Spatio-temporal Contrastive Domain Adaptation for Action Recognition

Xiaolin Song[1], Sicheng Zhao[2], Jingyu Yang[1,*], Huanjing Yue[1], Pengfei Xu[3,*], Runbo Hu[3], Hua Chai[3]
[1]Tianjin University [2]University of California, Berkeley [3]Didi Chuxing
{songxl, yjy, huanjing.yue}@tju.edu.cn, schzhao@gmail.com,
{xupengfeipf, hurunbo, chaihua}@didiglobal.com

Abstract

Compared with image-based UDA, video-based UDA is comprehensive to bridge the domain shift on both spatial representation and temporal dynamics. Most previous works focus on short-term modeling and alignment with frame-level or clip-level features, which is not discriminative sufficiently for video-based UDA tasks. To address these problems, in this paper we propose to establish the cross-modal domain alignment via self-supervised contrastive framework, i.e.*, spatio-temporal contrastive domain adaptation (STCDA), to learn the joint clip-level and video-level representation alignment. Since the effective representation is modeled from unlabeled data by self-supervised learning (SSL), spatio-temporal contrastive learning (STCL) is proposed to explore the useful long-term feature representation for classification, using self-supervision setting trained from the contrastive clip/video pairs with positive or negative properties. Besides, we involve a novel domain metric scheme,* i.e.*, video-based contrastive alignment (VCA), to optimize the category-aware video-level alignment and generalization between source and target. The proposed STCDA achieves stat-of-the-art results on several UDA benchmarks for action recognition.*

1. Introduction

Unsupervised domain adaptation (UDA) has made great progress on computer vision tasks with the improvement of the representation ability on convolutional neural networks (CNNs). It aims at transferring the knowledge from the labeled source domain with specific supervision to the target domain with unlabeled data and the different domain distribution, for reducing dependencies on the comprehensive annotations and particular datasets. A great number of UDA methods have been proposed for image-based benchmarks, *e.g.*, image classification [39, 34, 26], object detection [11, 1, 13], and semantic segmentation [3, 6], which supply the applications with unsupervised learning and leverage impressive performance. However, the progress on UDA for video analysis is still limited, since video-based UDA tasks are more challenging. Firstly, video-based tasks need to model the multi-dimension information, which includes richer spatial appearance and temporal dynamics than images. Secondly, they require exploration of the association and interaction in the space and time dimensions. Finally, video UDA is essential to optimize the domain alignment in both spatial and temporal association. Even though larger datasets have been released with great diversity for video understanding tasks, the applications on different scenes are promoted in slow progress with limited generalization, which rely on numerous unannotated videos for representation in the corresponding feature spaces.

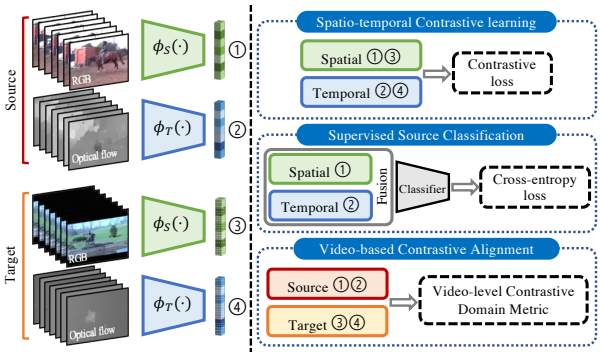

Figure 1. Overview of STCDA framework. Spatio-temporal contrastive learning (STCL) and video-based contrastive alignment (VCA) are proposed to model and align cross-domain features with long-term spatio-temporal representation.

In this paper, we address the challenging and valuable task of UDA for action recognition in videos, and explore

* Corresponding authors.
This work was done during Xiaolin Song's internship at Didi Chuxing.

the spatio-temporal representation to facilitate video-level modeling. Most previous UDA methods for action recognition [15, 4] focus on the short-term temporal representation with frame-level or clip-level modeling, using adversarial learning to align source and target features following image-based UDA works. However, it is unreasonable to ignore long-term modeling and alignment for video prediction, which are essential for video-based UDA tasks.

Integrating self-supervision module is a creative exploration for UDA to analyze the unannotated data for more effective feature. With the development of self-supervised algorithms, conducting supervised tasks with some customized rules becomes essential to explore the intrinsic information and statistical characteristics of unlabeled data. And these customized tasks are suitable for UDA to learn the implicit properties of source and target data without any labels. These tasks are suitable for UDA to learn the implicit properties of source and target data without any labels. Furthermore, self-supervision on the target data alone would not exploit the performance in UDA tasks, and therefore it is more reasonable to apply self-supervision on both the source and target data. Most self-supervised learning (SSL) methods [24, 9, 22, 18] are based on generative/predictive tasks, with particular supervised functions. In contrast, contrastive methods learn representations by contrasting positive and negative samples, which are flexible for improving the capacity to model correlations or complex structures with latent generalization, rather than overly pay attention to detail tasks in generative methods.

Compared with pre-training on labeled data, CNNs could obtain promising representing capability with contrastive-learning-based pre-training. MoCo [10] benefits downstream tasks with contrastive pre-training on ImageNet dataset, and outperforms most image-based supervised pre-training approaches. Thus the selection of positive-negative samples plays a decisive role in ensuring the quality of feature expression. Recent methods [32, 31] exploit contrastive modeling for videos using coupling networks, which are trained with frame-level or clip-level positive-negative examples. However, self-supervised contrastive learning for videos has been not leveraged for long-term video-level representation. From this point of view, we propose a novel decoupling framework for video understanding based on spatio-temporal contrastive learning, with jointly global(video-level)-local(clip-level) modeling in time dimension.

In addition, to optimize the feature alignment between source and target domains, we propose a category-aware video-level contrastive domain distance metric for UDA, which aims at improving the discrimination via cross-domain feature alignment. The alignment on clip and video levels are utilized to draw closer the samples under the same class and push apart the samples among different classes between source and target domains. Specifically, we measure self-modal and cross-modal metric respectively for reducing spatio-temporal domain shift. As shown in Figure 1, a novel spatio-temporal contrastive domain adaptation (STCDA) framework is proposed to overcome the misalignment in feature space and category confusion in different domains for UDA on action recognition.

In summary, our contributions of the proposed framework are as follows: (1) We design a novel spatio-temporal contrastive learning (STCL) framework, for learning joint clip-level and video-level feature representations by self-supervision, which improves the generalization of local-global temporal content modeling. (2) We propose a novel domain metric, *i.e.*, video-based contrastive alignment (VCA), to measure the video-level discrepancy between source and target domains. (3) Our proposed framework achieves stat-of-the-art results on several domain adaptation benchmarks for action recognition.

2. Related Work

Supervised action recognition. Supervised action recognition methods have made great progress and achieved impressive results using deep learning algorithms, especially by leveraging CNNs for spatio-temporal information modeling with various network architectures. Two-stream networks [29] utilize the multi-modal architecture and fuse the prediction of each modality, *i.e.*, appearance stream and motion stream. TSN [35] is an impressive extension of the two-stream networks and leverages the long-term temporal modeling as the input via sparse sampling of each video in time domain. Besides, for temporal modeling via 3D convolution, C3D [33] uses a full 3D convolutional-layer architecture for spatio-temporal feature extraction. I3D [2] uses an inflated Inception architecture with 3D convolutional layers utilized in each stage. In addition, some methods aim at long-term temporal context modeling, *e.g.*, non-local network [36]. However, these supervised action recognition approaches are still limited with the dependency on annotated labels for each clip. There is no guarantee of the robust performance if the algorithms are directly transferred to another domain, due to the presence of domain shift.

Self-supervision. Self-supervision is used to learn the feature representation with the prior characters supervised instead of the annotation supervised. Image-based self-supervised methods address the spatial content association to generalize the image representation, *e.g.*, image colorization [17] to perceive the color prior of natural images, and jigsaw puzzle [24] and rotation prediction [9] to learn the relative position correspondence. Furthermore, video-based self-supervised methods aim at exploring the content and association in both space and time dimensions. For conducting generative/predictive tasks, Misra *et al*. [22] propose an unsupervised sequential task for temporal order

verification. OPN [18] predicts the correct order of temporally shuffled sequences for frame sorting. Xu *et al*. [38] propose a clip order prediction framework to learn the spatio-temporal video representation by sorting the order of shuffled clips. MocycleGAN [5] uses optical-flow-based correspondence warping to optimize the pixel-wise detail characterization for unpaired video translation. For conducting contrastive tasks, Wang and Gupta [37] propose a Siamese-triplet network with a ranking loss to learn visual positive-negative representations from videos. MoCo [10] uses a momentum contrast mechanism to update the query encoder in the decoupling networks. CMC [32] utilizes a multi-view self-supervised contrastive framework, and uses RGB and optical flow for frame-level contrastive learning. IIC [31] leverages an inter-intra contrastive framework with preset positive-negative samples for clip-level representation, which however uses a coupled network and modified optical flow clips of only one direction (u in x-axis direction or v in y-axis direction).

Unsupervised Domain Adaptation. Unsupervised domain adaptation (UDA) methods have been widely proposed for image-based tasks. DAN [20] and JAN [21] are proposed to align the joint distributions by minimizing maximum mean discrepancy (MMD) and joint maximum mean discrepancy (JMMD) between source and target domains, respectively. CAN [16] models the domain discrepancy based on inter-intra classes. Recently, there are several works on video domain adaptation. DAAA [15] utilizes an adversarial learning framework with 3D CNN to align source and target domains. TA^{3}N [4] leverages a multi-level adversarial framework with temporal relation and attention mechanism to align the temporal dynamics of feature space for videos. TCoN [25] matches the feature distributions between source and target domains, for temporal alignment using the cross-domain co-attention mechanism.

Several methods also leverage video domain adaptation using SSL auxiliary. MM-SADA [23] learns the multi-modal correspondence of RGB and optical flow for self-supervised domain adaptation. SAVA [7] proposes a self-supervised predictive method for video domain adaptation, which aims to predict the clip order with the adversarial loss. However, there is no exploration on video domain adaptation with self-supervised contrastive learning. We propose a contrastive framework to model decoupled representation of different modalities, and combine with the proposed video-based contrastive metric for self-supervised action recognition.

3. Method

In this section, we present the main building blocks of the proposed spatio-temporal contrastive domain adaptation (STCDA) framework for action recognition. The overview of our STCDA is shown in Figure 1. For UDA, we conduct the network in two parts to capture the semantic information from input videos. First, we build a video-based contrastive learning framework for self-supervised learning (SSL), aiming at improving the clip-level and video-level generalization capacity in the target domain. Second, we propose the video-based contrastive distance metric for mitigating the domain shift of data distribution between two domains. Each operation is adapted in both clip and video levels, and the details of the framework are presented in Figure 2. Formally, denote the source samples as $\mathcal{S} = \{(X_1^{\mathcal{S}}, y_1^{\mathcal{S}}), ..., (X_{Ns}^{\mathcal{S}}, y_{Ns}^{\mathcal{S}})\}$, and the target samples as $\mathcal{T} = \{X_1^{\mathcal{T}}, , ..., X_{N\tau}^{\mathcal{T}}\}$. In each domain, x is denoted as a clip of the entire video X. Denote the feature extractor as $\phi_m(\cdot)$, and $m \in \{\mathbb{S}, \mathbb{T}\}$ is the modality of the input frames, with $\mathbb{S}$ of spatial RGB and $\mathbb{T}$ of temporal optical flow.

3.1. Spatio-temporal Contrastive Learning

To conduct a video-based self-supervision task for UDA, there are several key points to consider: (1) The built task is helpful to promote the network to mine the essential video-level representation of various inputs and spatio-temporal association for video classification. (2) For domain adaptation objective, the proposed SSL framework should be able to optimize the model to reduce the domain shift between source and target in feature space, and tries to learn the representation in shared feature distribution. (3) The SSL task should not bias the objective DA task overly, *i.e*., action recognition, and it provides a classification-aware structure basis for the design to optimize video DA system. Taking above points into consideration, we propose a spatio-temporal contrastive learning (STCL) network for self-supervised action recognition.

Unlike CMC [32] and IIC [31], we design a decoupling SSL task for symmetrical modeling in spatial and temporal dimensions from RGB and optical flow individually due to the different characteristics of the features. We involve the clip-level and video-level contrastive losses for local-global temporal content SSL expression, respectively. Especially, we expand the positive-negative selection for each level. The positive clip samples are leveraged with the help of video-level samples to improve the modeling robustness from the disturbance of individual clips. At the clip level, the positive samples are corresponding frames in another modality or video aggregation in the same modality with correct time order, and the negative samples are with wrong time order or wrong pose frame. Similarly, at the video level, the positive sample is the corresponding video in another modality, and the negative samples are the aggregation samples with at least one negative clips. The intra-negative samples on the clip and video levels are selected as Figure 3. Formally, the clip-level feature is denoted as v and $v = \phi_m(x)$, and the video-level feature V is the aggregation of multiple clip feature $\{v_i\}$. The spatial and temporal

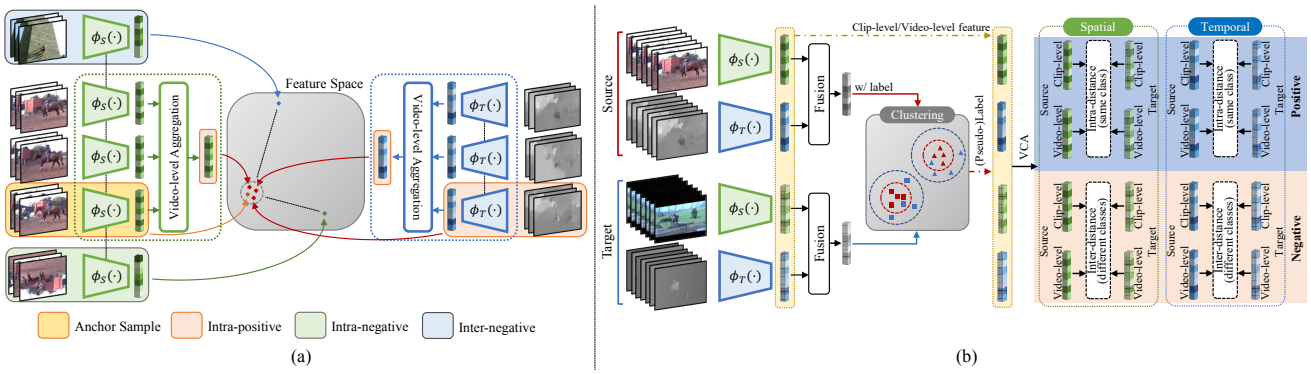

Figure 2. The details of STCDA with two key operations. (a) Spatio-temporal Contrastive Learning (STCL) to the essential video-level representation by self-supervision (Use the RGB modality as the example). (b) Video-based contrastive alignment (VCA) to bridge the domain shift by video-level category-aware measurement.

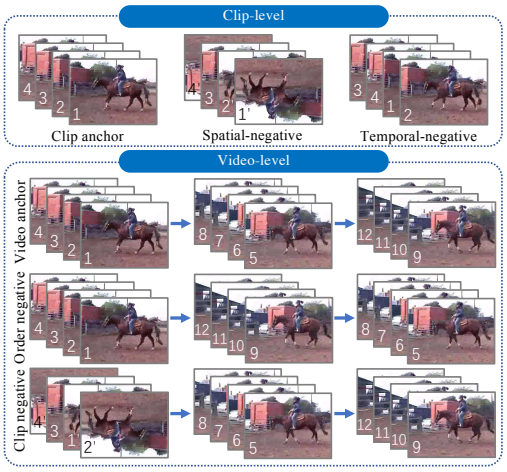

Figure 3. Intra-negative samples at the clip/video level. At the clip level, 'Spatial-negative' is the negative sample with wrong pose frames, like random horizontal/vertical flip and rotation of some frames in the clip, shown as 1', 2', and 4', and 'Temporal-negative' is with the wrong time order. At the video level, 'Order negative' is with wrong clip order in the video, and 'Clip negative' is with at least one negative clip.

networks are the two views of contrastive learning here, and the clip-level and video-level losses are defined as follows respectively (Take the first view (v^1 and V^1) as the sample):

$$\mathcal{L}_{clip}^{v_i^1} = -\log \frac{\Theta(\{v_i^1, V^1\}) + \Theta(\{v_i^1, v_i^2\})}{\Theta(\{v_i^1, V^1\}) + \sum_{j=1}^{k+1} \Theta(\{v_i^1, v_j^2\}) + \sum_{j=1}^{k+1} \Theta(\{v_i^1, v_j^{neg}\})}, \tag{1}$$

$$\mathcal{L}_{vid}^{V^1} = -\log \frac{\Theta(\{V^1, V^2\})}{\Theta(\{V^1, V^2\}) + \Theta(\{V^1, V^{neg}\})}, \tag{2}$$

where v^{neg} and V^{neg} are intra-negtive features at the clip level and the video level, respectively. The measurement $\Theta(\{\cdot, \cdot\})$ is defined as follows:

$$\Theta(\{v_i^1, v_j^2\}) = \exp\left(\frac{v_i^1 \cdot v_j^2}{\|v_i^1\| \cdot \|v_j^2\|} \cdot \frac{1}{\tau}\right), \tag{3}$$

where τ is a scalar hyper-parameter. The contrastive loss for unlabeled videos $X \in \{\mathcal{S} \cup \mathcal{T}\}$ is combined with the clip-level and video-level terms as follows:

$$\mathcal{L}_{Contrast} = \mathop{\mathbb{E}}_{X} \sum_{m \in \{\mathbb{S}, \mathbb{T}\}} \left[\mathop{\mathbb{E}}_{i}\left(\mathcal{L}_{clip}^{v_i^m}\right) + \mathcal{L}_{vid}^{V^m}\right]. \tag{4}$$

3.2. Video-based Contrastive Domain Alignment

To reduce the domain gap in distributions of different domains, the network should be able to capture the content and build up the relevance for generalization by UDA. However, self-supervision task is class-agnostic without any annotation. To compensate the representation on recognition for the network, we propose a video-level category-aware distance metric, *i.e.*, video-based contrastive alignment (VCA), to map the source and target data to a unified feature space, and measure the inter-intra distance with the auxiliary of the class label on source data and pseudo-label on target data. This operation aims to jointly maximize the discrepancy and separate the representations from different categories, and minimize the discrepancy and compact the features from the same class in the feature space of clip/video samples.

Similar to MMD [20], VCA defines the difference of features extracted from source and target in the reproducing kernel Hilbert space (RKHS). Formally, denote κ_m as the kernel of the feature from the spatial/temporal network. Here clip-level and video-level features share the same kernel in each modality m. Given a pair of video

data and the labels (or pseudo-labels)(X, y), for domain $d_1, d_2 \in \{\mathcal{S}, \mathcal{T}\}$ (allow that $d_1 = d_2$), define a relation of (X^{d_1}, y^{d_1}) and (X^{d_2}, y^{d_2}) as follows:

$$r\left((X^{d_1}, y^{d_1}), (X^{d_2}, y^{d_2}); c_1, c_2\right) = \sum_{i=1}^{N^1}\sum_{j=1}^{N^2} \frac{\mu_{c_1c_2}(y_i^{d_1}, y_j^{d_2})}{\sum_{i=1}^{N^1}\sum_{j=1}^{N^2}\mu_{c_1c_2}(y_i^{d_1}, y_j^{d_2})} \cdot \Big[\kappa_m(v_i^{d_1}, v_j^{d_2}) + \kappa_m(V_i^{d_1}, V_j^{d_2})\Big], \tag{5}$$

where c_1 and c_2 are two classes. $c_1 = c_2$ and $c_1 \neq c_2$ means the intra-category and inter-category settings, respectively. And $\mu_{c_1c_2}(y_1, y_2)$ is defined for the judgement of the prediction as follows:

$$\mu_{c_1c_2}(y_1, y_2) = \mathbf{1}_{\left[\substack{y_1=c_1 \\ y_2=c_2}\right]} = \begin{cases} 1, \text{ if } y_1 = c_1, y_2 = c_2, \\ 0, \text{ otherwise.} \end{cases} \tag{6}$$

Considering that the target labels are not available, the evaluation of the target label $\hat{y}$ is essential through iterative optimization. Specially, both the spatial and temporal videos are annotated by the pseudo-labels from the decision of spatio-temporal feature fusion, and then K-means algorithm is adopted for the clustering of target samples individually, in each modality and attach corresponding labels. Formally, for modality m, the cluster center of target data $C_m^{\mathcal{T}}$ is initialized with the source data category center $C_m^{\mathcal{S}}$, of which features fused from spatial and temporal feature extraction, *i.e.*, $C_m^{\mathcal{T}} \leftarrow C_m^{\mathcal{S}} = \sum_{i=1}^{N^{\mathcal{S}}} \mathbf{1}_{[y_i^{\mathcal{S}}=c]} \frac{V_i}{||V_i||}$, where c is the labeled categories. Then the iteration for optimizing the pseudo-labels is as follows during the training process. (1) The given pseudo-label $\hat{y}^{\mathcal{T}}$ is updated as $\hat{y}^{\mathcal{T}} \leftarrow \arg\min_c \sum_m \Phi(V^{\mathcal{T}}, C_m^{\mathcal{T}})$, which jointly employs the spatial and temporal representation for video-level prediction robustness. The operator $\Phi(\cdot, \cdot)$ is the distance measurement in the feature level, and here we use the Euclidean distance, *i.e.*, $\Phi(a, b) = ||a - b||_2$. (2) Because of the feature space alignment from STCL between source and target, the target cluster center $C_m^{\mathcal{T}}$ is updated by the target feature expression as $C_m^{\mathcal{T}} \leftarrow \sum_{i=1}^{N^{\mathcal{T}}} \mathbf{1}_{[\hat{y}_i^{\mathcal{T}}=c]} \frac{V_i}{||V_i||}$ till the convergence or the iteration ending.

According to the predicted pseudo-labels, the distance calculation of video-level domain discrepancy is defined as follows:

$$\begin{aligned} \mathcal{D}(\hat{y}^{\mathcal{T}}; c_1, c_2) = & \, r\left(\{X^{\mathcal{S}}, y^{\mathcal{S}}\}, \{X^{\mathcal{S}}, y^{\mathcal{S}}\}; c_1, c_1\right) \\ & + r\left(\{X^{\mathcal{T}}, \hat{y}^{\mathcal{T}}\}, \{X^{\mathcal{T}}, \hat{y}^{\mathcal{T}}\}; c_2, c_2\right) \\ & - 2r\left(\{X^{\mathcal{S}}, y^{\mathcal{S}}\}, \{X^{\mathcal{T}}, \hat{y}^{\mathcal{T}}\}; c_1, c_2\right). \end{aligned} \tag{7}$$

Considering the inter-intra contrastive setting for optimizing the distribution of category-aware features, the definition of VCA is calculated as follows:

$$\mathcal{D}_{VCA} = \mathop{\mathbb{E}}_{c} \mathcal{D}(\hat{y}^{\mathcal{T}}; c, c) - \mathop{\mathbb{E}}_{c}\left[\mathop{\mathbb{E}}_{c' \neq c} \mathcal{D}(\hat{y}^{\mathcal{T}}; c, c')\right]. \tag{8}$$

3.3. Overall Objective

For labeled source data, we train the network as the traditional supervision task through minimizing the cross-entropy loss for classification as follows:

$$\mathcal{L}_{CE} = - \mathop{\mathbb{E}}_{x^{\mathcal{S}}} \sum_{n=1}^{N^{\mathcal{S}}} \Big[y^{\mathcal{S}} \log \phi_m(x^{\mathcal{S}})\Big]. \tag{9}$$

Therefore, the overall loss can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{Contrast} + \beta \mathcal{D}_{VCA}, \tag{10}$$

where α and β are the parameters to balance the weights of each term.

4. Experiments

4.1. Datasets

We evaluate our approach on four DA datasets: UCF–HMDB$_{small}$, UCF–Olympic, UCF–HMDB$_{full}$ and EPIC Kitchens (D1, D2 and D3).

UCF–Olympic and **UCF–HMDB$_{small}$**. These two datasets are small-scale with a higher degree of distinction between the categories. UCF–Olympic [15] have 6 shared classes from UCF50 and Olympic datasets, and UCF–HMDB$_{small}$ [30] have 5 shared classes from UCF101 and HMDB51.

UCF–HMDB$_{full}$. UCF–HMDB$_{full}$ [4] is a complementary version of UCF–HMDB$_{small}$, with 12 shared categories from UCF101 and HMDB51, respectively.

EPIC Kitchens. EPIC Kitchens [23] is a fine-grained action recognition dataset of three domain partitions (D1, D2, and D3) with 8 categories in varying amounts. It contains different actions with fine details under the view of the first person in indoor kitchen scenes.

4.2. Implementation Details

We use the BN-Inception [14] and I3D [2] architectures as the backbone feature extractors for each clip of both source and target videos for different datasets, and the two backbones are initialized with the ImageNet [8] dataset and the Kinetics dataset [2] pre-trained models, respectively. The size of input clips is 16 frames with 224×224 pixels. Specially, the channel numbers of RGB and optical flow frame stacks are 3 (Red, Green and Blue) and 2 (u and v), respectively. For the video level, the segment number of clips per video is set to 3 for both training and testing.

For self-supervised spatio-temporal contrastive learning, we adopt the memory bank training scheme following CMC

[32], which stores the latent features to the memory bank for each training input sample, to retrieve and compare positive/negative samples efficiently from the memory bank buffer, without extra cost on feature extraction computation.

We train our proposed STCDA framework in three stages as follows: (1) We firstly train the feature extractor of RGB and optical flow with STCL under the contrastive loss in Eq. 4, based on both source and target data without any labels. (2) After contrastive learning, we further train the model with cross-entropy loss and contrastive loss with labeled source data and unlabeled target data, where the hyper-parameters are set to $\alpha = 2$ and $\beta = 0$ in Eq. 10. (3) We further train the full model with cross-entropy loss, contrastive loss and VCA, with the data used as step (2), where the hyper-parameters are set to $\alpha = 2$ and $\beta = 0.5$ in Eq. 10. The optimizer is stochastic gradient descent (SGD) with momentum of 0.9 to train the network, and the weight decay is set to 10^{-7}. The batch size is set to 64, and the model is trained in 300, 300, and 400 epochs at three stages, respectively. The learning rate starts from 0.01, 0.01, and 0.001, respectively, and it is divided by 10 after 100 epochs and 200 epochs.

4.3. Comparison with Stat-of-the-art Methods

We compare our proposed network with state-of-the-art approaches for UDA on video action recognition, *e.g.*, DAAA [15], TA^{3}N [4], TCoN [25], SAVA [7] and MM-SADA [23] in several benchmarks — UCF–HMDB$_{small}$ and UCF–Olympic in Table 1, UCF–HMDB$_{full}$ in Table 2, and EPIC Kitchens in Table 3. In Table 1 and Table 2, "Source only" and "Target only" mean the network is trained with labeled source data (lower bound) and labeled target data (upper bound), respectively. For a fair comparison, note that the modality of "RGB" and "R" in each table means the network is only trained from RGB frames, without optical flow used. As shown in these tables, we can see that our proposed STCDA obtains the stat-of-the-art preformance in various scenarios.

UCF–HMDB$_{small}$ and UCF–Olympic. We utilize BN-Inception architecture as the feature extractor. With the strong representation capability of CNNs, our STCDA achieves promising performance with 2D CNN in these two small-scale datasets, *i.e.*, BN-Inception against 3D CNN (C3D). In particular, we obtain stat-of-the-art results on HMDB→UCF with accuracy of 100% and UCF→Olympic with accuracy of 98.1%.

UCF–HMDB$_{full}$. BN-Inception and I3D are used for feature extraction. Compared with other methods, STCDA achieves stat-of-the-art results on HMDB→UCF with accuracy of 91.9% with only RGB modality used. In addition, with efficient multi-modal contrastive learning and feature fusion, STCDA obtains higher performance with accuracy of 83.1% and 92.1% on UCF→HMDB and HMDB→UCF, respectively. Besides, we can observe that the larger network architecture with 3D modeling (I3D) would obtain the higher accuracy than the 2D CNN (BN-Inception) with large margins, *e.g.*, 83.1% vs. 80.0% on UCF→HMDB and 92.1% vs. 87.7% on HMDB→UCF, even though the upper bounds of these two networks are similar. Compared with 2D CNNs, 3D CNNs can directly model spatio-temporal video clips with richer representation and larger reception field in space and time domain, which are significantly useful for video understanding tasks.

EPIC Kitchens. We utilize I3D as the backbone following MM-SADA [23]. In Table 3, we can see that STCDA obtains the stat-of-the-art results with the mean accuracy on six domain settings of 51.2%, achieving 0.9% improvement

Table 1. Comparison of accuracy (%) on UCF–HMDB$_{small}$ and UCF–Olympic.

Method	Backbone	U→H	H→U	U→O	O→U
DAAA [15] (R)	C3D	—	—	91.6	90.0
TA^{3}N [4] (R)	R-TRN	**99.3**	99.5	**98.1**	92.9
TCoN [25] (R+F)	BNIncep	—	96.8	95.8	94.1
TCoN [25] (R+F)	C3D	—	—	95.9	94.8
TCoN [25] (R)	B-TRN	—	—	96.8	**96.8**
Source only (R)	BNIncep	94.7	97.7	91.6	90.4
STCDA (R)	BNIncep	97.3	99.3	94.4	93.3
Target only (R)	BNIncep	98.7	99.5	96.3	98.3
Source only (R+F)	BNIncep	96.7	99.3	94.4	92.9
STCDA (R+F)	BNIncep	98.7	**100**	**98.1**	96.3
Target only (R+F)	BNIncep	100	100	98.1	100

"R" and "F" denote the RGB and optical flow modalities. "R-TRN" and "B-TRN" denote ResNet-101-based TRN and BN-Inception-based TRN respectively, and "BN-Incep" denotes BN-Inception.

Table 2. Comparison of accuracy (%) on UCF–HMDB$_{full}$.

Method	Backbone	Pre-train	U→H	H→U
Source only (R)	R-TRN	ImgNet	71.7	73.9
TA^{3}N [4] (R)	R-TRN	ImgNet	78.3	81.8
Target only (R)	R-TRN	ImgNet	82.8	94.9
TCoN [25] (R)	R-TRN	ImgNet	87.2	89.1
Source only (R)	I3D	K400	80.3	88.8
SAVA [7] (R)	I3D	K400	82.2	91.2
Target only (R)	I3D	K400	95.0	96.8
Source only (R)	BNIncep	ImgNet	74.1	82.5
STCDA (R)	BNIncep	ImgNet	76.9	85.1
Target only (R)	BNIncep	ImgNet	91.7	94.7
Source only (R+F)	BNIncep	ImgNet	76.1	85.8
STCDA (R+F)	BNIncep	ImgNet	80.0	87.7
Target only (R+F)	BNIncep	ImgNet	94.2	96.8
Source only (R)	I3D	K400	80.8	88.4
STCDA (R)	I3D	K400	81.9	91.9
Target only (R)	I3D	K400	94.4	96.3
Source only (R+F)	I3D	K400	82.8	89.8
STCDA (R+F)	I3D	K400	**83.1**	**92.1**
Target only (R+F)	I3D	K400	95.8	97.7

"R" and "F" denote the RGB and optical flow modalities. "R-TRN" denotes ResNet-101-based TRN, and "BNIncep" denotes BN-Inception. "ImgNet" indicates ImageNet dataset, and "K400" indicates Kinetics400 dataset.

Table 3. Comparison of accuracy (%) on EPIC Kitchens. In each modality, STCDA enables the network to achieve better performance.

Method	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	Mean
AdaBN [19]	44.6	47.8	47.0	54.7	40.3	48.8	47.2
MMD [20]	43.1	48.3	46.6	55.2	39.2	48.5	46.8
MCD [27]	42.1	47.9	46.5	52.7	43.5	51.0	47.3
MM-SADA (RGB) [23]	41.7	42.1	45.0	48.4	39.7	46.1	43.9
MM-SADA (Flow) [23]	45.0	45.7	49.0	58.9	44.8	52.1	49.3
MM-SADA (RGB + Flow) [23]	48.2	50.9	49.5	56.1	44.1	**52.7**	50.3
STCDA (RGB)	44.4	41.1	47.7	45.5	41.2	47.6	44.6
STCDA (Flow)	45.3	52.2	45.1	**59.5**	44.0	51.2	49.6
STCDA (RGB + Flow)	**49.0**	**52.6**	**52.0**	55.6	**45.5**	52.5	**51.2**

Table 4. Ablation study of accuracy (%) on hyper-parameters on UCF–HMDB$_{full}$.

Method	α	β	U→H	H→U
Source only	0	0	76.1	85.8
STCL	1	0	77.5	86.2
STCL	2	0	**78.1**	**86.5**
VCA	0	0.5	**78.3**	**86.9**
VCA	0	1	77.8	86.5
STCL+VCA	2	0.5	**80.0**	**87.7**

Table 5. Ablation study of accuracy (%) on spatio-temporal contrastive learning (STCL) on UCF–HMDB$_{full}$.

Method	STCL on Source	STCL on Target	U→H	H→U
Source only	×	×	76.1	85.8
VCA	×	×	78.3	86.9
Clip+VCA	✓	✓	79.7	87.2
Video+VCA	✓	✓	79.4	87.0
Clip+Video+VCA	✓	×	79.2	87.2
Clip+Video+VCA	×	✓	78.9	87.0
Clip+Video+VCA	✓	✓	**80.0**	**87.7**

than MM-SADA. In each individual modality, STCDA also enables the network to achieve better performance with the mean accuracy of 44.6% and 49.6% in RGB and optical flow, respectively.

4.4. Ablation Study

We leverage ablation experiments on UCF–HMDB$_{full}$, to analyze contributions of each component for UDA.

Hyper-parameters. As shown in Table 4, we evaluate the hyper-parameters α and β in Eq. 10. According to the contributions of STCL and VCA individually used, we set the trade-off weights of α and β to 2 and 0.5, respectively, which obtains the best performance with the accuracy of 80.0% on UCF→HMDB and 87.7% on HMDB→UCF.

Contribution of STCL. As shown in Table 5, the results show the effect of STCL. "Source only" is trained on the source data without adaptation components, from the BN-Inception backbone within multi-modal setting. Here we perform the experiments on clip-level and video-level contrastive learning. Besides, with the flexible structure of STCL, we compare the results using different input data with the combination of source and target optionally. The combination of clip-level and video-level contrastive learning obtains significant improvement over source only by 1.7% (78.3% to 80.0%).

Table 6. Comparison of accuracy (%) effect on (a) video aggregation in STCL, (b) video fusion in VCA, and (c) different kinds of pseudo-labels used in VCA on UCF–HMDB$_{full}$.

Method		U→H	H→U
Source only		76.1	85.8
(a)	Mean	79.7	87.4
	Concat + FC	79.2	87.6
	GRU	**80.0**	**87.7**
(b)	w/ adversarial loss	77.7	86.5
	w/o spatio-temporal fusion	78.3	87.4
	w/ spatio-temporal fusion	**79.7**	**87.6**
(c)	w/ fixed pseudo-labels	76.9	86.2
	w/ updated pseudo-labels	**78.3**	**86.9**

Besides, there are three candidate options for video aggregation operation, *i.e.*, *Mean* for calculating the average value for each clip feature, *Concat+FC* for concatenating each clip feature and feeding to a fully-connected layer, and *GRU* for recurrent modeling on each clip feature by Gated Recurrent Unit (GRU) with 512 hidden units, and obtaining the output from the last GRU node. The accuracy comparison of these operators is indicated in Table 6(a) We can see that the recurrent module of GRU is effective for temporal aggregation and video-level representation on self-supervised contrastive learning. Note that *Mean* aggregation is not used with time-order negative samples for contrastive learning.

Contribution of VCA. For video-based contrastive distance metric, the spatio-temporal fusion is important for robust video prediction of pseudo-labels, which are allocated to both spatial and temporal inputs for higher confidence in clustering. In Table 6(b), we observe that the cross-modal video fusion is beneficial for better representation,

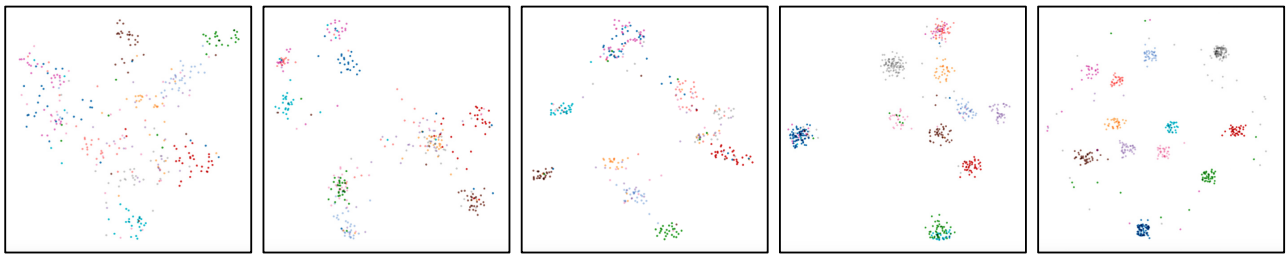

Figure 4. Visualization of t-SNE on HMDB→UCF, and the stages are (a) source only; (b) w/ STCL; (c) w/ VCA; (d) the final result; (e) target only. Features are extracted from the last fully-connected layer.

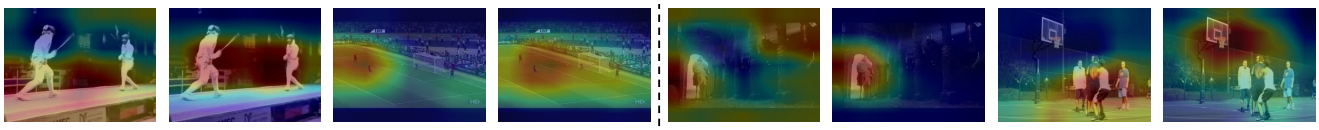

Figure 5. Visualization of Grad-CAM [28] on UCF–HMDB$_{full}$. Examples are sampled from (a) UCF and (b) HMDB datasets.

even than the adversarial losses used individually with the gradient reversal layer (GRL).

Furthermore, we evaluate the clustering effect with different ways of pseudo-labels generation in Table 6(c). "fixed pseudo-labels" means that the pseudo-labels of target is obtained from the initial clustering with $C_m^{\mathcal{S}}$ and fixed in the training process. With the iterative updated pseudo-labels, the prediction of target data would be accurate with better performance than the fixed pseudo-labels, where the proposed VCA is efficient to adapt the target data with higher robustness.

4.5. Visualization

We visualize the distribution of learned features by t-SNE [12] embedding. As shown in Figure 4, we can observe that components of STCL and VCA achieve the contributions of mixed clusters in each class (shown in different colors) (Figure 4(b)(c)), for video-level feature alignment and classification generalization. Particularly, in Figure 4(d), our proposed framework leads to the discriminative distribution of target data, with the similar distribution of supervised target only setting in Figure 4(e). However, in terms of the limitation of confused category-aware clustering, the similar action would be closer, *e.g.*, *ride_bike* and *ride_horse* marked in green and cyan at the bottom in Figure 4(d). And it is meaningful to explore effective class-aware discrimination algorithms for accurate classification.

Furthermore, we indicate some samples of target videos and predictions in Figure 5. Grad-CAM [28] is used to present the activation region for the video under the prediction. The visualization results show that the network focuses on the irrelevant scene or objects without any domain adaptation modules, while it pays more attention to key actions with the proposed STCDA framework, which aims at transferring local-global temporal content learning, *e.g.*, in Figure 5(a) right and Figure 5(b) right, the baseline network focuses on the persons, while STCDA makes a decision based on the persons and the discriminative scene/object (the goal in *kick_ball* and the rim in *shoot_ball*).

5. Conclusion

In this paper, we propose a self-supervised contrastive network for videos, *i.e.*, spatio-temporal contrastive learning (STCL), for learning joint clip-level and video-level representations to improve the generalization of local-global temporal content modeling. Besides, we propose the video-based contrastive alignment (VCA) for multi-modal domain metric to measure the video-level discrepancy between source and target domains. Our spatio-temporal contrastive domain adaptation (STCDA) framework with STCL and VDA achieves stat-of-the-art results on several UDA benchmarks of action recognition, *e.g.*, UCF–HMDB, UCF–Olympic and EPIC Kitchens. Furthermore, we will explore the video-level spatio-temporal interaction for UDA, and extend STCDA to other cross-domain video tasks.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61771339.

References

[1] Qi Cai, Yingwei Pan, Chong Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *CVPR*, 2019. 1

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 5

[3] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *ICCV*, 2019. 1

[4] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*, 2019. 2, 3, 5, 6

[5] Yang Chen, Yingwei Pan, Ting Yao, Xinmei Tian, and Tao Mei. Mocycle-gan: Unpaired video-to-video translation. In *ACMMM*, 2019. 3

[6] Yun Chun Chen, Yen Yu Lin, Ming Hsuan Yang, and Jia Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *CVPR*, 2019. 1

[7] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *ECCV*, 2020. 3, 6

[8] Jia Deng, Wei Dong, Richard Socher, Li Jia Li, and Fei Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[9] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. 2016. 2

[10] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 3

[11] Zhenwei He and Lei Zhang. Multi-adversarial faster-rcnn for unrestricted object detection. In *ICCV*, 2019. 1

[12] G. E. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9(2):2579–2605, 2008. 8

[13] Han Kai Hsu, Chun Han Yao, Yi Hsuan Tsai, Wei Chih Hung, and Ming Hsuan Yang. Progressive domain adaptation for object detection. In *WACV*, 2020. 1

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. 2015. 5

[15] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, 2018. 2, 3, 5, 6

[16] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, 2019. 3

[17] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017. 2

[18] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 2, 3

[19] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *PR*, 80:109–117, 2018. 7

[20] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 3, 4, 7

[21] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. 2017. 3

[22] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, 2016. 2

[23] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020. 3, 5, 6, 7

[24] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. 2016. 2

[25] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *AAAI*, 2020. 3, 6

[26] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 1

[27] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 7

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020. 8

[29] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NIPS*, 2014. 2

[30] J. Tang, H. Jin, S. Tan, and D. Liang. Cross-domain action recognition via collective matrix factorization with graph laplacian regularization. volume 55, pages 119–126, 2016. 5

[31] Li Tao, Xueting Wang, and Toshihiko Yamasaki. Self-supervised video representation learning using inter-intra contrastive framework. In *ACMMM*, 2020. 2, 3

[32] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2, 3, 6

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 2

[34] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 1

[35] 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 *ECCV*, 2016. 2

[36] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. 2

[37] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *ICCV*, 2015. 3

[38] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *CVPR*, 2019. 3

[39] Jing Zhang, Wanqing Li, and Philip Ogunbona. Joint geometrical and statistical alignment for visual domain adaptation. In *CVPR*, 2017. 1