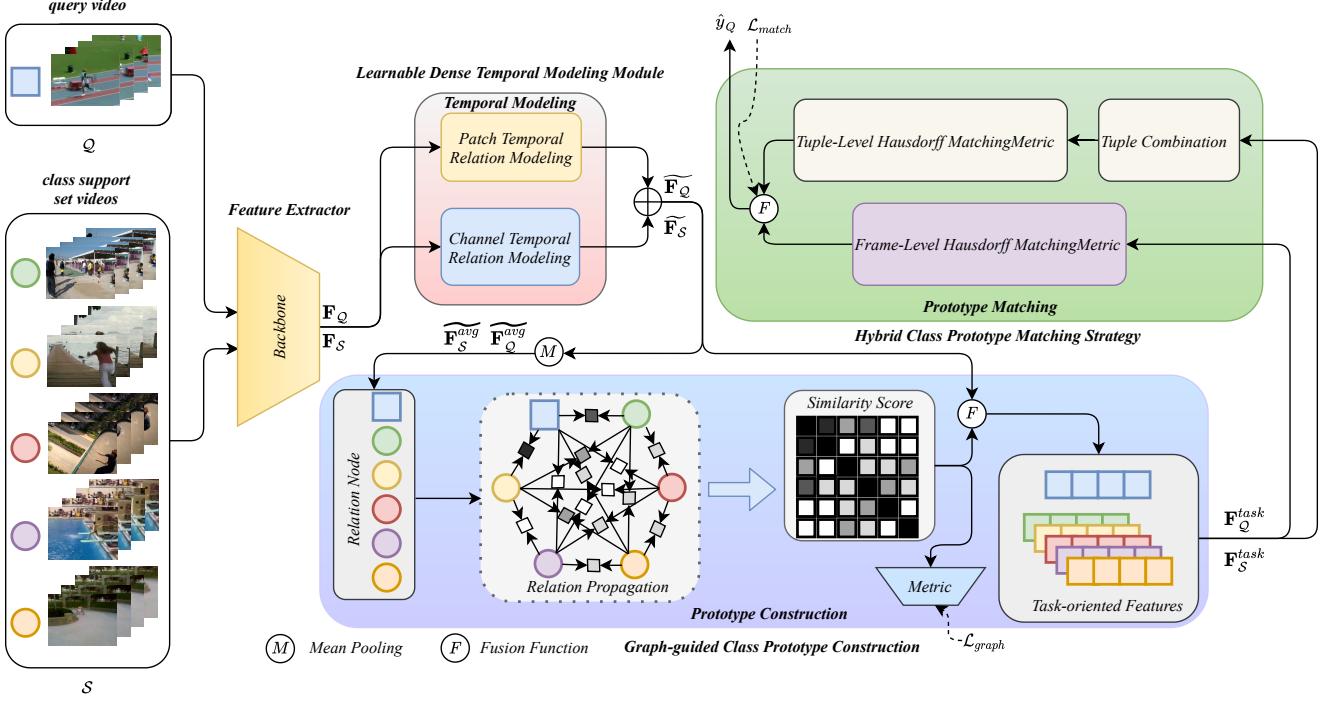


Figure 2. Overview of GgHM. For simplicity and convenience, we discuss the case: the 5-way 1-shot problem and the query set \mathcal{Q} with a single video. The support set video features \mathbf{F}_S and query video feature \mathbf{F}_Q are obtained by the feature extractor. The enhanced temporal features $\widetilde{\mathbf{F}}_S$ and $\widetilde{\mathbf{F}}_Q$ are obtained by the learnable dense temporal modeling module. The task-oriented features \mathbf{F}_S^{task} and \mathbf{F}_Q^{task} are obtained by the graph-guided prototype construction module. The \hat{y}_Q is the class prediction of the query video, and the loss \mathcal{L}_{match} and \mathcal{L}_{graph} are the standard cross-entropy loss. \oplus indicates element-wise weighted summation.

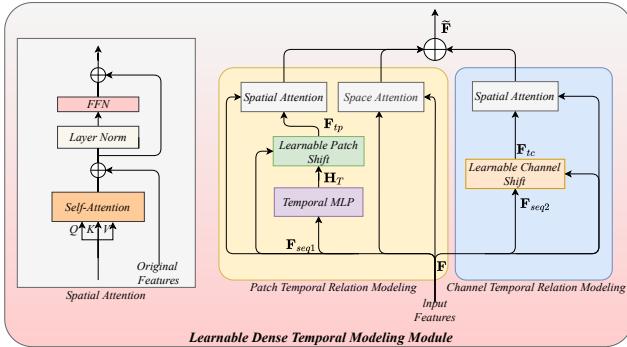


Figure 3. The architecture of learnable dense temporal modeling module. \oplus denotes element-wise summation.

MLP to get hidden temporal feature \mathbf{H}_T :

$$\mathbf{H}_T = \text{relu}(\mathbf{W}_{t1}\mathbf{F}_{seq1})\mathbf{W}_{t2} + \mathbf{F}_{seq1} \quad (1)$$

where \mathbf{W}_{t1} and $\mathbf{W}_{t2} \in \mathbb{R}^{T \times T}$ are the learnable weights for temporal information interaction of different video frames. Then, \mathbf{H}_T with rich video spatiotemporal information are inserted into the original features \mathbf{F}_{seq1} , making the single-frame video feature contain semantic information for all video frames. The temporal patch relation modeling feature

\mathbf{F}_{tp} is obtained by:

$$\mathbf{F}_{tp}[:, n, :, :] = \begin{cases} \mathbf{F}_{seq1}[:, n, :, :] & \text{if } n \% \text{gap} = 0 \\ \mathbf{H}_T[:, n, :, :] & \text{if } n \% \text{gap} \neq 0 \end{cases} \quad (2)$$

where n is the patch index and gap is a positive integer to control the frequency of the patch shift. After the learnable patch shift operation, the feature \mathbf{F}_{tp} is reshaped as $\mathbf{F}_{tp}^* \in \mathbb{R}^{NT \times HW \times C}$ and do spatial self attention. This way collects the temporal information of the different video frames sparsely within the frame but sacrifices the original spatial information within every frame. To alleviate this problem, we do the weighted summation between spatial-only and spatiotemporal attention results, given by:

$$\mathbf{F}_{tp} = \gamma SA_{spa}(\mathbf{F}_{tp}^*) + (1 - \gamma) SA_{spa}(\mathbf{F}^*) \quad (3)$$

where SA_{spa} stands for the spatial attention operation, $\mathbf{F}^* \in \mathbb{R}^{NT \times HW \times C}$ is reshaped from \mathbf{F} and $\gamma \in [0, 1]$ is a hyperparameter.

Channel Temporal Relation Modeling (CTRM). We first reshape \mathbf{F} as $\mathbf{F}_{seq2} \in \mathbb{R}^{NHW \times C \times T}$. Then it is fed into a learnable channel shift operation to obtain the temporal channel relation modeling feature \mathbf{F}_{tc} . Concretely, the learnable channel shift operation is a 1D channel-wise temporal convolution adopted to learn independent kernels for

each channel. Formally, the learnable channel shift operation can be formulated as:

$$\mathbf{F}_{tc}^{t,c} = \sum_i \mathbf{K}_{c,i} \mathbf{F}_{seq2}^{c,t+i} \quad (4)$$

where t and c denote the temporal and channel dimensions of the feature map, respectively. $\mathbf{K}_{c,i}$ indicates the temporal kernel weights of the c -th channel, $\mathbf{F}_{seq2}^{c,t+i} \in \mathbf{F}_{seq2}$ is the input c -th channel feature and $\mathbf{F}_{tc}^{t,c} \in \mathbf{F}_{tc}$ is the output c -th channel feature. After that, the final temporal channel relation modeling feature \mathbf{F}_{tc} is obtained through a spatial attention and we do the weight summation between \mathbf{F}_{tp} and \mathbf{F}_{tc} to obtain the final enhanced temporal features $\tilde{\mathbf{F}}$ as follows:

$$\tilde{\mathbf{F}} = \beta \mathbf{F}_{tp} + (1 - \beta) \mathbf{F}_{tc} \quad (5)$$

where $\beta \in [0, 1]$ is a hyperparameter.

In summary, PTRM aggregates temporal information for parts of patches while CTRM learns the temporal shift of channels. As a result, our LDTM could achieve sufficient temporal relation modeling in both the spatial and channel dimensions in a dense and learnable way.

3.4. Graph-guided Prototype Construction(GgPC)

We design a graph-guided prototype construction module to enhance the priori knowledge of the unknown video and explicitly optimize the intra- and inter-class correlation within video features. We draw inspiration from few-shot image classification methods based on graph neural networks [12, 15, 22, 6], which utilize graph networks to optimize intra-cluster similarity and inter-cluster dissimilarity and transform the image classification problems into node or edge classification problems. Different from this, directly feeding the video features (usually after the temporal pooling operation) into the graph network can lead to unsatisfactory results due to the loss of temporal information. Therefore, we only use graph networks as guidance to optimize features' intra- and inter-class correlation.

The overall framework of the proposed graph-guided prototype construction module is shown in Fig.4, and the overall algorithm is summarized in Algorithm.1. For simplicity and convenience, we discuss the process of the N_S -way 1-shot problem and consider that the query set \mathcal{Q} contains N_Q videos. This process can be divided into two stages: Graph neural network (GNN) propagation and task-oriented features obtaining. For GNN propagation, the temporally enhanced features $\tilde{\mathbf{F}}$ after doing the Mean Pooling operation in the temporal dimension \mathbf{F}^{avg} are used as node features \mathbf{V} for graph network initialization. Edge features \mathbf{A} represent the relationship between two nodes, i.e., the strength of intra- and inter-class relationships, and their initialization depends on the labels. The propagation includes the node aggregation and edge aggregation process. After

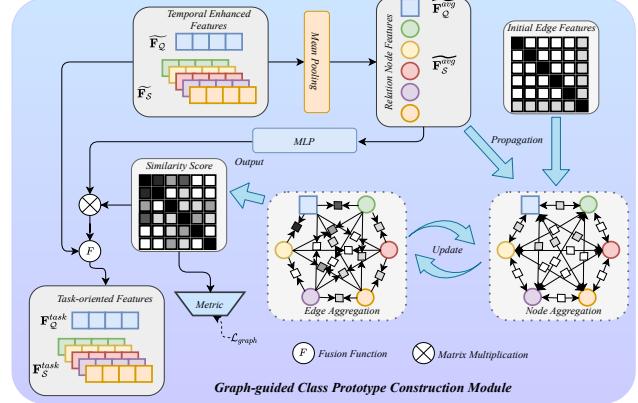


Figure 4. The overall framework of the proposed graph-guided prototype construction model. Consider that the query set \mathcal{Q} contains one video for simplicity and convenience.

completing the graph propagation, we use a Select operation to extract the similarity score from the updated edge features in the last layer. Select means that the edge features related to each query video feature are selected from the output entire edge features, and a total of N_Q new edge features are formed further. For task-oriented features obtaining, the details are shown in Algorithm.1 where f_{FNN} is a feed-forward network, f_{emb} and f_{fuse} are MLPs, and \otimes indicates the matrix multiplication. Meanwhile, the Select process is summarized in Algorithm.2. For K -shot ($K > 1$) tasks, when constructing node features, we perform mean pooling on the features of support videos of the same category in the feature dimension, while keeping other aspects consistent with the 1-shot task.

To sum up, the task-oriented features \mathbf{F}^{task} are obtained by fusing enhanced temporal features $\tilde{\mathbf{F}}$ with features \mathbf{F}^{graph} guided by graph networks to preserve the temporality of features. Through the guidance of GNN, every query video feature has its special support features, and the class correlation within video features is optimized explicitly.

3.5. Hybrid Prototype Matching Strategy (HPM)

Frame-level matching uses single individual frames, while tuple-level matching combines several frames into a tuple as the matching unit. HyRSM [35] applies the Hausdorff Distance metric as the prototype matching method, which can alleviate the strictly ordered constraints of acquiring better query-support correspondences, but it fails to capture temporal order. This matching metric is easily confused for actions with similar action scenes but strongly depends on temporal order, e.g., *pick up a glass of water* and *put down a glass of water*. To solve this problem, we design a hybrid prototype matching strategy that combines frame-level and tuple-level matching based on the bidirectional Hausdorff Distance. This approach effectively copes with video tasks of diverse styles. Given the task-oriented features $\mathbf{F}_S^{task}, \mathbf{F}_Q^{task}$, the m -th support video

Algorithm 1: The process of graph-guided prototype construction(GgPC)

```

1 Us indicates the unsqueeze operation, R indicates
    the repeat operation.
2 Input:  $\widetilde{\mathbf{F}}_S \in \mathbb{R}^{N_S \times T \times C}$ ,  $\widetilde{\mathbf{F}}_Q \in \mathbb{R}^{N_Q \times T \times C}$ ,
     $\widetilde{\mathbf{F}} = \widetilde{\mathbf{F}}_S \cup \widetilde{\mathbf{F}}_Q = \widetilde{\mathbf{F}} \in \mathbb{R}^{(N_S+N_Q) \times T \times C}$ 
3 Output:
     $\mathbf{F}_Q^{task} \in \mathbb{R}^{N_Q \times T \times C}$ ,  $\mathbf{F}_S^{task} \in \mathbb{R}^{N_Q \times N_S \times T \times C}$ 
4 Initialize:  $\widetilde{\mathbf{F}}^{avg} = \text{Mean\_pool}(\widetilde{\mathbf{F}}, dim = 1)$ 
    /* GNN Propagation */
5 Graph:  $\mathbf{G} = (\mathbf{V}, \mathbf{A}; \mathcal{S} \cup \mathcal{Q})$ ,  $\mathbf{v}_i^0 = \widetilde{\mathbf{F}}_i^{avg}$ ,  $\mathbf{a}_{ij}^0$ 
     $\forall i, j \in \mathcal{S} \cup \mathcal{Q}$ 
6 for  $l = 1, \dots, L$  do
7 for  $i = 1, \dots, |\mathcal{V}|$  do
8 |  $\mathbf{v}_i^l = \text{NodeAggregation}(\mathbf{v}_j^{l-1}, \mathbf{a}_{ij}^{l-1})$ 
9 end
10 for  $(i, j) = 1, \dots, |\mathcal{A}|$  do
11 |  $\mathbf{a}_{ij}^l = \text{EdgeAggregation}(\mathbf{v}_j^l, \mathbf{a}_{ij}^{l-1})$ 
12 end
13 end
14 Similarity Score:
     $\mathbf{M}_{siam} = \text{Select}(\mathbf{a}_{ij}^L[0]) \in \mathbb{R}^{N_Q \times (N_S+1) \times (N_S+1)}$ 
    /* Get Task-Oriented Features */
15 Optimized Features:
     $\mathbf{F}_S^{node} = \widetilde{\mathbf{F}}_S^{avg}.\text{Us}(0).\mathbf{R}(N_Q, 1, 1)$ 
     $\mathbf{F}^{node} = \text{Cat}([\mathbf{F}_S^{node}, \widetilde{\mathbf{F}}_Q^{avg}.\text{Us}(1)], dim = 1)$ 
     $\mathbf{F}^{graph} = f_{FFN}(\mathbf{M}_{siam} \otimes f_{emb}(\mathbf{F}^{node}))$ 
     $\mathbf{F}_S^{graph} = \mathbf{F}^{graph}[:, :, N_S, :].\text{Us}(1).\mathbf{R}(1, T, 1, 1)$ 
     $\mathbf{F}_Q^{graph} = \mathbf{F}^{graph}[:, N_S :, :].\text{Us}(1).\mathbf{R}(1, T, 1)$ 
16 Task-oriented Features:
     $\mathbf{F}_S^{hid} = \widetilde{\mathbf{F}}_S.\text{Us}(0).\mathbf{R}(N_Q, 1, 1, 1)$ 
     $\mathbf{F}_S^{task} = f_{fuse}(\text{Cat}([\mathbf{F}_S^{hid}, \mathbf{F}_S^{graph}], dim = 2))$ 
     $\mathbf{F}_Q^{task} = f_{fuse}(\text{Cat}([\mathbf{F}_Q, \mathbf{F}_Q^{graph}], dim = 2))$ 

```

feature in the k class and the p -th query video feature indicates $\mathbf{s}_m^k \in \mathbb{R}^{T \times C}$, $\mathbf{q}_p \in \mathbb{R}^{T \times C}$, respectively. For single-frame matching, we apply a bidirectional Mean Hausdorff metric as follow:

$$\mathcal{D}_{frame} = \frac{1}{T} \left[\sum_{\mathbf{s}_{m,i}^k \in \mathbf{s}_m^k} \left(\min_{\mathbf{q}_{p,j} \in \mathbf{q}_p} \|\mathbf{s}_{m,i}^k - \mathbf{q}_{p,j}\| \right) + \sum_{\mathbf{q}_{p,j} \in \mathbf{q}_p} \left(\min_{\mathbf{s}_{m,i}^k \in \mathbf{s}_m^k} \|\mathbf{q}_{p,j} - \mathbf{s}_{m,i}^k\| \right) \right] \quad (6)$$

where $\mathbf{s}_{m,i}^k$ represents the i -th frame feature of \mathbf{s}_m^k , $\mathbf{q}_{p,j}$ indicates the j -th frame feature of \mathbf{q}_p , and they have a total T frames. For tuple-level prototype matching, we combine two frames into one tuple and iterate through all combina-

Algorithm 2: The process of Select operation

```

1 Input:  $\mathbf{a}_{ij}^L[0] \in \mathbb{R}^{(N_S+N_Q) \times (N_S+N_Q)}$ 
2 Output:  $\mathbf{M}_{siam} \in \mathbb{R}^{N_Q \times (N_S+1) \times (N_S+1)}$ 
3 Similarity Score:  $\mathbf{M}_{siam} = \text{List}()$ 
4 for  $n_Q = 1, \dots, N_Q$  do
5 |  $\mathbf{m}_{siam} = \text{Zeros}((N_S + 1) \times (N_S + 1))$ 
6 |  $\mathbf{m}_{siam}[:, N_S, : N_S] = \mathbf{a}_{ij}^L[0][: N_S, : N_S]$ 
7 |  $\mathbf{m}_{siam}[:, N_S, -1] = \mathbf{a}_{ij}^L[0][: N_S, N_S + n_Q]$ 
8 |  $\mathbf{m}_{siam}[-1, :, N_S] = \mathbf{a}_{ij}^L[0][N_S + n_Q, :, N_S]$ 
9 |  $\mathbf{m}_{siam}[-1, -1] = \mathbf{a}_{ij}^L[0][N_S + n_Q, N_S + n_Q]$ 
10 |  $\mathbf{M}_{siam}.\text{Append}(\mathbf{m}_{siam})$ 
11 end
12  $\mathbf{M}_{siam} = \text{Stack}(\mathbf{M}_{siam})$ 

```

tions to get $L = \frac{1}{2}(T-1)T$ tuples for T frames, given by:

$$\begin{aligned} \mathbf{ts}_{m,i}^k &= [\mathbf{s}_{m,i_1}^k + \text{PE}(i_1), \mathbf{s}_{m,i_2}^k + \text{PE}(i_2)] \quad 1 \leq i_1 \leq i_2 \leq T \\ \mathbf{tq}_{p,j} &= [\mathbf{q}_{p,j_1} + \text{PE}(j_1), \mathbf{q}_{p,j_2} + \text{PE}(j_2)] \quad 1 \leq j_1 \leq j_2 \leq T \end{aligned} \quad (7)$$

where $\mathbf{ts}_{m,i}^k, \mathbf{tq}_{p,j} \in \mathbb{R}^{2C}$, and each tuple follows the temporal information of the original frame. To this end, the Mean Hausdorff metric based on tuples can be formulated as:

$$\begin{aligned} \mathcal{D}_{tuple} &= \frac{1}{L} \left[\sum_{\mathbf{ts}_{m,i}^k \in \mathbf{ts}_m^k} \left(\min_{\mathbf{tq}_{p,j} \in \mathbf{tq}_p} \|\mathbf{ts}_{m,i}^k - \mathbf{tq}_{p,j}\| \right) \right. \\ &\quad \left. + \sum_{\mathbf{tq}_{p,j} \in \mathbf{tq}_p} \left(\min_{\mathbf{ts}_{m,i}^k \in \mathbf{ts}_m^k} \|\mathbf{tq}_{p,j} - \mathbf{ts}_{m,i}^k\| \right) \right] \end{aligned} \quad (8)$$

Finally, the hybrid matching metric can be formulated as:

$$\mathcal{D}_{hybrid} = \alpha \mathcal{D}_{tuple} + (1 - \alpha) \mathcal{D}_{frame} \quad (9)$$

where $\alpha \in [0, 1]$ is a hyperparameter.

In a word, our proposed hybrid prototype matching strategy combines the advantages of both frame- and tuple-level matching to cope with video tasks of multivariate styles well.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate the performance of our method on four few-shot datasets, including Kinetics [5], HMDB51 [16], UCF101 [29], and SSv2 [13]. For Kinetics and SSv2, we use the splits provided by [4] and [47], where 100 classes were selected and divided into 64/12/24 action classes as the meta-training/meta-validation/meta-testing set. Additionally, for UCF101 and HMDB51, we

Methods	Reference	Backbone	HMDB51		UCF101		SSv2		Kinetics	
			1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
MatchingNet [31]	NeurIPS(16)	ResNet-50	-	-	31.3	45.5	-	-	53.3	74.6
MAML [10]	ICML(17)	ResNet-50	-	-	30.9	41.9	-	-	54.2	75.3
ProtoNet [28]	NeurIPS(17)	C3D	54.2	68.4	74.0	89.6	33.6	43.0	64.5	77.9
TRN++ [45]	ECCV(18)	ResNet-50	-	-	-	-	38.6	48.9	68.4	82.0
CMN++ [46]	ECCV(18)	ResNet-50	-	-	-	34.4	43.8	-	57.3	76.0
TARN [3]	BMVC(19)	C3D	-	-	-	-	-	-	64.8	78.5
ARN [41]	ECCV(20)	C3D	45.5	60.6	66.3	83.1	-	-	63.7	82.4
OTAM [4]	CVPR(20)	ResNet-50	54.5	68.0	79.9	88.9	42.8	52.3	73.0	85.8
TTAN [19]	ArXiv(21)	ResNet-50	57.1	74.0	80.9	93.2	46.3	60.4	-	-
ITANet [42]	IJCAI(21)	ResNet-50	-	-	-	-	49.2	62.3	73.6	84.3
TRX [27]	CVPR(21)	ResNet-50	54.9*	75.6	81.0*	96.1	42.0	64.6	65.1*	85.9
TA2N [20]	AAAI(22)	ResNet-50	59.7	73.9	81.9	95.1	47.6	61.0	72.8	85.8
STRM [30]	CVPR(22)	ResNet-50	57.6*	<u>77.3</u>	82.7*	96.9	43.5*	66.0*	65.1*	86.7
MTFAN [36]	CVPR(22)	ResNet-50	59.0	74.6	84.8	95.1	45.7	60.4	<u>74.6</u>	87.4
HyRSM [35]	CVPR(22)	ResNet-50	<u>60.3</u>	76.0	83.9	94.7	<u>51.5*</u>	67.5*	73.7	86.1
HCL [44]	ECCV(22)	ResNet-50	59.1	76.3	82.5	93.9	47.3	64.9	73.7	85.8
Huang <i>et al.</i> [14]	ECCV(22)	ResNet-50	60.1	77.0	71.4	91.0	49.3	66.7	73.3	86.4
Nguyen <i>et al.</i> [23]	ECCV(22)	ResNet-50	59.6	76.9	84.9	95.9	43.8	61.7	74.3	87.4
SloshNet [38]	AAAI(23)	ResNet-50	59.4	77.5	86.0	97.1	46.5	<u>68.3</u>	70.4	<u>87.0</u>
GgHM	-	ResNet-50	61.2	76.9	85.2	96.3	54.5	69.2	74.9	87.4

Table 1. State-of-the-art comparison on the 5-way k-shot benchmarks of HMDB51, UCF101, SSv2, Kinetics. The **boldfaced** and underline font indicate the highest and the second highest results. Note: * means our implementation.

evaluate our method on the splits provided by [41]. **Network Architectures.** We utilize the ResNet-50 as the feature extractor with ImageNet pre-trained weights [7]. For LDTM, $\mathbf{W}_{t1}, \mathbf{W}_{t2}$ are two one-layer MLPs, and gap is set to 2. For GgPC, we apply one-layer GNN to obtain task-oriented features. More implementation details can be found in the appendix.

Training and Inference. Followed by TSN [33], we uniformly sample 8 frames ($T=8$) of a video as the input augmented with some basic methods, e.g. random horizontal flipping, cropping, and color jit in training, while multi-crops and multi-views in inference. For training, SSv2 were randomly sampled 100,000 training episodes, and the other datasets were randomly sampled 10,000 training episodes. Moreover, we used the Adam optimizer with the multi-step scheduler for our framework. For inference, we reported the average results over 10,000 tasks randomly selected from the test sets in all datasets.

4.2. Results

As shown in Tab.1, our method **GgHM** achieves impressive results against the state-of-the-art methods in all datasets and few-shot settings. Our method especially achieves new state-of-the-art performance on Kinetics and SSv2 in all few-shot settings and HMDB in the 5-way 1-shot task, respectively. In other tasks, our method either achieves the second-highest result or achieves results that are very close to the SOTA. Our method per-

forms impressively without any preference for datasets or the few-shot settings. In contrast, some methods perform unsatisfactorily in the 1-shot task (e.g., TRX [27], STRM [30], SloshNet [38]) or particular datasets (e.g., Nguyen *et al.* [23] on SSv2, MTFAN [36] on SSv2, Huang *et al.* [14] on UCF101). In addition, compared to our baseline HyRSM [35], which also utilizes the Hausdorff Distance metric as the class prototype matching strategy and focuses on building the task-oriented feature, the effect of our method is significantly improved. Specifically, compared to HyRSM, our method brings 0.9%, 1.3%, 3.0%, and 0.3% performance improvements in the 1-shot task of HMDB51, UCF101, SSv2, and Kinetics, respectively. In the 5-shot task, our method outperforms HyRSM significantly, bringing 0.3%, 1.6%, 2.7%, and 0.7% gain on HMDB51, UCF101, SSv2, and Kinetics, respectively.

4.3. Ablation Study

Impact of the proposed components. To validate the contributions of each module (i.e. LDTM, GgPC, HPM) in our method, we experiment under 5-way 1-shot and 5-way 5-shot settings on the SSv2 dataset. Our baseline method only utilizes the frame-level bidirectional Mean Hausdorff metric as the prototype matching strategy without any extra modules. As shown in Tab. 2, we observe that each component is effective. Specifically, compared to the baseline, the HPM module can bring 0.6% and 0.7% accuracy improvement on 1-shot and 5-shot tasks, the GgPC module

LDTM	GgPC	HPM	1-shot	5-shot
✗	✗	✗	44.6	56.0
✗	✗	✓	45.2	56.7
✗	✓	✗	49.0	61.5
✓	✗	✗	51.8	64.9
✗	✓	✓	50.1	63.4
✓	✗	✓	52.2	65.8
✓	✓	✗	53.9	68.7
✓	✓	✓	54.5	69.2

Table 2. The impact of proposed modules on SSV2 in the 5-way 1-shot and 5-way 5-shot settings.

can bring 4.4% and 5.5% performance improvement on two tasks, and the LDTM module can bring 7.2% and 8.9% performance gain on two tasks. Additionally, stacking modules can enhance performance, indicating the complementarity between components. Combining all modules can get the best results, bringing 9.9% and 13.2% performance improvement on 1-shot and 5-shot tasks over the baseline.

Impact of temporal modeling integration. To explore the impact of each temporal modeling module in LDTM and demonstrate their effectiveness, we experiment on the 5-way 1-shot and 5-way 5-shot tasks of SSV2 to ablate our proposed temporal relation modeling blocks. The PTRM block includes spatial attention, which indicates doing Self-Attention only on the spatial dimension. As shown in Tab.3, the CTRM block brings about a 1.0% and 1.9% accuracy improvement on the 1-shot and 5-shot tasks over the baseline. Moreover, the PTRM block obtains 1.5% and 2.3% gain on the 1-shot and 5-shot tasks over the baseline. The integration of these two blocks results in 2.9% and 3.7% gain on two tasks, respectively.

Analysis of building the task-oriented features. To demonstrate the necessity of constructing task-specific features and compare the efficacy of various methods for constructing them, we conduct experiments on the 5-way 1-shot task of Kinetics and SSV2. Building task-oriented features can be divided into two categories: unsupervised and supervised. The critical difference between them is whether label information is used directly to constrain the construction of features. The Self-Attention method(HyRSM [35]) means that the task features (the set of support and query video features) do self-attention without using the label information to supervise. In contrast, our GNN method directly applies label information to do supervision, which

Spatial Attention	PTRM	CTRM	1-shot	5-shot
✓	✗	✗	51.6	65.5
✓	✗	✓	52.6	67.4
✗	✓	✗	53.1	67.8
✗	✓	✓	54.5	69.2

Table 3. The impact of temporal modeling blocks integration on SSV2 in the 5-way 1-shot and 5-way 5-shot settings.

Method	Type	Kinetics	SSV2
None	-	72.9	52.2
Self-Attention	unsupervised	74.1	53.7
GNN	supervised	74.6	54.0
GNN(Transduction)	supervised	74.9	54.5

Table 4. Analysis of building the task-oriented features on Kinetics and SSV2 in the 5-way 1-shot setting.

Metric	Kinetics	SSV2
Frame-level matching	74.3	53.9
Tuple-level matching	74.1	54.2
Hybrid matching	74.9	54.5

Table 5. Comparisons of different prototype matching strategies on Kinetics and SSV2 in the 5-way 1-shot setting.

Param α	0	0.2	0.4	0.6	0.8	1.0
Kinetics	74.3	74.6	74.9	74.5	74.3	74.1
SSV2	53.9	54.1	54.2	54.5	54.3	54.2

Table 6. The impact of the varying fusion parameter α of hybrid prototype matching on Kinetics and SSV2 in the 5-way 1-shot setting.

can explicitly optimize the video features' intra- and inter-class correlation. As shown in Tab.4, the Self-Attention method can bring 1.2% and 1.5% gain on Kinetics and SSV2 over the baseline each, which can demonstrate the necessity of building task-oriented features. Moreover, our GNN method(each query feature owns a graph) can bring 1.7% and 1.8% gain over the baseline on two datasets, respectively, showing the advantage of the supervised method. Moreover, our GNN method with transduction(all query features in the same graph) brings a 2.0% and 2.3% accuracy improvement on two datasets.

Comparisons of different prototype matching strategies. To analyze different prototype matching strategies, we experiment on the 5-way 1-shot task of Kinetics and SSV2 with different prototype matching methods to evaluate the

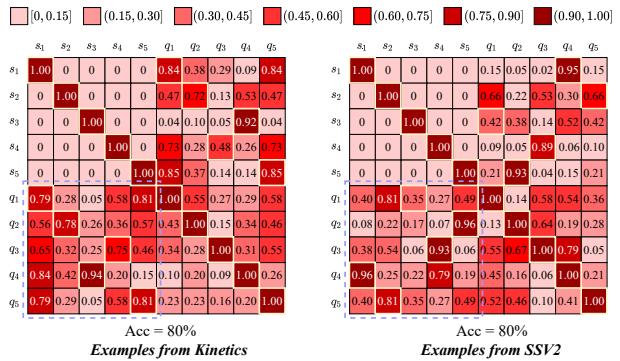


Figure 5. Visualization of the updated edge features output by the GNN. GT stands for the ground truth and ACA represents the accuracy calculation area. A higher score indicates a greater degree of similarity. We can use the features in the accuracy calculation area directly to obtain task recognition results.

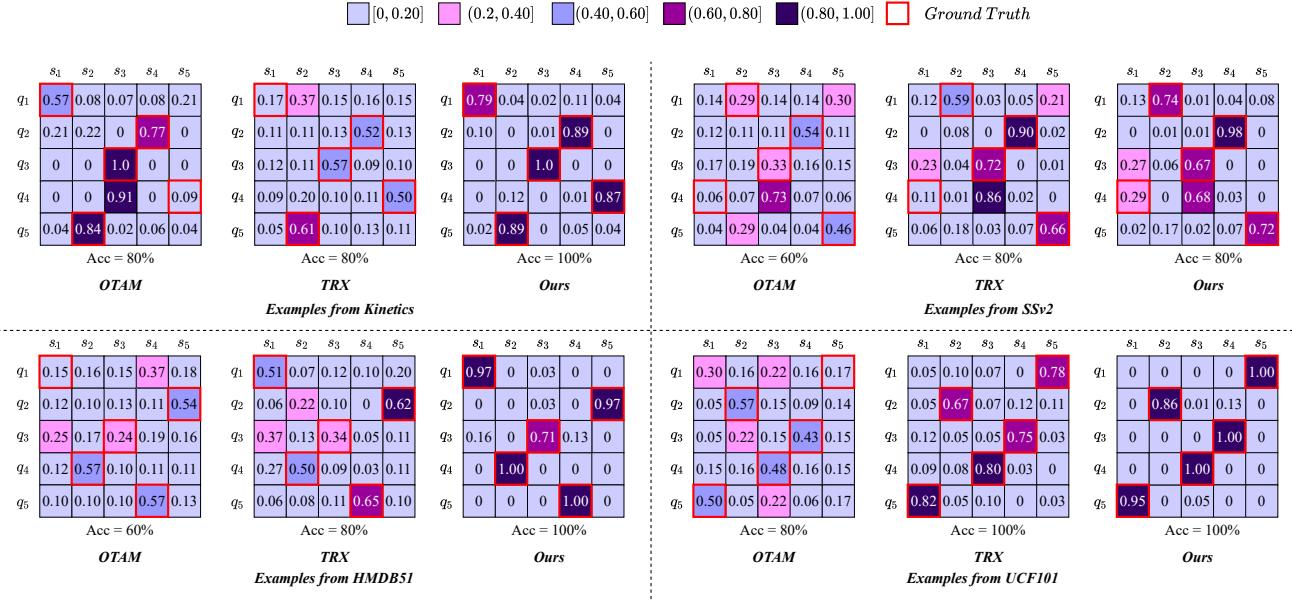


Figure 6. Similarity visualization between query and support videos with different methods on the 5-way 1-shot task of Kinetics, SSv2, HMDB51, and UCF101. A higher score indicates a greater degree of similarity.

effectiveness of our hybrid matching strategy. All the methods are based on the bidirectional Mean Hausdorff metric and the experiment results are shown in Tab.5. Our hybrid matching strategy brings a 0.6% and 0.6% accuracy improvement on two datasets over the frame-level matching strategy. Meanwhile, it obtains 0.8% and 0.3% gain on two datasets over the tuple-level matching strategy, respectively.

Impact of the varying fusion parameter of hybrid prototype matching. Tab.6 shows the impact of the varying fusion parameter α in hybrid prototype matching. As part of our experiments, we perform the 5-way 1-shot task on Kinetics and SSv2. The parameter α denotes the weight assigned to the frame- and tuple-level matching in the final fusion. From the results, the optimal values of parameter α are 0.4 for Kinetics and 0.6 for SSv2.

Visualization of the update edge features output by GNN. As shown in Fig.5, we visualize two examples of the updated edge features output by the GNN and the ground truth on Kinetics and SSv2 in the 5-way 1-shot setting. The edge features' value can be seen as the similarity score between two video features. From the visualization, GNN as guidance can well optimize video features' inter- and intra-class correlation, in which updated edge features are very close to the similarity matrix corresponding to the ground truth. Meanwhile, the intermediate output recognition results of GNN obtained by the edge features in the accuracy calculation area can also achieve high accuracy.

Similarity visualization. Fig.6 visualizes the predicted similarities between query and support videos with different methods on the 5-way 1-shot task of Kinetics, SSv2,

HMDB51, and UCF101. Our method achieves more discriminative results for similar videos in each task compared to OTAM [4] and TRX [27]. The results presented here demonstrate the effectiveness of our method in distinguishing videos from similar categories, as it has significantly improved both the prediction accuracy and intra-/inter-class correlation within video features.

5. Conclusion

In this work, we have presented a novel few-shot action recognition framework, GgHM, leading to impressive performance in recognizing similar categories in every task without any datasets or task preference. Concretely, we learn task-oriented features by the guidance of a graph neural network during class prototype construction, optimizing the intra- and inter-class feature correlation explicitly. Next, we propose a hybrid class prototype matching strategy that leverages both frame- and tuple-level prototype matching to effectively handle video tasks with diverse styles. Besides, we propose a dense temporal modeling module consisting of a temporal patch and temporal channel relation modeling block to enhance the video feature temporal representation, which helps to build a more solid foundation for the matching process. GgHM shows consistent improvements over other challenging baselines on several few-shot datasets, demonstrating the effectiveness of our method.

Acknowledgement

This work is partly supported by the following grant: Key R&D Program of Zhejiang (No.2022C03126).

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.
- [2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.
- [3] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019.
- [4] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020.
- [5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [6] Chaofan Chen, Xiaoshan Yang, Changsheng Xu, Xuhui Huang, and Zhe Ma. Eckpn: Explicit class knowledge propagation network for transductive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6596–6605, 2021.
- [7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [8] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. *Advances in Neural Information Processing Systems*, 33:21981–21993, 2020.
- [9] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.
- [10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.
- [11] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. Depth guided adaptive meta-fusion network for few-shot video recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1142–1151, 2020.
- [12] Victor Garcia and Joan Bruna. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*, 2017.
- [13] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The “something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.
- [14] Yifei Huang, Lijin Yang, and Yoichi Sato. Compound prototype matching for few-shot action recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 351–368. Springer, 2022.
- [15] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11–20, 2019.
- [16] Hildegard Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.
- [17] Changzhen Li, Jie Zhang, Shuzhe Wu, Xin Jin, and Shiguang Shan. Hierarchical compositional representations for few-shot action recognition. *arXiv preprint arXiv:2208.09424*, 2022.
- [18] Hongyang Li, David Eigen, Samuel Dodge, Matthew Zeiler, and Xiaogang Wang. Finding task-relevant features for few-shot learning by category traversal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1–10, 2019.
- [19] Shuyuan Li, Huabin Liu, Rui Qian, Yuxi Li, John See, Mengjuan Fei, Xiaoyuan Yu, and Weiyao Lin. Ttan: Two-stage temporal alignment network for few-shot action recognition. *arXiv preprint*, 2021.
- [20] Shuyuan Li, Huabin Liu, Rui Qian, Yuxi Li, John See, Mengjuan Fei, Xiaoyuan Yu, and Weiyao Lin. Ta2n: Two-stage action alignment network for few-shot action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1404–1411, 2022.
- [21] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Metasgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*, 2017.
- [22] Yuqing Ma, Shihao Bai, Shan An, Wei Liu, Aishan Liu, Xiantong Zhen, and Xianglong Liu. Transductive relation-propagation network for few-shot learning. In *IJCAI*, volume 20, pages 804–810, 2020.
- [23] Khoi D Nguyen, Quoc-Huy Tran, Khoi Nguyen, Binh-Son Hua, and Rang Nguyen. Inductive and transductive few-shot video classification via appearance and temporal alignments. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XX*, pages 471–487. Springer, 2022.
- [24] Xinze Ni, Hao Wen, Yong Liu, Yatai Ji, and Yujiu Yang. Multimodal prototype-enhanced network for few-shot action recognition. *arXiv preprint arXiv:2212.04873*, 2022.
- [25] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018.
- [26] Jay Patravali, Gaurav Mittal, Ye Yu, Fuxin Li, and Mei Chen. Unsupervised few-shot action recognition via action-appearance aligned meta-adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8484–8494, 2021.

- [27] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 475–484, 2021.
- [28] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.
- [29] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
- [30] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19958–19967, 2022.
- [31] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.
- [32] Jiaxing Wang, Jiaxiang Wu, Haoli Bai, and Jian Cheng. Mnas: Meta neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6186–6193, 2020.
- [33] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.
- [34] Xiao Wang, Weirong Ye, Zhongang Qi, Xun Zhao, Guangge Wang, Ying Shan, and Hanzi Wang. Semantic-guided relation propagation network for few-shot action recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 816–825, 2021.
- [35] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19948–19957, 2022.
- [36] Jiamin Wu, Tianzhu Zhang, Zhe Zhang, Feng Wu, and Yongdong Zhang. Motion-modulated temporal fragment alignment network for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9160, 2022.
- [37] Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Xian-Sheng Hua, and Lei Zhang. Spatiotemporal self-attention modeling with temporal patch shift for action recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 627–644. Springer, 2022.
- [38] Jiaheng Xing, Mengmeng Wang, Boyu Mu, and Yong Liu. Revisiting the spatial and temporal modeling for few-shot action recognition. *arXiv preprint arXiv:2301.07944*, 2023.
- [39] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8808–8817, 2020.
- [40] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *International conference on machine learning*, pages 7115–7123. PMLR, 2019.
- [41] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *European Conference on Computer Vision*, pages 525–542. Springer, 2020.
- [42] Songyang Zhang, Jiale Zhou, and Xuming He. Learning implicit temporal alignment for few-shot video classification. *arXiv preprint arXiv:2105.04823*, 2021.
- [43] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13577–13587, 2021.
- [44] Sipeng Zheng, Shizhe Chen, and Qin Jin. Few-shot action recognition with hierarchical matching and contrastive learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 297–313. Springer, 2022.
- [45] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018.
- [46] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018.
- [47] Linchao Zhu and Yi Yang. Label independent memory for semi-supervised few-shot video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):273–285, 2020.