CROSS-DOMAIN ACTION RECOGNITION VIA PROTOTYPICAL GRAPH ALIGNMENT

Anonymous ICME submission

ABSTRACT

Compared with the well-explored cross-domain image recognition, cross-domain action recognition is a more challenging task because not only spatial but also temporal domain gaps exist across domains. Previous works attempt to bridge the temporal domain gap by aligning the domain-related key segments of videos from source and target domains. However, such practice overlooks the heterogeneous temporal domain gaps among different categories and presents temporal alignment strategies in a class-irrelevant manner. To address this issue, we propose to achieve class-wise temporal alignment for cross-domain action recognition via prototypical graph alignment (PGA). Concretely, we generate segment-level prototypes for the classes of both domains to capture per-class temporal dynamics. Furthermore, intra-domain and inter-domain prototypical graphs are established to mine the temporal relationships between each input video and its corresponding intra-domain and inter-domain prototypes. In this way, a discriminative and domain adaptive video representation is obtained by holistically reasoning cross-domain temporal dynamics. To class-wisely align the cross-domain video representations, each action category is equipped with a customized class-specific domain discriminator for temporal alignment with adversarial learning. Extensive experiments on three benchmarks show that PGA yeilds state-of-the-art performance on the task of cross-domain action recognition.

Index Terms**—** Cross-domain action recognition, Prototypical graph alignment, Adversarial Learning

1. INTRODUCTION

Cross-domain learning [1, 2, 3, 4], also known as domain adaptation, aims to mitigate the problem of domain shift by learning domain-invariant features or aligning distributions across domains. Thanks to the development of convolutional neural networks and adversarial learning, cross-domain image recognition [1, 2, 3, 4] has already witness appealing performance. By contrast, cross-domain action recognition [5, 6, 7, 8] is a more challenging task because cross-domain videos differ in both spatial appearance and temporal dynamics. To be specific, besides the appearance and spatial contexts, the start time and duration of an action may be different in videos of source and target domains. Therefore, temporal alignment is essential for diminishing the temporal domain gap between source and target videos.

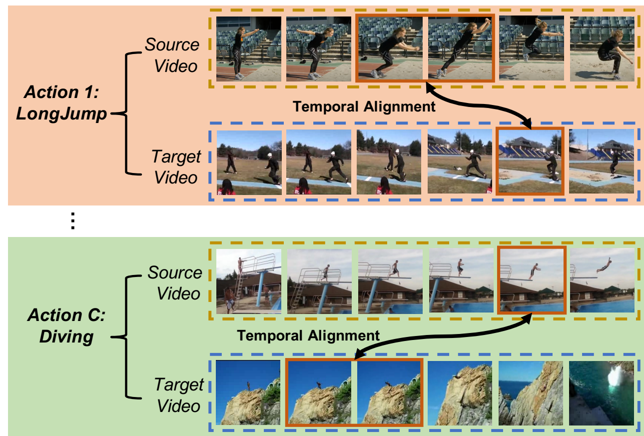

Fig. 1. Illustration of class-wise temporal alignment for cross-domain action recognition. Different classes of actions have distinct key frames / segments, which are vital for recognition. Heterogeneous temporal domain gaps exist among these classes. Thus, customized cross-domain alignment strategy is essential for the temporal dynamics of each category. Key segments are outlined with red boxes. Best viewed in color.

Prior works on cross-domain action recognition (video domain adaptation) enable their models to attend to the domain-related key segments that are crucial for recognizing the actions in source and target videos via variants of temporal attention mechanisms. Moreover, adversarial learning is used to bridge the spatial and temporal domain gaps based on the attended important segments. These solutions implicitly assume that different classes of videos in a domain have similar temporal dynamics, simply following the spatial domain consistency assumption in cross-domain image recognition: all classes of images in a domain have similar spatial contexts and styles. Unfortunately, such temporal domain consistency assumption does not hold for action recognition because classes of videos in a domain may have distinct key frames / segments although they perhaps share similar backgrounds. As shown in Fig. 1, the key action stages of *LongJump* and *Diving* are distinct and the temporal domain shifts of these two actions are different. If a class-irrelevant temporal alignment strategy is employed, the heterogeneous temporal domain gaps among classes would be overlooked and the alignment of these key segments can not be well handled.

To address the aforementioned issue, we propose to achieve class-wise temporal alignment in a fine-grained manner for cross-domain action recognition via prototypical graph alignment (PGA). To be specific, we extract the semantics of segments of each action category by learning their corresponding high-level prototypical representations. The prototypes are learnable and updated as the moving average of the segment-level features during the training stage. In this way, the temporal dynamics of each class can be simply represented by such segment-level prototypes. Going one step further, the intra-domain and inter-domain prototypical graphs are established to exploit the class-aware temporal relationships between the segment-level features of the input video and intra-domain / inter-domain prototypes. Graph convolutional networks are introduced to obtain discriminative and domain adaptive video representations by message propagation among graph vertices. Besides, we equip a customized class-specific domain discriminator for each action category to realize class-wise temporal alignment in a manner of adversarial learning. To this end, PGA is able to fully exploit the cross-domain class-aware temporal relationships and perform a better domain adaptive training for action recognition.

To summarize, the major contributions of our work can be listed as follows:

- To the best of our knowledge, we make the first attempt to achieve class-wise temporal alignment for cross-domain action recognition via prototypical graph alignment (PGA).
- We propose a novel cross-domain temporal graph reasoning to explore the class-aware fine-grained temporal relationships among intra-domain and inter-domain classes of videos.
- We propose a class-wise alignment via adversarial training with customized domain discriminator for each category.
- Comprehensive experiments demonstrate the effectiveness of the proposed framework on three large-scale cross-domain action recognition benchmarks.

2. RELATED WORK

2.1. Action Recognition

Thanks to the rise of deep learning, vast progress has been made for action recognition over the last decades. Two-stream networks [9] and TSN [10] fuse the prediction from appearance and motion streams with RGB and optical flow features. C3D [11] and I3D [12] explore the 3D convolutional architectures for action recognition. TRN [13] learns multi-scale temporal relations for discovering knowledge over time. However, these works follow the standard supervised training paradigm and heavily rely on annotations, which brings new challenges of cross-domain action recognition.

2.2. Cross-domain Action Recognition

Cross-domain action recognition is a challenging problem because both spatial and temporal domain gaps should be eliminated. TA^{3}N [5] proposes to align the temporal dynamics of the videos with temporal relation and attention mechanism. TCoN [6] leverages cross-domain co-attention mechanism to align key segments. SAVA [8] introduces a self-supervision task to learn more robust features for foreground objects and align important segments. STCDA [7] employs spatial-temporal contrastive self-supervised learning to improve the generalization of video representation. ABG [14] jointly models source domain and target domain data as bipartite graphs and aligns features via conditional adversarial learning. Such existing methods follow main spirit of aligning the class-irrelevant domain-related key segments of source and target videos by adversarial learning and temporal aggregation. Nevertheless, the heterogeneous domain gaps among different action categories are overlooked. In this work, we explore to achieve class-wise temporal alignment for cross-domain videos via prototypical graph alignment (PGA).

3. METHODOLOGY

3.1. Problem Setup

Given a set of N^s labeled source domain videos $\mathcal{D}^s = \{(v_i^s, y_i^s)\}_{i=1}^{N^s}$ and another set of N^t unlabeled target domain videos $\mathcal{D}^t = \{v_i^t\}_{i=1}^{N^t}$ which share the same label space $\mathcal{C} = \{1, 2, ..., C\}$, the goal of cross-domain action recognition is to adapt the action recognition model trained on the source domain to the previously unseen target domain. Here, we represent a video v_i^s as a collection of m segment-level features, *i.e.*, $v_i^s = [v_{i,1}^s, v_{i,2}^s, ..., v_{i,m}^s]$, where $v_{i,j}^s$ denotes the feature of the j^{th} segment in the i^{th} source video.

3.2. Prototypical Graph Representation

Formally, we represent a graph as $G = (V, A, X)$, where V, A and X denote the set of vertices, the adjacency matrix and the features corresponding to the vertices in V, respectively.

In order to exploit the underlying temporal relationships among different classes of actions, we establish intra-domain and inter-domain prototypical graphs and obtain discriminative video representations by message propagation among vertices. Concretely, intra-domain and inter-domain graphs are constructed for each source and target input videos: $\mathcal{G}_{intra}^s = \{V_{intra}^s, A_{intra}^s, X_{intra}^s\}$, $\mathcal{G}_{intra}^t = \{V_{inra}^t, A_{intra}^t, X_{intra}^t\}$, $\mathcal{G}_{inter}^s = \{V_{intra}^s, A_{inter}^s, X_{inter}^s\}$ and $\mathcal{G}_{inter}^t = \{V_{inter}^t, A_{inter}^t, X_{inter}^t\}$. For simplicity, we take an input video of target domain v_i^t as an example and following elaborate the construction of its intra-domain and inter-domain prototypical graphs, *i.e.*, $\mathcal{G}_{intra}^t$ and $\mathcal{G}_{inter}^t$. The prototypical graphs $\mathcal{G}_{intra}^s$ and $\mathcal{G}_{inter}^s$ for the source video v_j^s can be defined analogously.

We first extract the semantics of the segments of each action category by learning their corresponding high-level prototypical representations. To be specific, prototypes $p_c^s = [p_{c,1}^s, p_{c,2}^s, ..., p_{c,m}^s]$ and $p_c^t = [p_{c,1}^t, p_{c,2}^t, ..., p_{c,m}^t]$ are used to

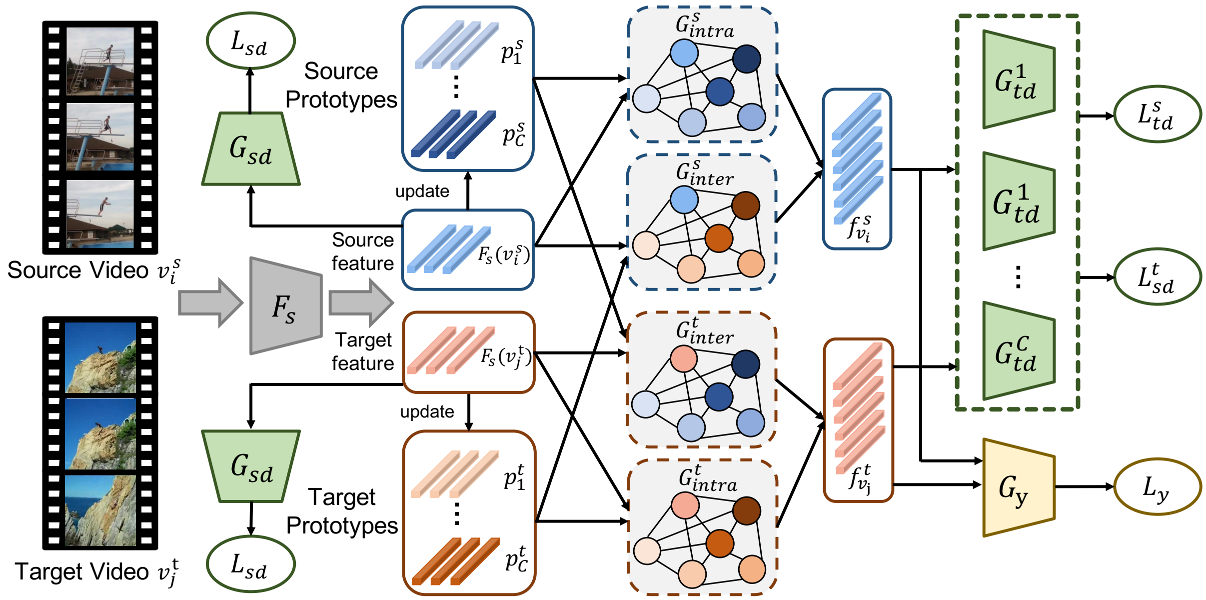

Fig. 2. Overview of our proposed PGA. For each source video v_i^s and target video v_j^t, their corresponding prototypical graphs $\mathcal{G}_{intra}^s$, $\mathcal{G}_{inter}^s$ and $\mathcal{G}_{intra}^t$, $\mathcal{G}_{inter}^t$ are established, respectively. Segment-level prototypes $\{p_c^s, p_c^t | c = 1, 2, ..., C\}$ are learned to represent the temporal dynamics of action categories in source and target domains. Prototypical graphs capture the class-aware temporal relationships among prototypes and segment-level video features. Spatial domain discriminator G_{sd} and multiple class-specific temporal domain discriminators $\{G_{td}^c | c = 1, 2, ..., C\}$ are introduced for spatial and class-wise temporal alignment via adversarial learning. Best viewed in color.

represent the semantics of action c in source and target domains, respectively, where $p_{c,k}^s$ and $p_{c,k}^t$ are the prototypes of the k^{th} segment. The prototype $p_{c,k}^s$ of source domain is computed as the average segment-level feature vector of all samples of the corresponding k^{th} segment and class c:

$$p_{c,k}^s = \frac{1}{N_c^s} \sum_{i=1}^{N_c^s} F_s(v_{i,k}^s), \tag{1}$$

where F_s is the feature extractor for segment-level features, N_c^s is the numbers of samples of class c in source and target domains. However, $p_{c,k}^t$ can not directly computed in a same way since the videos in target domain are unlabeled. Hence, we use the pseudo-labels given by the label classifier G_y for each video v_i^t, *i.e.*, $\{\hat{y}_{i,1}^t, \hat{y}_{i,2}^t, ..., \hat{y}_{i,C}^t\}$ to generate the prototypes $p_{c,k}^t$:

$$p_{c,k}^t = \frac{1}{N^t} \sum_{i=1}^{N^t} \hat{y}_{i,c}^t F_s(v_{i,k}^t), \tag{2}$$

where $\hat{y}_i^t = [\hat{y}_{i,1}^t, \hat{y}_{i,2}^t, ..., \hat{y}_{i,C}^t]$, $\hat{y}_{i,c}^t$ is the pseudo-label (softmax probability) of class c ($\sum_{c=1}^{C} \hat{y}_{i,c}^t = 1$). To avoid the computational-intensive update in the training stage, prototypes $p_{c,k}^s$ and $p_{c,k}^t$ evolve in the way of moving average:

$$\begin{aligned} p_{c,k}^s &\leftarrow (1-\eta) p_{c,k}^s + \eta \left[\frac{1}{N_c^s} \sum_{i=1}^{N_c^s} F_s(v_{i,k}^s) \right] \\ p_{c,k}^t &\leftarrow (1-\eta) p_{c,k}^t + \eta \left[\frac{1}{N^t} \sum_{i=1}^{N^t} \hat{y}_{i,k}^t F_s(v_{i,k}^t) \right], \end{aligned} \tag{3}$$

where η is the momentum hyper-parameter.

Based on the segment-level prototypes $\{p_c^s, p_c^t | c = 1, 2, ..., C\}$, intra-domain and inter-domain prototypical graphs $\mathcal{G}_{intra}^t$ and $\mathcal{G}_{inter}^t$ are established for an input video of target domain v_i^t. Their vertex features are defined as: $X_{intra}^t = [F_s(v_i^t) \parallel p_1^t \parallel p_2^t \parallel ... \parallel p_C^t] \in \mathbb{R}^{(C+1) \cdot m \times d}$ and $X_{inter}^t = [F_s(v_i^t) \parallel p_1^s \parallel p_2^s \parallel ... \parallel p_C^s] \in \mathbb{R}^{(C+1) \cdot m \times d}$, where $\parallel$ denotes the concatenation operation, d denotes the vertex feature dimension, $F_s(v_i^t) = [F_s(v_{i,1}^t), F_s(v_{i,2}^t), ..., F_s(v_{i,m}^t)]$.

Then, we perform graph convolution over the vertex features X_{intra}^t and X_{inter}^t for message propagation and class-aware temporal reasoning:

$$\begin{aligned} Z_{intra}^t &= \sigma(A_{intra}^t X_{intra}^t W_{intra}^t) \\ Z_{inter}^t &= \sigma(A_{inter}^t X_{inter}^t W_{inter}^t), \end{aligned} \tag{4}$$

where σ is ReLU activation function, A_{intra}^t and A_{inter}^t are adjacency matrices for intra-domain and inter-domain graphs, W_{intra}^t and W_{inter}^t are learnable weight matrices.

For better reasoning the temporal relationships between the input video and prototypes of different classes, we introduce graph attention mechanism to dynamically capture their segment-paired relations and induce the adjacency matrices A_{intra}^t and A_{inter}^t. For clear formulation, given any two segment-level vertex representations x_i and x_j in X^t ($X \in \{X_{intra}^t, X_{inter}^t\}$), the affinity edge $\alpha_{i,j}$ is defined as:

$$\alpha_{i,j} = \frac{\exp(\sigma(W_{att}[x_i || x_j]))}{\sum_{k=1}^{|X^t|} \exp(\sigma(W_{att}[x_i || x_k]))}, \tag{5}$$

where W_{att} is a learnable weight matrix for graph attention, $|X^t| = (C+1) \cdot m$ is the number of vertices.

Thanks to the aforementioned temporal graph reasoning, a discriminative representation of the video v_i^t can derive from the updated vertex features Z_{intra}^t and Z_{inter}^t given in Equation 4. Formally, we first rewrite Z_{intra}^t and Z_{inter}^t in the form of concatenated features:

$$\begin{aligned} Z_{intra}^t &= \left[z_{intra;v}^t \parallel z_{intra;p_1}^t \parallel z_{intra;p_2}^t \parallel ... \parallel z_{intra;p_C}^t\right] \\ Z_{inter}^t &= \left[z_{inter;v}^t \parallel z_{inter;p_1}^t \parallel z_{inter;p_2}^t \parallel ... \parallel z_{inter;p_C}^t\right], \end{aligned} \tag{6}$$

where $\parallel$ denotes the concatenation operation, $z_{intra;v}^t$ and $z_{intra;p_c}^t$ are the updated vertex features corresponding to $F(v_i^t)$ and p_c^t, other notations are similarly defined. Then we generate the video representation $f_{v_i}^t$ as the concatenation of these refined intra-domain and inter-domain features:

$$f_{v_i}^t = \left[z_{intra;v}^t \parallel z_{inter;v}^t\right]. \tag{7}$$

As for a given input video of source domain v_j^s, we also establish its intra-domain and inter-domain prototypical graphs, *i.e.*, $\mathcal{G}_{intra}^s$ and $\mathcal{G}_{inter}^s$ and generate its semantic representation $f_{v_j}^s$ in a similar way:

$$f_{v_j}^s = \left[z_{intra;v}^s \parallel z_{inter;v}^s\right]. \tag{8}$$

3.3. Cross-domain Alignment

Besides using our proposed prototypical graphs to explore the class-aware cross-domain temporal relationships, we aim to further equip each action category with a customized adaptation strategy to handle the heterogeneous temporal domain gaps among different classes. To achieve this, we adopt adversarial learning to align the video representations across domains and design multiple class-specific temporal domain discriminators $\{G_{td}^c | c = 1, 2, ..., C\}$, each is used for mitigating the temporal domain discrepancy related to the action category c.

For a source video v_i^s of class y_i^s, the temporal domain discriminator G_{td}^c is applied to its visual feature $f_{v_i}^s$:

$$L_{td}^s = -\mathbb{E}_{(v_i^s, y_i^s)\sim\mathcal{D}^s} \log\left[G_{td}^{y_i^s}(F_G(v_i^s))\right], \tag{9}$$

where $f_{v_i}^s = F_G(v_i^s)$, F_G which has been elaborated in Sec. 3.2 is the feature extractor for graph representations $f_{v_i}^s$ and $f_{v_j}^t$. However, for an unlabeled target video v_j^t, we can not exactly determine which adaptation strategy is applicable to it. Thus, we resort to the pseudo-labels $\{\hat{y}_{j,c}^t | c = 1, 2, ..., C\}$ provided by the label classifier G_y and assigning multiple domain discriminators to v_j^t in a relaxation way:

$$L_{td}^t = -\mathbb{E}_{v_j\sim\mathcal{D}^t} \sum_{c=1}^{C} \hat{y}_{i,c}^t \log\left[1 - G_{td}^c(F_G(v_j^t))\right], \tag{10}$$

where $f_{v_j}^t = F_G(v_j^t)$, $\hat{y}_j^t = [\hat{y}_{j,1}^t, \hat{y}_{j,2}^t, ..., \hat{y}_{j,C}^t] = G_y(f_{v_j}^t)$.

Additionally, we also introduce a spatial domain discriminator G_{sd} to align the segment-level features following [5]:

$$\begin{aligned} L_{sd} = &-\mathbb{E}_{(v_i^s, y_i^s)\sim\mathcal{D}^s} \sum_{k=1}^{m} \log\left[G_{sd}(F_s(v_{i,k}^s))\right] \\ &- \mathbb{E}_{v_j^t\sim D^t} \sum_{k=1}^{m} \log\left[(1 - G_{sd}(F_s(v_{j,k}^t)))\right], \end{aligned} \tag{11}$$

where F_s is the feature extractor for segment-level features, m is the number of segments, $v_{i,k}^s$ and $v_{j,k}^t$ are the segments of the source and target videos v_i^s and v_j^t, respectively.

To learn discriminative features, the label classifier G_y is trained using the label information from source domain:

$$L_y = -\mathbb{E}_{(v_i^s, y_i^s)\sim\mathcal{D}^s} L_{ce}(G_y(F_G(v_i^s)), y_i^s), \tag{12}$$

where L_{ce} is the cross-entropy loss.

Integrating all these components together, the overall optimization of PGA is performed in adversarial learning manner:

$$\begin{cases} \min\limits_{F_G, F_s, G_y} L_y - \lambda_d(L_{sd} + L_{td}^s + L_{td}^t) \\ \min\limits_{G_{sd}, G_{td}^1, G_{td}^2, ..., G_{td}^C} L_{sd} + L_{td}^s + L_{td}^t, \end{cases} \tag{13}$$

where λ_d is the trade-off hyper-parameter for adversarial learning.

4. EXPERIMENTS

4.1. Experimental Setup

Datasets We conduct our experiments on three publicly available datasets for cross-domain action recognition: UCF-HMDB [5], Jester (S)→Jeseter (T) [6] and Kinetics→NEC-Drone [8]. For fairness, we follow standard train/test split strategries provided by the authors. UCF-HMDB contains 12 classes of videos from UCF [15] and HMDB [16]. Jester (S)→Jeseter (T) is a large scale cross-domain gesture recognition dataset and contains 7 classes of videos collected from Jester [17]. Kinetics→NEC-Drone is a more challenging dataset containing 7 classes of action because its domain gap is much more significant compared to other datasets.

Implementation Details We use I3D [12] and TRN [13] as the backbone feature extractors F_s for extracting segment-level features. In our experiments, we set the number of segments $m = 5$ and each segment is comprised of 16 frames. We only extract RGB features for video representations and do not use optical flow. The momentum hyper-parameter η, vertex feature dimension d and trade-off hyper-parameter for adversarial learning λ_d are set to 0.1, 128 and 0.3, respectively. SGD with momentum of 0.9 and weight decay of 10^{-4} is used to train the network. The network is trained for 100 epochs with batch size of 64 and learning rate of 3×10^{-2}.

4.2. Comparison with Existing Methods

We mainly compare our method with other approaches using the same backbone feature extractors for fairness on three benchmarks: UCF-HMDB, Jester (S)→Jester (T) and Kinetics→NEC-Drone. To be specific, we use I3D backbone on UCF-HMDB and Kinetics→NEC-Drone, and use TRN backbone on Jester (S)→Jester (T), following the settings of

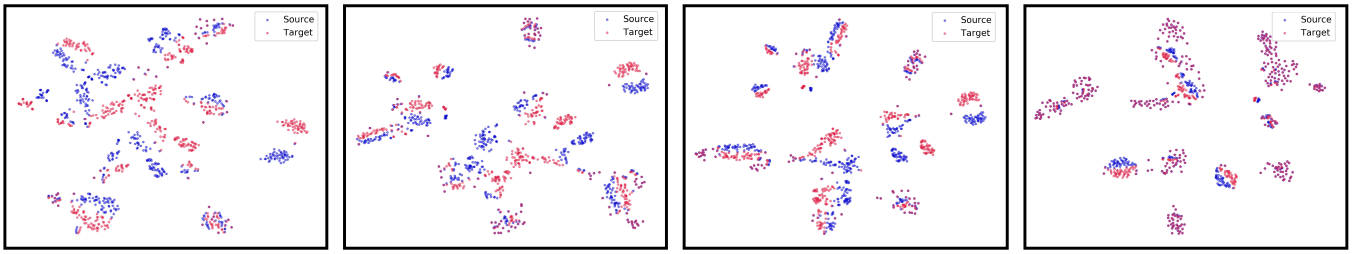

Fig. 3. The comparison of t-SNE visualization on HMDB→UCF. The blue and red dots represent source and target data, respectively. Best viewed in color.

Method	Backbone	U→ H	H → U	Average
TA^{3}N [5]	ResNet-101	78.3	81.8	80.1
ABG [14]		79.1	85.1	82.1
TCoN [6]	TRN	87.2	89.1	88.1
Source Only	I3D	80.3	88.8	84.5
TA^{3}N [5]		81.4	90.5	85.9
SAVA [8]		82.2	91.2	86.7
STCDA [7]		81.9	91.9	86.9
PGA (Ours)	I3D	**86.7**	**94.1**	**90.4**
Target Only	I3D	95.0	96.8	95.9

Table 1. Accuracy (%) on UCF-HMDB.

previous works. The results of the "Source Only" and supervised "Target Only" baselines with the same backbone are also provided for comparison.

Table 1 exhibits the results of different methods on UCF→HMDB and HMDB→UCF. We can observe that our PGA performs best in both directions, compared with other methods using I3D backbone. Especially, PGA achieves remarkable 3.5% performance improvement over STCDA [7], in terms of the average accuracy.

Method	Backbone	Jester (S)→Jester (T)
DAAA [1]	TSN	56.5
CDAN [2]		58.3
TCoN [6]		61.8
Source Only	TRN	51.2
TA^{3}N [5]		60.1
TCoN [6]		62.5
PGA (Ours)	TRN	**65.7**
Target Only	TRN	94.4

Table 2. Accuracy (%) on Jester (S)→Jester (T).

Table 2 shows the comparison on Jester (S)→Jester (T), which is a more challenging benchmark because the domain gap arises from action dynamics rather than spatial appearance [6]. It can be noticed that PGA greatly outperforms other approaches and achieves the best adaptation performance of 65.7% on Jester (T), which demonstrates the effectiveness of our temporal alignment strategy.

Method	Backbone	Kinetics→NEC-Drone
Source Only	I3D	17.2
DANN [3]		22.3
ADDA [4]		23.7
SAVA [8]		31.6
PGA (Ours)	I3D	**35.0**
Target Only	I3D	81.7

Table 3. Accuracy (%) on Kinetics→NEC-Drone.

In Table 3, we observe a large domain discrepancy on Kinetics→NEC-Drone, *i.e.*, 64.5% performance gap between "Source Only" and "Target Only" baselines. The experiment results show that PGA also achieve new state-of-the-art performance (35.0%) on this challenging cross-domain action recognition task.

Method	U→ H	H → U	Average
PGA	86.7	94.1	90.4
PGA w/o Intra-Graph	84.8	93.3	89.0
PGA w/o Inter-Graph	82.2	91.4	86.8
PGA w/o Graphs	80.7	89.4	85.1
PGA w/o L_{td}	83.5	92.1	87.8
PGA w/o L_{sd}	85.9	93.5	89.7

Table 4. Ablation study on UCF-HMDB.

4.3. Ablation Study

To evaluate the effectiveness of each component of our proposed PGA, we conduct ablation experiments on the UCF-HMDB dataset. To analyze the effectiveness of intra-domain / inter-domain prototypical graph representations and spatial / class-wise temporal domain discriminators, we study five variants of PGA in Table 4: (1) PGA w/o Intra-Graph: the variant without intra-domain graph reasoning $\mathcal{G}^s_{intra}$ and $\mathcal{G}^t_{intra}$; (2) PGA w/o Inter-Graph: the variant without inter-domain graph reasoning $\mathcal{G}^s_{inter}$ and $\mathcal{G}^t_{inter}$; (3) PGA w/o Graphs: the variant without prototypical graph representations; (4) PGA w/o L_{sd}: the variant without spatial domain discriminator G_{sd} for aligning segment-level features; (5)

PGA w/o L_{td}: the variant without class-wise temporal domain discriminator $\{G_{td}^c | c = 1, 2, ..., C\}$.

From Table 4, we can conclude that: (1) both intra-domain and inter-domain prototypical graphs contribute to the temporal segment-level relational reasoning within each domain and across different domains, which benefits exploiting class-aware relationships for temporal domain adaptation; (2) class-wisely temporal alignment is beneficial for adapting cross-domain temporal dynamics; (3) spatial domain discriminator helps capturing the spatial domain gap and further improve the capacity of spatial alignment.

Moreover, we plot the t-SNE visualization of the features learned on HMDB→UCF by the "Source Only" model and variants of PGA. As shown in Fig. 3, we can observe that prototypical graphs provide more discriminative video representations and class-wise temporal domain alignment enhance the cross-domain clustering in each class.

5. CONCLUSION

In this paper, we propose prototypical graph alignment (PGA) to achieve class-wise temporal alignment for cross-domain action recognition. We establish intra-domain and inter-domain prototypical graphs based on segment-level prototypes of different classes and exploit the class-aware fine-grained temporal relationships via graph convolutional networks. Furthermore, we propose to class-wisely align the video representations via class-specific domain discriminators. Experiments on three benchmarks validate the effectiveness of our proposed PGA, which achieves new state-of-the-art performance.

6. REFERENCES

[1] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh, "Deep domain adaptation in action space.," in *BMVC*, 2018, vol. 2, p. 5.

[2] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan, "Conditional adversarial domain adaptation," in *NeurIPS*, 2018.

[3] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[4] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[5] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng, "Temporal attentive alignment for large-scale video domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6321–6330.

[6] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles, "Adversarial cross-domain action recognition with co-attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11815–11822.

[7] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai, "Spatio-temporal contrastive domain adaptation for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9787–9795.

[8] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang, "Shuffle and attend: Video domain adaptation," in *European Conference on Computer Vision*. Springer, 2020, pp. 678–695.

[9] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, 2014, pp. 568–576.

[10] 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*. Springer, 2016, pp. 20–36.

[11] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[12] 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*, 2017, pp. 6299–6308.

[13] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba, "Temporal relational reasoning in videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 803–818.

[14] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh, "Adversarial bipartite graph learning for video domain adaptation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 19–27.

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," 2012.

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.

[17] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.