Modality-Collaborative Test-Time Adaptation for Action Recognition

Baochen Xiong[1,2,3], Xiaoshan Yang[1,2,3], Yaguang Song[2], Yaowei Wang[2], Changsheng Xu[1,2,3*]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA)
[2]Peng Cheng Laboratory, China
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
bcxiong@yeah.net, xiaoshan.yang@nlpr.ia.ac.cn, songyg01@pcl.ac.cn, wangyw@pcl.ac.cn, csxu@nlpr.ia.ac.cn

Abstract

Video-based Unsupervised Domain Adaptation (VUDA) method improves the generalization of the video model, enabling it to be applied to action recognition tasks in different environments. However, these methods require continuous access to source data during the adaptation process, which are impractical in real scenarios where the source videos are not available with concerns in transmission efficiency or privacy issues. To address this problem, in this paper, we focus on the Multimodal Video Test-Time Adaptation (MVTTA) task. Existing image-based TTA methods cannot be directly applied to this task because videos have domain shifts in multimodal and temporal, which brings difficulties to adaptation. To address the above challenges, we propose a Modality-Collaborative Test-Time Adaptation (MC-TTA) Network. MC-TTA contains maintain teacher and student memory banks respectively for generating pseudo-prototypes and target-prototypes. In the teacher model, we propose Self-assembled Source-friendly Feature Reconstruction (SSFR) to encourage the teacher memory bank to store features that are more likely to be consistent with the source distribution. Through multimodal prototype alignment and cross-modal relative consistency, our method can effectively alleviate domain shift in videos. We evaluate the proposed model on four public video datasets. The results show that our model outperforms existing state-of-the-art methods.

1. Introduction

Action recognition is a very challenging task, which requires complex motion analysis of video sequence information. It has broad application prospects and important research value in security monitoring, health management, smart home and other fields. With the development of multimodal technology, multimodal data fusion is a promising way to solve this task. In particular, RGB provides rich scene context information, while optical flow captures key motion attributes that complement the visual features for more accurate action understanding.

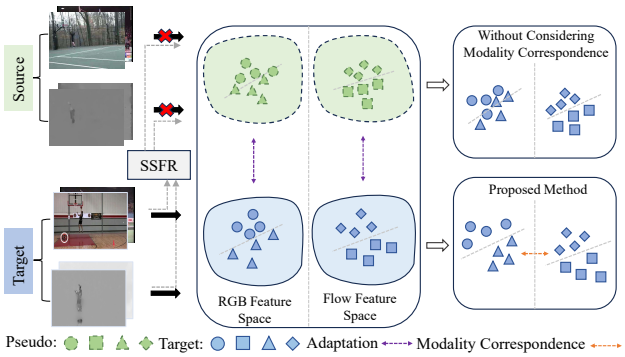

Figure 1. Main idea of the proposed Modality-collaborative Test-Time Tdaptation (MC-TTA). Only pre-trained source model and unlabeled target videos can be used for target model learning. We propose Self-assembled Source-friendly Feature Reconstruction (SSFR) module to construct pseudo-source domain features from target domain. In addition, modality correspondence is explored to maintain good discriminability for modalities susceptible to domain shifts.

Deep neural network has made significant progress in the supervised learning of large-scale labeled data, and it has achieved excellent performance in the task of action recognition [15, 37, 42, 46, 47]. However, when there is a distribution shift between the data of the test environment (i.e. the target domain) and the data of the training environment (i.e. the source domain), the multimodal action recognition model is more vulnerable to the impact of distribution changes [3]. Therefore, it is very important for the model to quickly adapt to the new multimodal data during the test to obtain better performance, i.e., test-time adaptation (TTA). This is different from the usual domain adaptive action recognition task settings [8, 11], which can access source data during training. In TTA, we only have access to model parameters pre-trained on the source data and then

*Corresponding author.

use unlabeled data on the target domain for fast adaptation. This usually refers to one training epoch, which is practical for real-world scenarios where the action recognition model needs to run online with minimal delay under strict hardware constraints [25].

Although the existing TTA-based action recognition methods have made important progress [25], they only consider the data with one modality in the source and target domains, such as RGB. Due to the multimodal and temporal characteristics of video, aligning multimodal video data without source domain data is more challenging. As shown in Figure 1, the domain shifts of different modalities are always diverse, which poses challenges to maintain the discriminability when adapting the model. For example, optical flow modality is more domain-invariant for action recognition [20, 28] in changing backgrounds compared with visual modality, which contains more domain-specific semantic information of action performers and context. Therefore, in video-based TTA task, it is very important to comprehensively consider the domain shifts of different modalities. Although multimodal alignment schemes have been widely studied in traditional video domain adaptation methods [33, 48], existing methods are not suitable for video-based TTA task because they cannot quickly eliminate multimodal domain shifts when the source video is unseen.

In this paper, we propose a Modality-Collaborative Test-Time Adaptation (MC-TTA) network, to solve the Multimodal Video Test-Time Adaptation task (MVTTA). Specifically, **in the pre-training step**, we utilize labeled source videos to separately learn the corresponding feature extractor and classifier for each modality, and the multimodal classifier for the fused multimodal learning. **In the adaptation step**, we construct the teacher and student models of the target domain through pre-trained model. The teacher model maintains a teacher memory bank to create pseudo-prototypes representing the pseudo-source domain feature distribution. The student model maintains a student memory bank to create target-prototypes representing the target domain feature distribution. We can mitigate domain shift by reducing the difference between the pseudo-source and target distributions. In the teacher model, we propose Self-assembled Source-friendly Feature Reconstruction (SSFR) module to encourage the teacher memory bank to store features that are more likely to be consistent with the source distribution. Specifically, to imitate the source distribution, SSFR utilizes the consistency and confidence scores of logits predicted for different modalities of a target video by the source classifiers to find the video clips that are more similar to source videos. Then, the features of selected clips are aggregated to represent the source-friendly features. Next, we use multimodal prototype alignment to push the target-prototypes closer to the pseudo-prototypes to reduce multimodal domain shift. Due to the lack of supervision in the target domain data, it may lead to the decrease of discriminability in adaptation. Therefore, we propose a cross-modal relative consistency loss to leverage the correspondence between modalities to maintain good discriminability for the modality that is susceptible to the domain shift. We conduct extensive experiments on four public video datasets, UCF–HMDB$_{small}$ [19], UCF–Olympic [36], UCF-HMDB$_{full}$ [8] and Epic-Kitchens [13]. The results demonstrate that the proposed MC-TTA achieves the state-of-the-art performance on the MVTTA task.

Our main contributions are summarized as follows:
(1) We propose Self-assembled Source-friendly Feature Reconstruction (SSFR) module to identify source-friendly video clips to generate pseudo-source distributions. (2) We propose multimodal prototype alignment and cross-modal consistency constraint to efficiently alleviate the domain shift among multimodal videos. (3) We evaluate the proposed method on four datasets and demonstrate its effectiveness with extensive experimental results.

2. Related Work

2.1. Action Recognition

Video action classification is more challenging than image recognition due to the high complexity of video data. TSN [39] obtains video-level representation by utilizing 2D convolutional networks in spatial and temporal dimensions, and then fusing features. TRN [49] generates video-level feature vectors through the temporal transformation and dependence of frames at different time scales. Another representative method, I3D [6], extends the idea of dual-stream networks by using 3D convolution kernels on RGB and optical flow. Recently, transformer-based models have also been applied to video recognition [1, 26, 29, 44]. For example, ViViT [1] adds several temporal transformer encoders based on the spatial encoder. Different from our work, all the above studies solve the traditional supervised action recognition problem (without domain shift).

2.2. Video Unsupervised Domain Adaptation

Video unsupervised domain adaptation(VUDA) aims to learn a model of labeled video samples from the source domain that generalizes well on the target domain with large distribution shift [16, 21, 43, 45]. VUDA research lags behind image-based UDA research, mainly due to the challenges brought by video temporal and multimodal. However, with the introduction of various cross-domain video datasets such as UCF-HMDB$_{full}$ [8] and Epic-Kitchens [13], there has been a significant increase in research interests for VUDA [9, 12, 27]. TA3N [8] uses an integrated temporal relation module that can simultaneously learn temporal dynamics and achieve domain align-

ment. TCoN [32] uses a deep architecture with a cross-domain attention module to match the distribution of temporally aligned features between source and target domains. SAVA [11] used an attention mechanism to determine discriminative clips and used this information for video-level alignment within an adversarial learning framework. MM-SADA [28] uses a domain adaptation method based on self-supervision and multi-modal learning (RGB+optical flow) for fine-grained first-person view action recognition. Although VUDA methods bring improvements in video model robustness, all these methods require access to source data during the adaptation process. Given the amount of private information surrounding the topics and scenes in the video, such a request could raise serious privacy concerns.

2.3. Test-time Adaptation

Test-time adaptation (TTA) is designed to enable existing models to quickly adapt to new target data without accessing source domain data. As an important challenge to deal with dynamic domain transfer in the real world, TTA has received increasing attention in many tasks [5, 7, 10, 23, 25, 34]. Tent [38] uses minimizing entropy to update trainable parameters in the batch normalization layer. SHOT [24] uses entropy minimization and diversity regularizers for test-time adaptation. LAME [4] uses Laplacian adjusted maximum likelihood estimation to adjust the output of the model rather than the parameters. There are some works [31, 40] combing test time adaptation and continual learning to maintain the performance on the source domain. TSD [41] uses test-time self-distillation to make the target features as consistent as possible during adaptation. MM-TTA [34] uses two complementary modules to obtain and select more reliable pseudo labels (from 2D and 3D modalities) as self-learning signals during TTA. ViTTA [25] is similar to our task, aligning the statistical data at different temporal augmentations of the same video consistent with the statistical data seen during training. Compared with the above work, we have studied similar TTA settings, but in the different context of using multimodal video action recognition, we consider the challenges of temporal and multi-modal in video.

3. Methodology

3.1. Problem Definition

In Multimodal Video Test-Time Adaptation(MVTTA) task, we use $D_s = \{x_s^i, y_s^i\}_{i=1}^{n_s}$ to denote the source domain dataset, where x_s^i denotes a multimodal video instance and n_s is the number of source video instances. $y_s^i \in \mathbb{R}^C$ is the corresponding class label, where C denotes the total number of classes. In addition, we denote the target domain by $D_t = \{x_t^i\}_{i=1}^{n_t}$, where n_t is the number of unlabelled video instances. We use the labelled target domain video instances only for evaluation. D_s and D_t have the same underlying label distribution, but belong to different data distributions. To comprehensively capture the important information of action recognition, we use the RGB modality and optical flow modality in video. For each multimodal video instance x, we first segment it into T equal-length clips. We can obtain T RGB clips and optical flow clips. The goal of MVTTA is to adapt the multimodal video classification model based on the unlabeled target domain video with the help of the pre-trained model in the source domain. It is worth noting that we can only access target domain unlabeled video instances D_t and the source model in an online manner.

3.2. Modality-Collaborative Test-Time Adaptation

To solve the MVTTA task, we propose a Modality-Collaborative Test-Time Adaptation (MC-TTA) network, as shown in Figure 2. To reduce the domain shift between the target domain and the source domain, we maintain separate memory banks for the teacher and student models. The teacher memory bank is used to generate pseudo-prototypes, and the student memory bank is used to generate target-prototypes. The purpose of the pseudo-prototype is to imitate the feature distribution of the pseudo-source domain, while the purpose of the target prototype is to represent the feature distribution of the target domain, reducing domain differences through alignment between prototypes. We hope that the teacher memory bank stores source-friendly features to make the pseudo-prototype closer to the source domain. We propose Self-assembled Source-friendly Feature Reconstruction (SSFR) module to encourage the teacher memory bank to store features that are more likely to be consistent with the source distribution. However, directly minimizing the feature distance between pseudo-prototypes and target-prototypes will lose the discriminative information. Due to the lack of supervision in the target domain data, it may lead to the decrease of discriminability in adaptation. Therefore, we propose a cross-modal relative consistency loss to leverage the correspondence between modalities to maintain good discriminability for the modality that is susceptible to the domain shift.

3.2.1 Network Architecture

We follow the setting of test-time adaptation, where we are not able to access the source data but only the source pre-trained action recognition classification model. This model consists of two branches and a multimodal classifier. Each branch includes a feature extraction network ϕ^r/ϕ^o and a classifier ψ^r/ψ^o for feature encoding and single-modality classification of RGB/optical flow. In addition, the multimodal classifier ψ^m achieves the final multimodal classification of video by concatenating the feature represen-

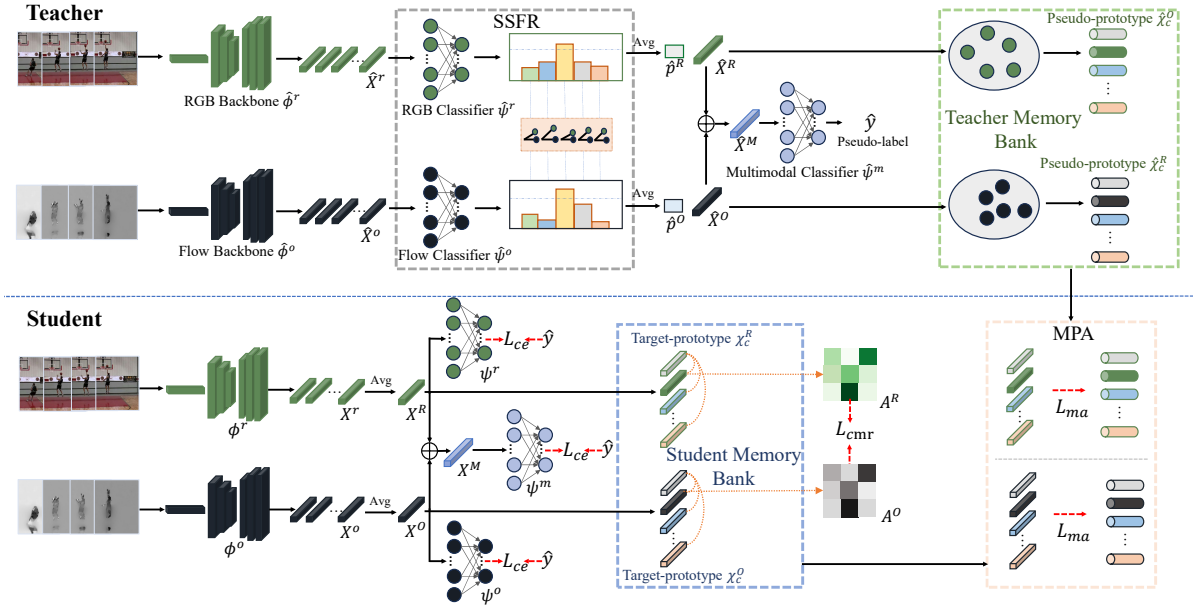

Figure 2. Overview of the proposed Modality-Collaborative Test-Time Adaptation (MC-TTA) network. SSFR:Self-Assembled Source-Friendly Feature Reconstruction. MPA:Multimodal Prototype Alignment.

tations obtained from the two branches. We use the ϕ^r and ϕ^o networks to extract the RGB clip-level features $X^r = [r_1, r_2, ..., r_T]^T$ and optical flow clip-level features $X^o = [o_1, o_2, \ldots, o_T]^T$ for all video clips, where T represents the number of video clips, $r_i, o_i \in \mathbb{R}^d$ and $X^r, X^o \in \mathbb{R}^{T\times d}$.

At the beginning of training, the the teacher model ($h_{\hat{\Theta}_t}$) and the student model (h_{Θ_t}) share the same weight, i.e., Θ_t=$\hat{\Theta}_t$=Θ_s, where Θ_s represents the source domain model parameters, Θ contains the parameters of ϕ^r, ϕ^o, ψ^r, ψ^o, and ψ^m.

3.2.2 Self-Assembled Source-Friendly Feature Reconstruction

In our method, the teacher model maintain a teacher memory bank that stores source-friendly features to simulate pseudo-source domain feature distribution. For each video clip, the high correlation between the predictions of the two modalities indicates that the source domain model performs similar judgments in predicting each modality. This shows the relationship between the two modalities of this clip is closer to the feature distribution of the source domain. Furthermore, we select predictions with low entropy, as lower entropy typically signifies that the target clip features are closer to the source domain feature distribution [18, 41]. Therefore, their features are more conducive for constructing pseudo-source domain features. Based on the above purposes, we propose Self-Assembled Source-Friendly Feature Reconstruction (SSFR) module. SSFR utilizes the consistency and confidence scores of logits predicted for different modalities of a target video by the source classifiers to find the video clips that are more similar to source videos. Then, the features of selected clips are aggregated to represent the source-friendly features. Specifically, each branch uses the teacher feature extractor $\hat{\phi}_t^r$/$\hat{\phi}_t^o$ to extract the features $\hat{X}_t^r$/$\hat{X}_t^o$ of the multimodal video instance in the target domain. Before averaging the clip-level features, each clip features input to the teacher single-modality classifier $\hat{\psi}_t^r$/$\hat{\psi}_t^o$ to obtain the clip-level logits. For each clip of RGB and optical flow logits, we calculate the modality correlation cosine distance between the two modalities:

$$d = cos(\hat{p}_t^r, \hat{p}_t^o), \tag{1}$$

Where $\hat{p}_t^r$ is RGB modality clip-level logits, and $\hat{p}_t^o$ is optical flow modality clip-level logits. d represents the similarity of the two modality classification results.

For each clip, we calculate the clip-level logits confidence score $conf\,(\hat{p}_t)$ of the teacher single-modality classification results:

$$conf\,(\hat{p}_t) = -\hat{p}_t^\top \log \hat{p}_t, \tag{2}$$

where $\hat{p}_t \in (p_t^r, p_t^o)$. Finally, we determine the clip is source-friendly through modality correlation and confidence score:

$$\mu = \begin{cases} 1, & \text{if } conf\,(\hat{p}_t) \geq \beta \text{ and } d \leq \alpha \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where α and β are the thresholds. For each clip of the target domain video, when the modality correlation cosine distance is less than α and the confidence score is greater than the β (i.e., $\mu = 1$), we consider this clip is source-friendly. By merging the video-level source-friendly features of both modalities, we obtain the multimodal feature representation $\hat{X}_t^M = concat(\hat{X}_t^R, \hat{X}_t^O)$, where $\hat{X}_t^R/\hat{X}_t^O$ means averaging all source-friendly clip features of the RGB/optical flow modality. Then, we can obtain the target video instances pseudo-labels $\hat{y} = \arg\max \hat{p}_t^m$, where $\hat{p}_t^m = \hat{\psi}_t^m(\hat{X}_t^M)$.

In the process of adaptation, given an unlabeled multimodal video target instance, we can obtain the video-level RGB and optical flow feature and logits through the teacher model. It is worth noting that the features and logits are averaged from the source-friendly clips obtained by the above method. We maintain a teacher memory bank $\hat{M}_t = \{(\hat{X}_t^R, \hat{p}_t^R)(\hat{X}_t^O, \hat{p}_t^O)\}$ to store the source-friendly features and logits of target domain video instances. Following T3A [18] and TSD [41], the teacher memory bank is initialized with the weights of the source single-modality classifier ψ_s^r and ψ_s^o. Through the teacher memory bank to build up the relations between the current instances and all of the previous instances, the pseudo-prototypes shall be generated for each class.

The prototype of class c can be formulated as:

$$\hat{\mathcal{X}}_c^R = \frac{\sum_i \hat{X}_{t,i}^R 1(\hat{y}_i = c)}{\sum_i 1(\hat{y}_i = c)}, \hat{\mathcal{X}}_c^O = \frac{\sum_i \hat{X}_{t,i}^O 1(\hat{y}_i = c)}{\sum_i 1(\hat{y}_i = c)}, \quad (4)$$

where $1(\cdot)$ is an indicator function, output value 1 if $\hat{y}_i = c$ or 0 otherwise. We hope to generate class prototypes from the teacher memory bank that have high similarity to source domain features. With the input of memory bank, we make entropy judgments based on the stored video-logits. To further improve the similarity, we select the top-K low entropy feature average to obtain pseudo-prototypes.

3.2.3 Relationship-Aware Multimodal Adaptation

In the teacher model, we propose self-assembled source-friendly feature reconstruction to target domain videos, extracting source-friendly video-level features that are stored in the teacher memory bank to simulate the distribution of source domain features.

To reduce domain shift, we need to push target domain features toward source domain features. Therefore, we maintain a student memory bank $M_t = \{X_t^R, X_t^O\}$ in the student model to store target domain features, where X_t^R/X_t^O represent the video-level RGB/optical flow features obtained by averaging all clip-level features. When each X_t^R/X_t^O comes, we average it with the corresponding class prototype to obtain new class target-prototype $\mathcal{X}_c^R/\mathcal{X}_c^O$, where the pseudo-label of X_t^R/X_t^O is generated by the teacher model. Compared to the teacher memory bank, the student memory bank differ in two aspects: (1) Its purpose is to store video-level features of target domain videos to obtain target-prototypes. These video-level features are obtained by averaging the features from all clips within each video; (2) It is initialized with zero weights.

Multimodal Prototype Alignment. To clearly reduce the multimodal domain shift between the source video domain and the target video domain, we need to push the target domain features to the source distribution. Since the teacher memory bank is composed of prototypes close to the distribution of source domain features, we can regard it as a pseudo-source domain for domain adaptation. Then we use a multimodal prototype alignment loss to align the class prototypes of the two modalities in the two memory banks, as follows:

$$\mathcal{L}_{ma} = \sum_{i=1}^{C} (\left\| \hat{\mathcal{X}}_i^R - \mathcal{X}_i^R \right\|_2 + \left\| \hat{\mathcal{X}}_i^O - \mathcal{X}_i^O \right\|_2), \quad (5)$$

where $\mathcal{X}_i^R$ and $\mathcal{X}_i^O$ represent the RGB prototype and optical flow prototype of i-th class in the student memory bank. With this constraint, the multimodal video feature can learn to explicitly reduce the domain shift.

Cross-Modal Relative Consistency. The domain shift of different modalities are always diverse, in MVTTA task, it is important to comprehensively consider the domain shift of different modalities. During the adaptation stage, the target domain data lacks supervisory information, which may lead to reduced class discriminability. We need to take advantage of the correspondence between modalities and let the modality with good discriminability guide the modality with serious loss of discriminability. For example, in scenes with fast motion or large background changes, the optical flow modality will perform more stably and robustly, while in scenes with rich textures and details, the visual information provided by the RGB modality will be more stable. Therefore, we propose a cross-modal relative consistency loss to leverage the correspondence between modalities to maintain good discriminability for the modality that is susceptible to the domain shift. Different from multimodal prototype alignment, this module focuses on modality collaboration, making it still have good discriminability after adaptation.

Specifically, we calculate the Euclidean distance between prototypes of each class in the student memory bank M_t and constructed a relationship matrix between prototypes as follows:

$$A_{ij}^R = E(\mathcal{X}_i^R, \mathcal{X}_j^R), A_{ij}^O = E(\mathcal{X}_i^O, \mathcal{X}_j^O), \quad (6)$$

where E$(\cdot)$ is the Euclidean distance formula, $A^R, A^O \in \mathbb{R}^{C \times C}$ represent the relationship matrices of RGB and optical flow respectively. $\mathcal{X}_i$, $\mathcal{X}_j$ is the prototype of class i and class j in student memory bank respectively.

We then implement adaptation of the relationships between modalities via consistency loss between the A^R and A^O relationship matrices:

$$\mathcal{L}_{cmr} = KL\left(A^R \| A^O\right) + KL\left(A^O \| A^R\right). \quad (7)$$

where KL(·) denotes the Kullback-Leibler divergence.

Classification constraint. In addition to the above operations, we compute the cross-entropy loss for two single-modality and one multimodal classifiers using pseudo-labels generated by the teacher model:

$$\mathcal{L}_{\text{CE}} = -\hat{y}(\log\sigma(p_t^R) + \log\sigma(p_t^O) + \log\sigma(p_t^m)), \quad (8)$$

where σ denotes the softmax operation. p_t^R/p_t^O is the prediction from the student single-modality classifier ψ_t^r/ψ_t^o, and p_t^m is the prediction from the student multimodal classifier ψ_t^m. Minimizing $\mathcal{L}_{\text{CE}}$ enhances consistency between teacher and student predictions.

3.2.4 Training Objective Function

Combing Eq.5, Eq.7, and Eq.8, we formulate the final objective function as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1\mathcal{L}_{ma} + \lambda_2\mathcal{L}_{cmr} \quad (9)$$

where λ is the trade-off parameter to balance different loss functions.

After the update of the student model h_{Θ_t} ($\Theta_{t,j} \rightarrow \Theta_{t,(j+1)}$), we update the weights of the teacher model $h_{\hat{\Theta}_t}$ using the student weights by exponential moving average (EMA):

$$\hat{\Theta}_{t,(j+1)} = \gamma\hat{\Theta}_{t,j} + (1-\gamma)\Theta_{t,(j+1)}, \quad (10)$$

where γ is a smoothing factor that controls the degree of change we require at each update.

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments on the following four common benchmarks in this field. (1) **UCF–Olympic** has 6 shared classes, respectively from the UCF101 dataset [35] and the Olympic dataset [30]. Following [17], we use a 7:3 train-test split, which results in 432/168 train/test action videos for the UCF domain and 260/55 train/test action videos for the Olympic domain. (2) **UCF–HMDB**$_{small}$ has 5 shared classes, respectively from the UCF101 dataset [35] and HMDB51 dataset [22], which contains 432/168 train/test action videos for the UCF domain and 482/189 train/test action videos for the HMDB domain. (3) **UCF–HMDB**$_{full}$ is one of the most widely used cross-domain video data sets. It contains a total of 3209 videos in 12 action class, with 2 cross-domain action recognition tasks. All videos are from the UCF101 [35] dataset and HMDB dataset [22]. We follow the split provided by [8], which results in 1438/571 train/test action videos for the UCF domain and 840/360 train/test action videos for the HMDB domain. (4) **Epic-Kitchens** is a more challenging dataset, which is a fine-grained action recognition dataset collected from a first-person view in a kitchen scene [13]. Following [28], we conducted experiments on three domain partitions (D_1, D_2, and D_3) of the 8 largest action classes. It contains 2495/313 train/test action videos on D_1, 1543/417 train/test action videos on D_2, and 3897/1030 on D_3 train/test action videos.

Table 1. Results on UCF–Olympic, UCF–HMDB$_{small}$ and UCF–HMDB$_{full}$ datasets.

	UCF–Olympic		UCF–HMDB$_{small}$		UCF–HMDB$_{full}$	
Methods	U→O	O→U	U→H	H→U	U→H	H→U
Source-only	92.73	90.48	93.65	94.05	81.39	85.29
Tent [38]	92.73	92.26	94.71	94.64	83.89	86.16
LAME [4]	92.73	93.45	95.24	95.23	84.44	87.39
T3A [18]	94.55	94.05	96.30	95.83	84.72	86.87
TSD [41]	92.73	94.05	95.76	97.02	85.27	87.92
ViTTA [25]	**94.55**	**95.23**	97.35	97.62	86.39	89.14
MM-TTA [34]	92.73	92.86	95.76	95.23	85.56	88.97
MC-TTA	**94.55**	**95.23**	**97.88**	**98.21**	**88.61**	**91.07**

Baseline. We compared MVTTA with start-of-the-art TTA methods, i.e., Tent [38], LAME [4], T3A [18], TSD [41], ViTTA [25], and MM-TTA [34]. Since existing TTA methods cannot be directly used in multimodal video scenarios, for a fair comparison, we combine the same model architecture with different TTA methods. In addition, we also compared Source-only, which means that only the source model is used for prediction without adaptation.

Implementation details. Given the success of video classification by CNNs, we use the I3D [6] architectures as the backbone feature extractors for each clip of both source and target videos for different datasets, and the backbone are initialized with the ImageNet [14] dataset and the Kinetics dataset [2] pre-trained models, respectively. Each video clip consists of 8 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. The number of clips T is set to 10. For RGB/optical flow single-modality classifier and multimodal classifier, we define them as two-layer perceptrons with ReLU activation functions and Softmax outputs. It should be noted that in the adaptation step, the parameters of the last classification layer are fixed, with the aim of aligning the representation of the target video with the source distribution, so that the domain shift is reduced. We extract the clip features with the same dimension (i.e., d=1024) for different modalities. We set the hidden layer dimension of all classifiers to 1024.

For all baselines, we use the publicly released code, and

Table 2. Results on Epic-Kitchens dataset.

Methods	$D_1 \rightarrow D_2$	$D_1 \rightarrow D_3$	$D_2 \rightarrow D_1$	$D_2 \rightarrow D_3$	$D_3 \rightarrow D_1$	$D_3 \rightarrow D_2$	Mean
Source-only	37.32	45.92	43.80	42.07	50.75	32.86	42.12
Tent [38]	36.60	45.72	43.96	41.43	49.87	32.49	41.68
LAME [4]	37.79	46.97	44.36	42.43	51.67	33.41	42.77
T3A [18]	39.05	47.93	45.24	42.96	51.61	33.96	43.46
TSD [41]	38.45	47.10	45.16	42.78	50.83	33.06	42.90
ViTTA [25]	38.99	48.80	**45.81**	43.30	52.26	34.62	43.96
MM-TTA [34]	37.93	46.55	44.87	42.36	50.30	32.82	42.47
MC-TTA	**40.84**	**50.73**	45.46	**43.94**	**53.68**	**36.46**	**45.19**

Table 3. Ablation studies on four datasets.

SSFR	Loss		UCF–Olympic		UCF–HMDB$_{small}$		UCF–HMDB$_{full}$		Epic-Kitchens
	$\mathcal{L}_{ma}$	$\mathcal{L}_{cmr}$	U→O	O→U	U→H	H→U	U→H	H→U	Mean
×	×	×	92.73	91.67	94.18	94.05	84.17	86.87	42.36
✓	×	×	92.73	92.26	94.71	94.64	85.00	87.22	42.65
✓	✓	×	92.73	94.05	96.30	97.02	86.94	88.44	43.96
✓	×	✓	94.55	94.64	96.83	97.62	87.50	89.31	44.52
✓	✓	✓	**94.55**	**95.23**	**97.88**	**98.21**	**88.61**	**91.07**	**45.19**

all the extra hyper-parameters involved in the compared methods use their best settings. Our model and baselines are all trained with SGD optimizer, where the weight decay is set to $1e-4$ and the momentum is set to 0.9. In the self-assembled source-friendly feature reconstruction module, the value of α is set to 0.3, and the value of β is set to 0.6. The average number of features K selected by the teacher memory bank is set to 5. On all datasets, the learning rate and the batch size are set to 0.01 and 64, respectively. The balance weights λ_1 and λ_2 of the loss function are set to 1.0 and 0.5.

4.2. Comparison with State-of-the-art Methods

We first reported the results obtained by comparing our method with state-of-the-art methods. Table 1 and table 2 respectively show our performance in UCF–Olympic, UCF–HMDB$_{small}$, UCF–HMDB$_{full}$, and Epic-Kitchens. In all tables, the best results are presented in bold.

As shown in table 1, the proposed MC-TTA is competitive compared to other state-of-the-art methods on these three datasets. The accuracy of the results in U→O and O→U is 94.55% and 95.23%, respectively. In addition, on the UCF-HMDB$_{small}$ dataset, the U→H and H→U settings performed better than all other methods, with accuracy rates of 97.88% and 98.21%, respectively. The UCF-HMDB$_{full}$ dataset has greater domain shift compared to the previous two datasets, making it more effective to evaluate the effectiveness of different methods in solving MVTTA task. Compared with other TTA methods, MC-TTA achieved better performance in U→H and H→U settings, with improvements of 2.22% and 1.93%, respectively. Because the proposed method can explicitly reduce the domain shift of multimodal information in the source and target video domains.

As shown in Table 2, since the Epic-Kitchens dataset is more challenging than the other three datasets, all methods cannot achieve high accuracy. MC-TTA achieves the best accuracy results on 5 domain adaptation tasks. On the $D_1 \rightarrow D_3$ and $D_3 \rightarrow D_2$ settings, the accuracy of MC-TTA is 1.93% and 1.84% higher than ViTTA. The mean accuracy of MC-TTA on the 6 domain adaptation tasks is 45.19%, outperforming the second-best ViTTA method by 1.23%.

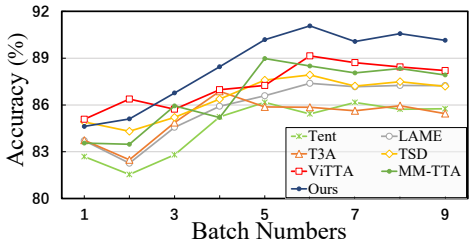

Figure 3. Visualization of the accuracy of different methods during adaptation on the UCF–HMDB$_{full}$ dataset (target domain:UCF). The batch numbers indicate the batch of videos for which the model has been updated.

It is worth noting that although MM-TTA [34] is a TTA method designed for multimodal, its performance is not ideal in temporal data scenarios. In addition, the effectiveness of the ViTTA [25] method is limited on datasets with large domain shift.

We also visualized the accuracy changes of different methods throughout the adaptation process, as shown in Figure 3. We can see that our method can adjust data faster and achieve higher accuracy in the target domain.

4.3. Ablation Study

In this section, the effectiveness of the proposed modality-collaborative test-time adaptation network is further evaluated by analyzing the impact of three key components of MC-TTA (i.e., self-assembled source-friendly feature re-

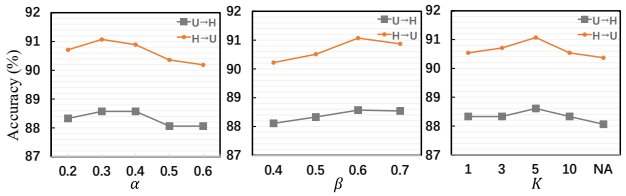

Figure 4. Sensitivity analysis about the two hyperparameters α and β in the SSFR module, and the number of Top-K features for the average prototype in the teacher memory bank.

construction (SSFR), multimodal prototype alignment, and cross-modal relative consistency) on four datasets. The ablation experiments are shown in Table 3. We find that our method performs better than source-only when the three components are not used, which illustrates the effectiveness of the self-distillation architecture. When we increase SSFR, we observe that model performance improves on all three datasets. This shows that this SSFR can first alleviate the problem of pseudo-label noise in multimodal videos. We found that the SSFR module is not effective on the UCF-Olympic datasets. The reason may be that the performance of the two small datasets is somewhat saturated. When using multimodal prototype alignment or cross-modal relative consistency schemes can improve the performance on four datasets. We observe that cross-modal relative consistency loss performs better on the UCF-HMDB$_{full}$ dataset than on the Epic-Kitchens dataset. We think that the UCF-HMDB$_{full}$ dataset has large differences in motion between different classes, making it easier to constrain the correspondence between modalities. Moreover, when using both these two schemes, the performance of our MC-TTA can be further improved. The above results demonstrate the importance of three components in our method.

4.4. Further Remarks

Sensitivity to hyperparameter. We analyzed three hyperparameters in our method on the UCF-HMDB$_{full}$ dataset: the two hyperparameters α and β in the self-assembled source-friendly feature reconstruction (SSFR) method, and the number of Top-K features for the average pseudo-prototypes in the teacher memory bank. Figure 4 shows the hyperparameter sensitivity analysis. Firstly, for the SSFR module, good performance is achieved when the cosine distance $d \in [0.2, 0.4]$ of the two modality prediction results and the prediction confidence score $conf\,(\hat{p}_t) \in [0.5, 0.7]$. However, $d \geq 0.4$ or $conf\,(\hat{p}_t) \leq 0.5$ results in performance degradation. We speculate that too large cosine distance and too small confidence score will not only increase the probability of pseudo-labels prediction error, but also select non-source-friendly clips. Then, we analyze the number of features $K \in (1, 3, 5, 10, \text{NA})$ in the teacher memory bank for constructing pseudo-prototypes, where NA denotes no feature filtering. We can observe that good performance is achieved when $K \in \{3, 5\}$. As K increases, using too many features may reduce the accuracy of pseudo-prototypes.

SSFR threshold selection strategy. In the proposed MC-TTA, we propose the SSFR module, SSFR utilizes the consistency and confidence scores of logits predicted for different modalities of a target video by the source classifiers to find the video clips that are more similar to source videos. In SSFR, we manually adjust the threshold to ensure the performance of the model. In addition, we designed an adaptive update threshold method that takes the cosine distance and confidence score average of a batch data as the threshold. This method eliminates the need for hyperparameter adjustment and makes the source-friendly selection process adaptive. As shown in Table 4, the adaptive results are not as effective as manually setting hyperparameters. We speculate that the mean is greatly affected by negative instances, leading to a decrease in performance.

Table 4. Hyperparameter Selection Strategy.

Selection Strategy	UCF–HMDB$_{full}$		Epic-Kitchens
	U→H	H→U	Mean
Adaptive	88.06	90.72	44.72
Hard(MC-TTA)	**88.61**	**91.07**	**45.19**

5. Conclusions

In this paper, we propose a simple and effective Modality-Collaborative Test-Time Adaptation (MC-TTA) to solve the multimodal video test-time adaptation (MVTTA) task, where only pre-trained source model and unlabeled target videos are available for learning the multimodal video classification model. In the adaptation stage, we construct the teacher and student models of the target domain through pre-trained source model. We maintain teacher and student memory banks respectively for generating pseudo-prototypes and target-prototypes. We propose Self-assembled Source-friendly Feature Reconstruction (SSFR) module to encourage the teacher memory bank to store features that are more likely to be consistent with the source distribution. We use multimodal prototype alignment to push the target-prototype closer to the pseudo-prototype to reduce multimodal domain shift. We propose a cross-modal relative consistency loss to maintain good discriminability for the modality that is susceptible to the domain shift. Extensive experimental results demonstrate the effectiveness of the proposed method. In future work, we would like to extend our MC-TTA to other applications, such as video segmentation and video retrieval.

Acknowledgments. This work was supported by National Natural Science Foundation of China (No.62036012, U23A20387, 62322212, 62072455), and was also supported by National Science and Technology Major Project(2021ZD0112200).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 6

[3] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021. 1

[4] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022. 3, 6, 7

[5] Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3582–3591, 2023. 3

[6] 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. 2, 6

[7] Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. Improved test-time adaptation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24172–24182, 2023. 3

[8] 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. 1, 2, 6

[9] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020. 2

[10] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9137–9146, 2021. 3

[11] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 678–695. Springer, 2020. 1, 3

[12] Victor G Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Dual-head contrastive domain adaptation for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1181–1190, 2022. 2

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 2, 6

[14] 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. 6

[15] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 1

[16] Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, and Bo Li. Dual alignment unsupervised domain adaptation for video-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18962–18972, 2023. 2

[17] Yi Huang, Xiaoshan Yang, Ji Zhang, and Changsheng Xu. Relative alignment network for source-free multimodal video domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1652–1660, 2022. 6

[18] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021. 4, 5, 6, 7

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

[20] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021. 2

[21] Wouter M Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*, 43(3):766–785, 2019. 2

[22] 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*, pages 2556–2563. IEEE, 2011. 6

[23] Yizhuo Li, Miao Hao, Zonglin Di, Nitesh Bharadwaj Gundavarapu, and Xiaolong Wang. Test-time personalization with a transformer for human pose estimation. *Advances in Neural Information Processing Systems*, 34:2583–2597, 2021. 3

[24] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020. 3

[25] Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video test-time adaptation for action recognition. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22952–22961, 2023. 2, 3, 6, 7

[26] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2

[27] Djebril Mekhazni, Maximilien Dufau, Christian Desrosiers, Marco Pedersoli, and Eric Granger. Camera alignment and weighted contrastive learning for domain adaptation in video person reid. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1624–1633, 2023. 2

[28] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 122–132, 2020. 2, 3, 6

[29] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3163–3172, 2021. 2

[30] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*, pages 392–405. Springer, 2010. 6

[31] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 3

[32] 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*, pages 11815–11822, 2020. 3

[33] Mohamed Ragab, Emadeldeen Eldele, Wee Ling Tan, Chuan-Sheng Foo, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Adatime: A benchmarking suite for domain adaptation on time series data. *ACM Transactions on Knowledge Discovery from Data*, 17(8):1–18, 2023. 2

[34] Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16928–16937, 2022. 3, 6, 7

[35] 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. 6

[36] 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. 2

[37] Thanh-Dat Truong, Quoc-Huy Bui, Chi Nhan Duong, Han-Seok Seo, Son Lam Phung, Xin Li, and Khoa Luu. Direcformer: A directed attention in transformer approach to robust action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20030–20040, 2022. 1

[38] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 3, 6, 7

[39] 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. 2

[40] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 3

[41] Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20050–20060, 2023. 3, 4, 5, 6, 7

[42] Zhengwei Wang, Qi She, and Aljosa Smolic. Action-net: Multipath excitation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13214–13223, 2021. 1

[43] Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 2023. 2

[44] Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. Client-adaptive cross-model reconstruction network for modality-incomplete multimodal federated learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1241–1249, 2023. 2

[45] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, and Zhenghua Chen. Source-free video domain adaptation by learning temporal consistency for action recognition. In *European Conference on Computer Vision*, pages 147–164. Springer, 2022. 2

[46] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022. 1

[47] Xiaoshan Yang, Baochen Xiong, Yi Huang, and Changsheng Xu. Cross-modal federated human activity recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[48] Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees GM Snoek. Audio-adaptive activity recognition across video domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13791–13800, 2022. 2

[49] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018. 2