Multi-level Attentive Adversarial Learning with Temporal Dilation for Unsupervised Video Domain Adaptation

Peipeng Chen[*], Yuan Gao[*], Andy J. Ma[*†✉]
[*]School of Computer Science and Engineering, Sun Yat-sen University, China.
[†]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
chenpp7@mail2.sysu.edu.cn, gaoy266@mail2.sysu.edu.cn, majh8@mail.sysu.edu.cn

Abstract

Most existing works on unsupervised video domain adaptation attempt to mitigate the distribution gap across domains in frame and video levels. Such two-level distribution alignment approach may suffer from the problems of insufficient alignment for complex video data and misalignment along the temporal dimension. To address these issues, we develop a novel framework of ***Multi-level Attentive Adversarial Learning with Temporal Dilation (MA^{2}L-TD)****. Given frame-level features as input, multi-level temporal features are generated and multiple domain discriminators are individually trained by adversarial learning for them. For better distribution alignment, level-wise attention weights are calculated by the degree of domain confusion in each level. To mitigate the negative effect of misalignment, features are aggregated with the attention mechanism determined by individual domain discriminators. Moreover, temporal dilation is designed for sequential non-repeatability to balance the computational efficiency and the possible number of levels. Extensive experimental results show that our proposed method outperforms the state of the art on four benchmark datasets.*[1]

1. Introduction

As one of the most important multimedia modalities, video data increases rapidly in recent years. Video analysis has become one of the most active research areas due to its wide range of applications including video retrieval [5], understanding [32], recommendation [37], etc. Inspired by the progress in deep convolutional neural networks designed for image classification [9, 12, 27], many deep architectures [6, 24, 29] have been developed for video action classification by taking temporal cues into account. There are two major factors for the success of recent works: (i) the data fitting capacity of deep learning models with practical optimization techniques, and (ii) large-scale annotated datasets for training. Nevertheless, it is very expensive and time-consuming to manually annotate not only the action labels but also the start/end times of the actions for video data. To reduce the cost of manual annotations, classifiers trained by the available labelled data (source domain) can be employed for the testing environment (target domain). Despite the simplicity of such an approach, the classification performance probably drops due to the distribution mismatch across domains caused by the domain gap (e.g., changes of background, illumination, camera pose, and so on).

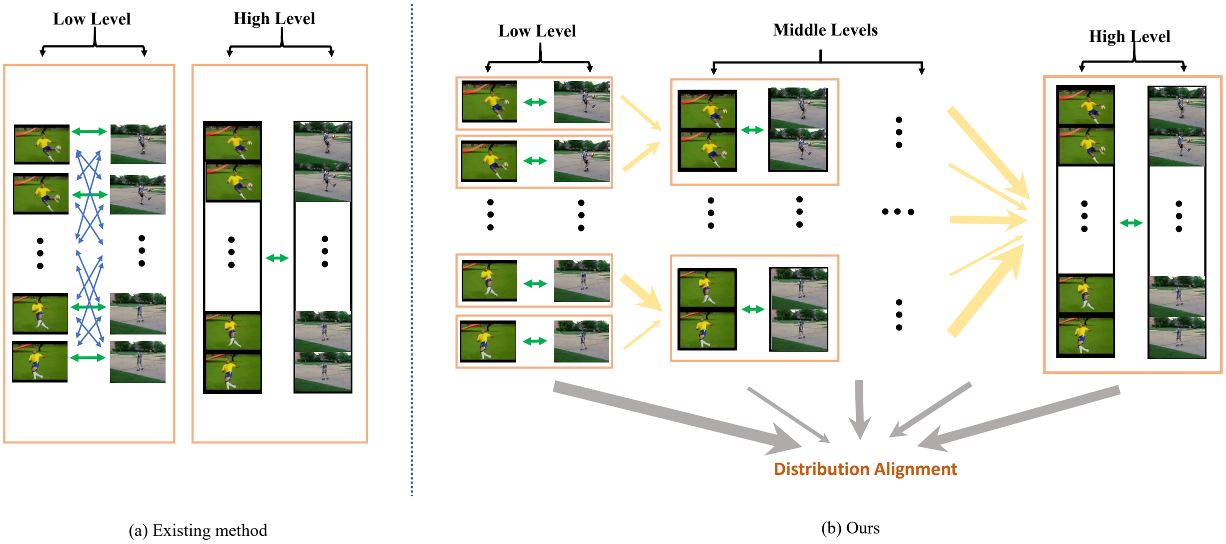
To address this issue, unsupervised domain adaptation (UDA) [7, 21, 23, 34] has been proposed, in which a set of labelled data in the source domain and a set of unlabelled data in the target domain are available for training. Though many UDA methods have achieved convincing results for image-based recognition tasks, they have not yet fully utilized temporal cues (which play an important role in understanding video data) for distribution alignment. To make use of temporal cues for domain discrepancy minimization, unsupervised video domain adaptation techniques [3, 18, 20] have been developed.

Though existing methods have advanced the task of unsupervised video domain adaptation, two limitations remain unsolved, as shown Fig. 1a. First, existing methods align distributions across domains in only frame and video levels. Due to the high complexity of video data, such two-level alignment methods may not guarantee that source and target distributions are sufficiently close. Second, it may be misleading to use only one domain discriminator for frame-level distribution alignment because the difference between temporal features in different time slots may be regarded as the domain gap and wrongly aligned.

To tackle the above-mentioned problems in existing methods, we develop a novel framework namely **Multi-level Attentive Adversarial Learning with Temporal Dilation (MA^{2}L-TD)** for unsupervised video domain adaptation. The main idea of the proposed method is shown

P. Chen, Y. Gao and A. J. Ma contribute to this work equally.

[1] Source code: https://github.com/justchenpp/MA2L-TD



(a) Existing method

(b) Ours

Figure 1. Main ideas of (a) existing method and (b) our proposed method. (a) The existing approach aligns features in only two levels, i.e. low level (frame) and high level (video), which may suffer from the problem of insufficient alignment due to the high complexity of video data. Moreover, frames may be misaligned at different time slots (blue arrows) by training only one domain discriminator (orange box) for low-level distribution alignment. (b) Our proposed method performs multi-level alignments from low level (frame), to middle levels (segment), and high level (video). Multiple domain discriminators (orange boxes) are trained with respect to different time-slots and levels (not shared even in the same level). Different attention weights (grey arrows) are assigned to each level for better distribution alignment. For feature aggregation, misaligned video segments are with smaller attention weights (yellow arrows).

in Fig. 1b. We propose to train multiple domain discriminators with respect to different time-slots and levels (not shared even in the same level) instead of only a shared one. For better distribution alignment, domain discriminators in different levels are weighted by the attention mechanism in adversarial training. The level-wise attention weights are computed by the degree of domain confusion in each level. To obtain multi-level temporal features, smaller attention weights determined by corresponding domain discriminators (not shared) are assigned to misaligned video segments for feature aggregation. At the same time, features are aggregated with temporal dilation for sequential nonrepeatability to balance the computational efficiency and the possible number of levels.

Major contributions of this work are summerized as follows:

- We develop a novel framework of **Multi-level Attentive Adversarial Learning with Temporal Dilation (MA^{2}L-TD)** for cross-domain action classification.
- We propose a multi-level attentive adversarial learning method for better distribution alignment. Attention weights in different levels are determined by the degree of domain confusion in each level.
- A new attentive temporally-dilated feature aggregation module is designed to generate multi-level temporal features, which mitigates the negative effect of misalignment and balances the computational efficiency and the possible number of levels.
- Extensive experiments show that our proposed method outperforms the state of the art on four benchmark datasets for unsupervised video domain adaptation.

2. Related work

2.1. Video Action Classification

Video action classification is more challenging than image recognition because of the higher complexity of video data. There are two widely used approaches to learn video representation for performance improvement by temporal cues. The first one is to use 2D convolution neural networks to model temporal relation. Based on the long-range temporal structure model, the temporal segment network (TSN) [30] utilize 2D convolution network in the spatial and temporal dimension respectively. Then, spatial and temporal features are combined to obtain the video-level representation. In [38], the temporal relation network (TRN) generates video-level feature vectors through temporal transformations and dependencies of frames in different time scales. Karen *et al.* [24] use RGB and optical flow to model spatial and temporal relations by 2D convolution neural network.

The second approach is based on 3D convolution neural networks. In the C3D [29], the 2D convolution kernel is inflated to 3D, such that the spatial and temporal information can be exploited simultaneously. The other representative method I3D [2] extends the idea of the two-stream network by using 3D convolutional kernels on RGB and optical flow.

2.2. Image-based Domain Adaptation

Domain adaptation aims to learn the knowledge from source domain and apply them to the target domain without serious performance degradation. To reduce the distribution mismatch, existing works attempt to align data distributions across domains for learning domain-invariant representations [15, 8, 16, 35, 36]. In the deep adaptation network (DAN) [15], transferable features are learned by minimizing the Maximum Mean Discrepancy (MMD) for distribution alignment. To match higher-order statistics across domains, Zellinge *et al.* [35] propose to a new distance function namely central moment discrepancy (CMD) for domain-invariant feature learning. Different from the moments-based approach, domain-adversarial neural network (DANN) [8] and conditional domain adversarial networks (CDANs) [16] learn a feature generator to deceive the domain discriminator. They extract discriminative features invariant from different domains by reversing gradients, where the gradient reversal layer (GRL) is optimized for the generator and discriminator simultaneously.

2.3. Unsupervised Video Domain Adaptation

Despite the great progress in image-based domain adaptation, there are less studies about the problem of unsupervised video domain adaptation. In [33, 28], this problem is addressed by using shallow learning methods. Collective matrix factorization or principal component analysis is utilized to learn the common latent semantic space for the source and target domain. With the success of deep neural networks for image applications, recent works [3, 20, 18] are proposed based on the development of general video action classification. The temporal attentive adversarial adaptation network (TA^{3}N) [3] aligns temporal relation features with attention computed by domain discrepancy. With the information of the temporal order and importance of video segments, (some of them may be irrelevant to the action, e.g., background frames), the temporal co-attention network (TCoN) [20] focuses on key segments shared by both domains for better alignment of temporal features. Instead of extracting domain-invariant representations, the frame-level and video-level bipartite graphs are used to model the relation between source-domain and target-domain features for recognizing domain-agnostic features in the adversarial bipartite graph (ABG) learning [18]. Nevertheless, existing methods may still suffer from the problems of feature mismatch between different time slots and insufficient distribution alignment in only two levels, i.e., frame (or segment) and video levels.

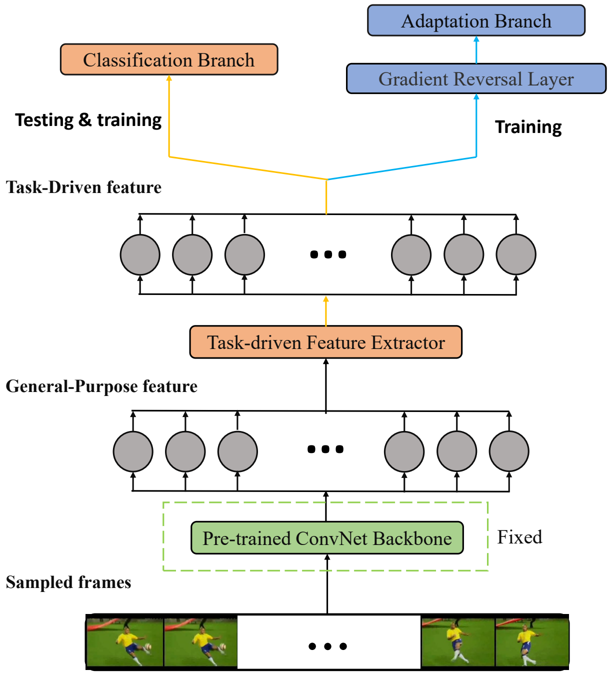


Figure 2. Network overview of the proposed method.

3. Method

In this section, we first give the problem definitions and network overview of the proposed **Multi-level Attentive Adversarial Learning with Temporal Dilation (MA^{2}L-TD)** (Section 3.1). The overall framework of the multi-level attentive adversarial learning is introduced in Section 3.2. Finally, Section 3.3 presents the attentive temporally-dilated aggregation module to obtain multi-level temporal features.

3.1. Definitions and Network Overview

The network overview of our proposed method is shown in Fig. 2. Suppose there are n_s labelled videos in the source domain and n_t unlabelled videos in the target domain. During training and testing, each video is divided into k segments with the same length and one frame in each segment is randomly sampled. For the jth frame of the ith (or i'th) video with k frames in the source (or target) domain, 2D convolutional backbone (e.g. ResNet) is used to extract the general-purpose feature vector f_{ij}^s (or $f_{i'j}^t$). Denote the source and target data as $\mathcal{D}_s = \{(F_i^s, y_i^s)\}_{i=1}^{n_s}$ and $\mathcal{D}_t = \{(F_{i'}^t)\}_{i'=1}^{n_t}$, where y_i^s is the label of the source domain data, $F_i^s = \{f_{ij}^s\}_{j=1}^k$ and $F_{i'}^t = \{f_{i'j}^t\}_{j=1}^k$ are the feature matrices in the source and target domain, respectively.

The objective of this work is to train a task-driven feature extractor by using multi-layer perceptron (MLP) for

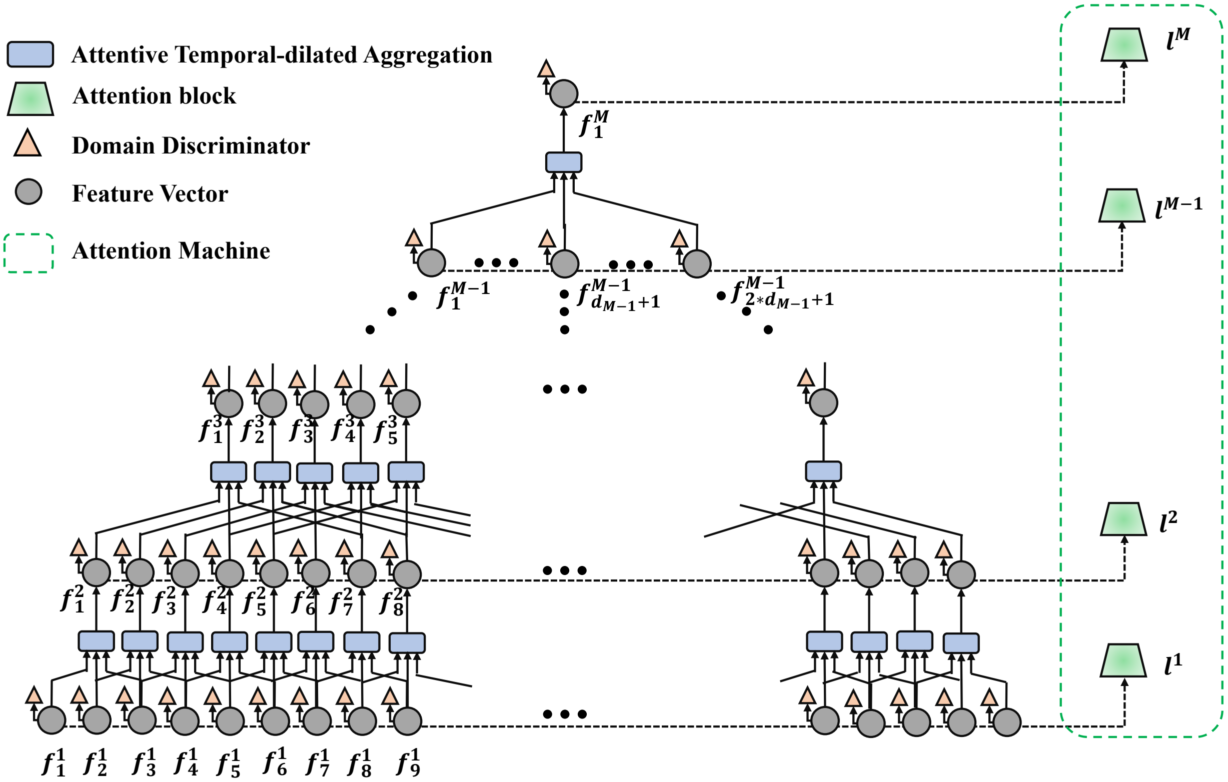

Attentive Temporal-dilated Aggregation
Attention block
Domain Discriminator
Feature Vector
Attention Machine
f_1^M
f_1^{M-1}
$f_{d_{M-1}+1}^{M-1}$
$f_{2*d_{M-1}+1}^{M-1}$
f_1^3 f_2^3 f_3^3 f_4^3 f_5^3
f_1^2 f_2^2 f_3^2 f_4^2 f_5^2 f_6^2 f_7^2 f_8^2
f_1^1 f_2^1 f_3^1 f_4^1 f_5^1 f_6^1 f_7^1 f_8^1 f_9^1
l^M
l^{M-1}
l^2
l^1

Figure 3. Framework of the proposed multi-level attentive adversarial learning. An individual domain discriminator D_j^m (not shared even in the same level) is adversarially trained for each feature vector f_j^m in time slot j of the mth level. The attention weight l^m determined by the degree of domain confusion is assigned to the mth level for distribution alignment. For attentive temporally-dilated aggregation, the kernel size is set as 3 for illustration. The dark/light and blue/gray colors refer to the temporal dilation operation. For example, in the 3rd level, the dark gray f_1^2, f_4^2, f_7^2 are aggregated to obtain f_1^3. Details of the aggregation module (blue block) can be referred to Fig. 4.

distribution alignment of video data. For convenience but without loss of generality, the task-driven feature vectors are denoted by f_{ij}^s and $f_{i'j}^t$ (the same as the general-purpose feature vectors). During training, the task-driven feature vectors are fed into both the classification and adaptation branches for optimization. For testing, only the classification branch is used.

In the classification branch, a classifier C is trained by minimizing the following loss functions of cross entropy L_s and soft entropy L_t for the source and target data respectively, i.e.,

$$L_s = -\frac{1}{n_s}\sum_{i=1}^{n_s} y_i^s log\, C(F_i^s). \tag{1}$$

$$L_t = -\frac{1}{n_t}\sum_{i'=1}^{n_t} C(F_{i'}^t) log\, C(F_{i'}^t). \tag{2}$$

For the adaptation branch, the distributions of task-driven feature vectors across domains are aligned by adversarial learning with the gradient reversal layer. The design of the adaptation branch includes multi-level attentive adversarial learning and attentive temporally-dilated aggregation, which will be detailed in Section 3.2 and Section 3.3, respectively.

3.2. Multi-level Attentive Adversarial Learning

The proposed adaptation branch of multi-level attentive adversarial learning is illustrated in Fig. 3. The input to this branch is the set of the task-driven feature vectors of a video in the source (or target) domain, i.e., f_{ij}^s (or $f_{i'j}^t$), $j = 1, \cdots, k$. They can be considered as the 1st-level features along the temporal dimension and denoted as $f_1^1, \cdots, f_k^1$. To obtain higher-level temporal features, we propose an attentive feature aggregation module with details elaborated in Section 3.3. Let the feature vector of the mth level in certain time slot j be f_j^m. For larger m (higher level), f_j^m is corresponding to a longer video segment, while smaller m (lower level) means shorter segments.

For better distributions alignment across domains, a domain discriminator D_j^m is adversarially trained for each time slot j and each level m. The loss function defined for the mth level is given as follows,

$$L^m = \sum_j L_b(D_j^m(f_j^m), d). \tag{3}$$

where d is the domain label for the input video clip and L_b is the binary entropy loss. By combining loss functions of multiple levels, the overall loss is defined as follows,

$$L_{adv} = \sum_m L^m. \tag{4}$$

For some videos, features can be better aligned in lower levels (short segments), while it is easier to align higher-level features (longer segments) for others. Therefore, we assign different weights to features in different levels and determine their importance by the attention mechanism (as shown in the rightmost column of Fig. 3). For this purpose, average pooling is used to attain the video presentation F^m in the mth level, i.e.,

$$F^m = average\_pooling(f_1^m, f_2^m, ..., f_{k_m}^m). \tag{5}$$

where k_m is the number of feature vectors in the mth level, such that $k = k_1 \geq k_2 \geq \cdots \geq k_M = 1$. To compute the attention weight l^m of the mth level, a domain discriminator $\mathcal{D}^m$ is learned by calculating the degree of domain confusion based on F^m (without back propagation for adversarial training). Then, the attention weight l^m is measured by the binary entropy loss of classifying F^m to the corresponding domain d, i.e.,

$$l^m = L_b(\mathcal{D}^m(F^m), d). \tag{6}$$

By eq. (6), the attention weight l^m is larger if F^m is more difficult be classified correctly by the domain discriminator $\mathcal{D}^m$. This means the domain gap is smaller in the mth level so that larger weight is assigned.

Before integrating the attention mechanism into the multi-level loss function, the weights are normalized, i.e.,

$$\omega^m = \frac{l^m}{\sum\limits_m l^m}. \tag{7}$$

With the normalized attention weights, the multi-level loss (4) becomes,

$$L_{adv} = \sum_m \omega^m L^m. \tag{8}$$

By combining eqs. (1) (2) (8), the optimization problem of our proposed method is given as follows,

$$\begin{gathered} \min_{\theta_d} L_{adv}, \\ \min_{\theta_c, \theta_a, \theta_{mlp}} L_s + L_t - L_{adv}. \end{gathered} \tag{9}$$

where θ_c, θ_d, θ_a, and θ_{mlp} respectively denote the learnable parameters of the action classifier C, domain discriminators D_j^m, feature aggregation modules A_j^m (with details in Section 3.3) and task-driven feature extractor based on MLP.

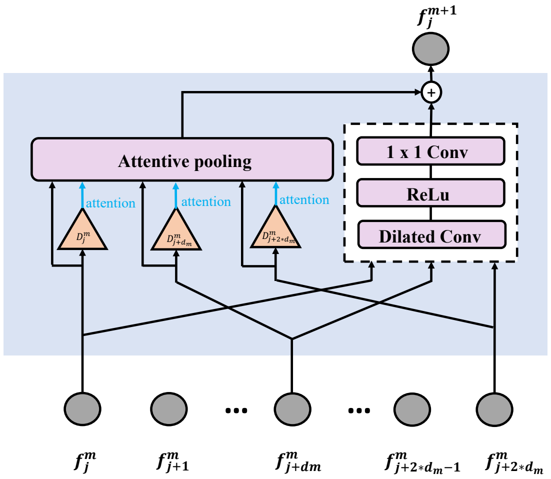


Figure 4. Attentive temporally-dilated aggregation: For illustration, the kernel size of the dilated convolution is set as 3 with the dilation rate d_m in the mth level.

3.3. Attentive Temporally-Dilated Aggregation

The proposed attentive temporally-dilated aggregation module is illustrated in Fig. 4. Different from dense feature aggregation which is computationally intensive, the temporal dilation operation can greatly reduce the computational complexity. Denote the dilation rate in the mth level as d_m and the size of the convolutional kernel as q. The input to the aggregation module A_j^m is the set of feature vectors in the mth level, i.e., $S_j^m = \{f_j^m, f_{j+d_m}^m, \cdots, f_{j+(q-1)d_m}^m\}$. The output to A_j^m is f_j^{m+1} in the $(m+1)$th level. As shown in Fig. 4, the aggregation module A_j^m is composed of an attention pool layer (left-hand side) and one-dimension residual convolution operation H_j^m (right-hand side), i.e.,

$$f_j^{m+1} = A_j^m(S_j^m) = AttPool(S_j^m) + H_j^m(S_j^m). \tag{10}$$

The dilation rate d_m is an important parameter to control the size of temporal features in multiple levels. For a fixed number of temporal features in the 1st level k_1, if d_m is too small, the model complexity will become intractable with a large number of levels and feature vectors in each layer. On the other hand, if d_m is too large, the number of levels will become small so that features across domains cannot be aligned sufficiently. As shown in Fig. 3, we set $d_1 = 1$ to aggregate successive temporal features in the 1st level to obtain temporal features in the 2nd level. For $m \geq 2$, d_m is set as the smallest number to ensure that feature vectors in the $(m-1)$th level are not repeatedly used in sequential order to obtain f_j^{m+1}. According to eq. (10), f_j^{m+1} is computed by $f_{j+sd_m}^m \in S_j^m, s = 0, \cdots, q-1$. The mth-level feature vector $f_{j+sd_m}^m$ is generated by $f_{j+sd_m}^{m-1}$, $f_{j+sd_m+d_{m-1}}^{m-1}$, $\cdots$, $f_{j+sd_m+(q-1)d_{m-1}}^{m-1}$ in the $(m-1)$th level. Therefore,

Table 1. Statistics of the four benchmark datasets

Dataset	$U-O$	$U-H_{small}$	$U-H_{full}$	$K-G$
Video Length	$<$ 39 s	$<$ 21 s	$<$ 33 s	$<$ 10 s
Class Number	6	5	12	30
Total Number	1145	1171	3209	49998
Training Number	U(601) O(250)	U(482) H(350)	U(1438) H(840)	K(43378) G(2625)
Testing Number	U(240) O(54)	U(189) H(150)	U(571) H(360)	K(3246) G(749)

we set $d_m = (q-1)d_{m-1} + 1$ as the smallest number for sequential non-repeatability.

On the other hand, some $f_{j+sd_m}^m \in S_j^m$ (corresponding to a video segment in a certain time slot) is probably misaligned along the temporal dimension between different videos. To mitigate the negative influence caused by the misalignment, attention weights are assigned to $f_{j+sd_m}^m$ to determine their importance for aggregation. For this purpose, the domain discriminator $D_{j+sd_m}^m$ is used to calculate the binary entropy loss of classifying $f_{j+sd_m}^m$ to the correspond domain as the attention weight $\ell_{j+sd_m}^m$, i.e.,

$$\ell_{j+sd_m}^m = L_b(D_{j+sd_m}^m(f_{j+sd_m}^m), d). \tag{11}$$

Similar to eq. (6), eq. (11) means that the domain gap is smaller if the attention weight $\ell_{j+sd_m}^m$ is larger. By normalizing the attention weights and substituting them into eq. (10), the proposed attentive temporally-dilated aggregation becomes,

$$\begin{aligned} f_j^{m+1} &= H_j^m(S_j^m) + \sum_{s=0}^{q-1} w_{j+sd_m}^m f_{j+sd_m}^m, \\ w_{j+sd_m}^m &= \ell_{j+sd_m}^m \Big/ \sum_{s=0}^{q-1} \ell_{j+sd_m}^m. \end{aligned} \tag{12}$$

4. Experiment

4.1. Datasets

We compare our proposed method with the state of the art on four benchmarks. 1) $UCF-Olympic$: There are 6 common classes from both the UCF101 [25] and Olympic Sports dataset [19]. With totally 1,145 videos, 601 and 240 videos are used for training and testing respectively in the UCF101, while there are 250 and 54 videos in Olympic Sport for training and testing respectively. 2) $UCF101-HMDB51_{small}$ [31]: The intersection subset of UCF101 [25] and HMDB51 dataset [13] has 5 classes and 1171 videos. There are 482 training videos and 189 testing videos in the UCF. In HMDB, 350 and 150 videos are used for training and testing respectively. 3) $UCF101-HMDB51_{full}$ [3]: The intersection subset of UCF101 [25] and HMDB51 dataset [13] has 12 classes and 3,209 videos. There are 1,438 training videos and 571 testing videos in the UCF. In HMDB, 850 and 350 videos are used for training and testing respectively. 4) $Kinetics-Gameplay$ [1, 3, 11]: It has 30 classes and 49,998 videos. There are 43,378 training videos and 3,246 testing videos in the Kinetics. In Gameplay, 2,625 and 749 videos are used for training and testing respectively. The statistics information of the four benchmark datasets are shown in Table 1.

4.2. Implementation Details

Our proposed method is implemented in the PyTorch framework and the source code is released here. For fair comparison with other methods, the backbone convolutional network is the Resnet101 [9] pretrained on ImageNet [4]. The sampling strategy is to divide each video into k segments and then randomly sample one frame from each segment. The kernel size in the attentive temporally-dilated aggregation is set as 3 in all experiments unless otherwise stated. In the $UCF-Olympic$ and $UCF101-HMDB51_{small}$ dataset, the batch size is 32 and 23 frames are sampled from each video with the number of levels equal to 4. In the $UCF101-HMDB51_{full}$ dataset, the batch size is 32 and 53 frames are sampled from each video with the number of levels equal 5. In the $Kinetics-Gameplay$ dataset, the batch size is 64 and 23 frames are sampled from each video with the number of levels equal to 4. In all the experiments, we use Adam as the optimizer and the learning rate is initiated as 3 x 10^{-4} and decays when the epoch increases.

4.3. Comparison with Other Methods

In this section, we first compare our proposed method of multi-level attentive adversarial learning with temporal dilation (MA^{2}L-TD) with existing works. The comparison results on the small datasets $UCF101-HMDB51_{small}$ and $UCF-Olympic$ are shown in Table 2. Source only means that the backbone network is trained only in the source domain and tested in target domain. From these results, we can see that our proposed method outperforms state-of-the-art methods on the two benchmark datasets. On the task of $UCF \rightarrow Olympic$, our method achieves 100% accuracy and is better then other methods by a margin of 1.85% (from 98.15% to 100%). On the two larger benchmark datasets $UCF101-HMDB51_{full}$ and $Kinetics-$

Table 2. Classification accuracy (%) on benchmark datasets: $UCF - Olympic$ and $UCF101 - HMDB51_{small}$

Method	Backbone	$U \rightarrow O$	$O \rightarrow U$	$U \rightarrow H$	$H \rightarrow U$
Source only	TSN	80.00	76.67	-	82.10
Source only	C3D	82.13	83.16	-	-
W.Sultani et al. [26]	-	33.33	47.91	68.70	68.67
Many-to-one [33]	action bank	87.00	75.00	82.00	82.00
AMLS(SA) [10]	C3D	84.65	86.44	89.53	95.36
AMLS(GFK) [10]	C3D	83.92	86.07	90.25	94.40
DAAA [10]	TSN	88.37	86.25	-	88.36
DAAA [10]	C3D	91.60	89.96	-	-
TcoN [20]	TSN	93.91	91.65	-	93.01
TA^3N [3]	ResNet-101	98.15	92.92	99.33	99.47
ABG [18]	ResNet-101	98.15	92.50	99.33	98.41
Ours	ResNet-101	**100**	**94.72**	**99.33**	**99.47**

Table 3. Classification accuracy (%) on benchmark dataset: $UCF101 - HMDB51_{full}$ and $Kinetics - Gameplay$

Method	$UCF101 \rightarrow HMDB51$	$HMDB51 \rightarrow UCF101$	$Gameplay \rightarrow Kinetics$
DANN [8]	75.28	76.36	20.56
JAN [17]	74.72	79.69	18.16
AdaBN [14]	72.22	77.41	20.29
MCD [22]	73.89	79.34	19.76
TA^3N [3]	78.33	81.79	27.50
ABG [18]	79.19	85.11	27.89
Source only	76.39	78.11	17.56
Ours	**85.00**	**86.59**	**31.45**

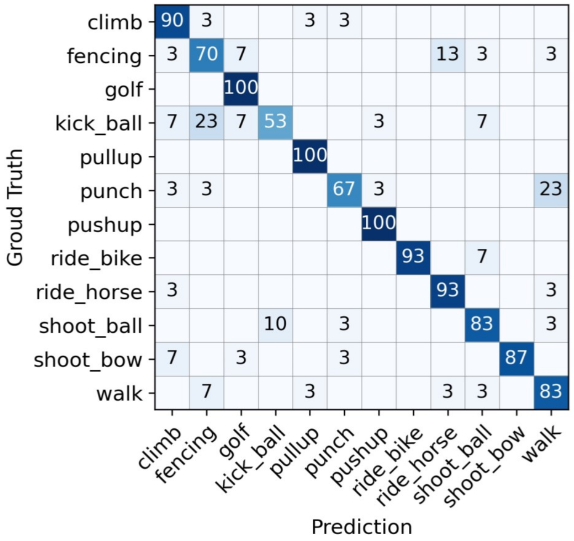


Figure 5. The confusion matrix (%) of our proposed method on benchmark dataset: $UCF101 \rightarrow HMDB51_{full}$.

$Gameplay$, results reported in Table 3 compare our proposed method with the state of the art by using the same backbone of ResNet101. From these results, we can see that our proposed method achieves a large margin (5.81%) of improvement on the task of $UCF101 \rightarrow HMDB51_{full}$. To evaluate the per-class recognition performance, the confusion matrix on $UCF101 \rightarrow HMDB51_{full}$ dataset is shown in Fig. 5.

4.4. Ablation Experiment

To investigate the effectiveness of the components in our proposed method, ablation experiments are performed on the $HMDB51 - UCF101_{full}$ dataset. The classification accuracy is shown in the Tabel 4. The proposed MA^{2}L-TD w/o individual domain discriminators in each level means that only one shared discriminator is trained for each level. Without distribution alignment in different time slots, the classification accuracy decreases by 3.43% and 4.35% on the tasks of the $UCF101 \rightarrow HMDB51$ and $HMDB51 \rightarrow UCF101$, respectively. This convinces that distributions can be better aligned by training individual domain discriminators with respect to different time slots. MA^{2}L-TD w/o multi-level alignment means that only the frame- and video-level features are aligned. For this experiment, the classification accuracy decreases by 2.59% and 3.81% in the two tasks. These results show that alignments of middle-level features can help for performance improvement. MA^{2}L-TD w/o level-wise attention means that every level has the same weight. Without level-wise attention, the classification accuracy decreases by 0.89% and 0.95% respectively, which verifies the usefulness of the attention mechanism in each level. MA^{2}L-TD w/o attentive

Table 4. Classification accuracy (%) of the ablation study on benchmark dataset: $UCF101 - HMDB51_{full}$

Method	$UCF101 \rightarrow HMDB51$	$HMDB51 \rightarrow UCF101$
MA^{2}L-TD w/o individual domain discriminators in each level	81.57	82.24
MA^{2}L-TD w/o multi-level alignment	82.41	82.78
MA^{2}L-TD w/o level-wise attention	84.11	85.64
MA^{2}L-TD w/o attentive temporally-dilated aggregation	81.67	82.39
MA^{2}L-TD	**85.00**	**86.59**

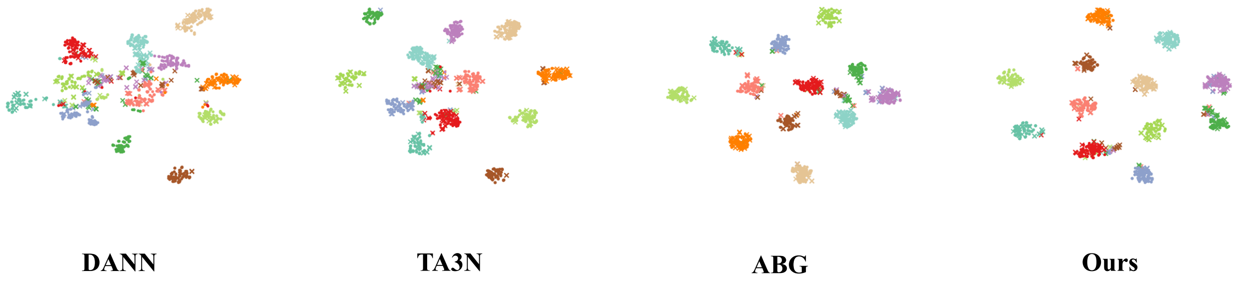


Figure 6. The t-SNE visualization of source and target domain features on the task: $UCF101 \rightarrow HMDB51_{full}$.

Table 5. Classification accuracy (%) with different number of level and kernel size on the task: $UCF101 \rightarrow HMDB51_{full}$

level / kernel	3	4	5
2	80.83	82.50	82.78
3	81.67	84.13	**85.00**
4	83.61	84.44	-

temporally-dilated aggregation means that average pooling is used instead for feature aggregation. With simple average pooling, the classification accuracy decreases by 3.33% and 4.20% in the two tasks. These results validate the effectiveness of the attentive temporally-dilated aggregation module. These ablation experiments show that all the components of our proposed method can help to improve the performance for cross-domain action recognition.

4.5. Parameter Sensitivity

To study the hyperparameters of kernel size and level number, experiments in the $UCF101 \rightarrow HMDB51_{full}$ task are conducted. Results with varying number of kernel size and levels are shown in Table 5. The accuracy is 82.50 when the kernel size is 2 and the number level is 4 with 7 sampled frames. For both kernel size and level number equal to 3 with 9 sampled frames, the accuracy is 81.67. These results show that larger number of levels can help to achieve higher accuracy. Besides, all the results in Table 5 still outperform TA^3N [3] and ABG [18]. This indicates that our method can robustly improve the performance with respect to different numbers of kernel size and level.

4.6. t-SNE Visualization

For visualization comparison, the t-SNE results of DANN, TA^3N, ABG and our method are shown in Fig. 6. In this figure, features in the last layer are used to generate the low-dimensional embeddings. Different colors stand for different classes, circle and cross represent the source and target domain respectively. From these results, we can see that features of the same class can be better clustered together while features of different classes are farther away by using our method.

5. Conclusion

In this work, we develop a novel framework namely Multi-level Attentive Adversarial Learning with Temporal Dilation (MA^{2}L-TD) for unsupervised video domain adaptation. In our method, multiple domain discriminators are trained for multi-level temporal features with the level-wise attention for better distribution alignment. To mitigate the negative effect of misalignment along the temporal dimension, features are aggregated by the attentive temporally-dilated aggregation module, which can balance the computational efficiency and the possible number of levels. Extensive experiments on the four benchmark datasets show that our proposed method outperforms the state of the art for cross-domain action recognition.

Acknowledgment. This work was supported partially by NSFC (No. 61906218), Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515011497) and Science and Technology Program of Guangzhou (No. 202002030371).

References

[1] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[2] 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.

[3] 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*, pages 6321–6330, 2019.

[4] 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.

[5] Yajiao Dong and Jianguo Li. Video retrieval based on deep convolutional neural network. In *Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing*, pages 12–16, 2018.

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[7] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of International conference on machine learning*, pages 1180–1189, 2015.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, volume 2, page 4, 2018.

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

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

[14] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

[15] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proceedings of International conference on machine learning*, pages 97–105, 2015.

[16] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Proceedings of Neural Information Processing Systems*, 2017.

[17] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[18] 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*, pages 19–27, 2020.

[19] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on computer vision*, pages 392–405. Springer, 2010.

[20] 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*, volume 34, pages 11815–11822, 2020.

[21] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017.

[22] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018.

[23] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision*, pages 153–168, 2018.

[24] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Proceedings of Neural Information Processing Systems*, 2014.

[25] 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.

[26] Waqas Sultani and Imran Saleemi. Human action recognition across datasets by foreground-weighted histogram decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 764–771, 2014.

[27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[28] Jun Tang, Haiqun Jin, Shoubiao Tan, and Dong Liang. Cross-domain action recognition via collective matrix factorization

with graph laplacian regularization. *Image and Vision Computing*, 55:119–126, 2016.

[29] 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*, pages 4489–4497, 2015.

[30] 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.

[31] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, et al. Hmdb: the human metabolome database. *Nucleic acids research*, 35(suppl_1):D521–D526, 2007.

[32] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[33] Tiantian Xu, Fan Zhu, Edward K Wong, and Yi Fang. Dual many-to-one-encoder-based transfer learning for cross-dataset human action recognition. *Image and Vision Computing*, 55:127–137, 2016.

[34] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2720–2729, 2019.

[35] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (cmd) for domain-invariant representation learning. *Proceedings of The International Conference on Learning Representations*, 2017.

[36] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *Proceedings of International Conference on Machine Learning*, pages 7523–7532, 2019.

[37] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 43–51, 2019.

[38] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision*, pages 803–818, 2018.