Synthesizing Videos from Images for Image-to-Video Adaptation

Anonymous Author(s)
Submission Id: 912*

ABSTRACT

We address the image-to-video adaptation task that aims to leverage labeled images and unlabeled videos for video recognition. There are two major challenges in this task, including the domain discrepancy between the two domains, and the modality gap between the image and video modalities. Existing methods mainly employ a two-stage paradigm by first adopting frame-level adaptation to reduce the domain discrepancy and then learning a spatio-temporal model to bridge the modality gap. In this paper, we provide a new perspective and propose a single-stage method that synthesizes video from the source static image and converts the image-to-video adaptation problem into a video-to-video adaptation problem. With the synthesized video, we present a simple baseline that a spatio-temporal model is trained with cross entropy loss with source labels and the Batch Nuclear norm Maximization to encourage the classification responses of target videos maintain the discriminability and diversity. We further propose a new pseudo label generation method that inherits the robustness of class prototype and the effectiveness of the small loss criterion. Based on the constructed baseline and the proposed pseudo label generation method, we are able to train a model that achieves state-of-the-art performances or gets comparable performances on three standard benchmarks.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Image-to-Video Adaptation, Domain Adaptation, Transfer Learning, Video Classification

ACM Reference Format:

Anonymous Author(s). 2023. Synthesizing Videos from Images for Image-to-Video Adaptation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Video classification is crucial for many multimedia applications such as video search, recommendation systems, and content-based retrieval. Training a deep video classification model that performs well typically relies on a substantial amount of labeled videos, but this can be impractical due to the labor-intensive and time-consuming process of collecting and annotating them. To address this issue, the research community and industry have shown interest in alternative methods such as zero-shot video classification [5], few-shot video classification [44], and semi-supervised video classification [38]. However, these methods still require large amounts of labeled videos of seen categories and support sets, or they may result in inferior performance. As an alternative, researchers have explored image-to-video adaptation techniques [2, 23, 40] that leverage existing labeled image datasets and unlabeled target videos for video recognition. Image-to-video adaptation is more feasible since it is easier to collect and annotate images than videos, which is the focus of this paper.

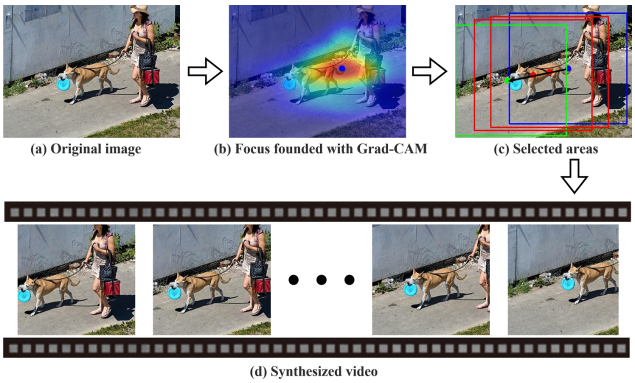


Figure 1: Synthesizing video from a single source static image. Grad-CAM is used to locate the start area which should contain the objects relevant to a category. With randomly selected end area, several intermediate areas along the path of the start and end areas are further selected. All selected areas form a video.

Image-to-video adaptation presents two significant challenges. The first challenge is the domain discrepancy that arises due to differences in background, lighting, image style, and camera perspectives between the source images and target video frames. This domain discrepancy causes a decline in the performance of well-trained source models over target video frames. The second challenge is the modality gap. The modality gap refers to the absence of temporal information in source images, which is present in videos.

Existing methods mainly employ a two-stage paradigm by first adopting frame-level adaptation to reduce the domain discrepancy and then learn a spatio-temporal model to bridge the modality gap. For example, HiGAN [41] and SymGAN [40] train JANs [27] to reduce domain discrepancy and employ generative model [14] for temporal information completion. Kae et al. [19] and Wei et al. [23] both learn a domain invariant frame-level model via domain adversarial learning [11] and then utilize self training via pseudo label on target videos to learn a spatio-temporal model.

Conference acronym 'XX, June 03–05, 2023, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Different from the existing two-stage paradigm, we propose a single-stage method that converts the static image to synthesized video which facilitate the image-to-video adaptation problem into a video-to-video adaptation problem. There are three advantages to our single-stage method. First, the accurate source labels can be directly used to train a spatio-temporal model without modality gap. Second, there are plenty of video based domain adaptation methods and even image based domain adaptation methods could be used to reduce the domain discrepancy without handling the modality gap. Third, single-stage method is much simpler as both stages of learning in existing methods require much effort like exploring hyperparameters.

To synthesize a video, we crop several areas of a static source image and treat these areas as frames. To maintain the semantic of the frames, we adopt Grad-CAM [31] to locate the major objects for a video category. The area that contains semantic objects is regarded as the start frame of the video and we randomly select another area as the end frame. We further randomly select several areas as intermediate frames, whose centers are along the path between centers of the start area and end area, to simulate camera movement and camera zooming in the video. Such synthesizing method is very simple and effective as the generated video preserves the semantic information for training a discriminative spatio-temporal model.

With the synthesized source videos and the unlabeled target video, we construct a simple baseline that a spatio-temporal model is trained with cross entropy loss and the typical Batch Nuclear norm Maximization (BNM) loss [6]. The cross entropy loss is computed with the labeled synthesized source video, and the BNM loss is employed over unlabeled target videos. Minimizing the BNM loss encourage the model's classification responses over target videos to maintain the discriminability and diversity. Nevertheless, the temporal information of synthesized video may not be appropriate for distinguishing categories like "run" and "walk" where both categories exhibit similar spatial appearance. To address this issue, we employ self training on target videos with their pseudo labels since the temporal information of target videos are reliable.

To reduce the negative influence of noise in pseudo labels, we first construct prototype for each class and then refine the pseudo labels with the help of the prototypes. As the prototype is more robust to noise, the refined pseudo labels are more reliable. To further reduce the adverse impact of noisy labels on training the spatio-temporal model, we resort to small loss criterion [42] to select parts of the target data as reliable samples. We conduct extensive experiments to evaluate our method that integrates the proposed self training into the constructed baseline. The experimental results show that our single-stage method performs favorably against the current state-of-the-art method. To summarize:

- We propose a single-stage method that synthesizes video from the source static image and convert the image-to-video adaptation problem into video-to-video adaptation problem. We present a simple spatio-temporal baseline that is trained with cross entropy loss and the BNM loss to encourage the classification responses on target videos maintain the discrimination and diversity.
- We propose a new pseudo label generation method that inherit the robustness of prototype and the effectiveness of small loss criterion. Such a generation method enables the model to be trained with more accurate supervision and achieves high performance.
- Based on the constructed baseline and the proposed pseudo label generation technique, our method outperforms the current state-of-the-art method or gets comparable performances on three standard benchmarks. We provide detailed ablation studies to validate the contributions of the proposed techniques.

2 RELATED WORK

Webly-supervised video classification. Using web images and videos as labeling-free data source to improve the performance of video classification has drawn considerable attention to alleviate the dependence on large amounts of labeled videos [8, 9]. Some approaches rely on the labeled target video and use noisy collected web data to boost the performance of the model. For example, Duan et al. [8] first trained a teacher model on labeled target videos and used the teacher model to filter collected web data of different formats. Then the filtered different formats of web data are transformed into trimmed video clips for training a student model in combination with the labeled target videos. Other methods only use collected web data to train a video classification model [9, 10] with explicitly reducing the domain discrepancy. For example, Gana et al. [9] proposed to mutually filter web images and web video frames by matching the distributions between the selected images and the selected frames via minimizing MMD [16]. Webly-supervised video classification methods are labeling-free and very flexible, but their performances are inferior to those of the supervised methods.

Image-to-video adaptation. Compared to the webly-supervised methods, existing unsupervised image-to-video adaptation methods assume that the source images are correctly labeled and the unlabeled target videos are available. Reducing the domain discrepancy and mitigating the modality gap are the focus of image-to-video adaptation approaches [19, 22, 40]. In [22], the authors assume that the attention map representation of the convolutional layer is more transferable. They define an energy score as the largest local activation over an attention map for a specific class, and the predicted class is inferred as the one with the highest score among all classes. To mitigate the modality gap, Hierarchical GAN [41], Symmetric GAN [40] and spatio-temporal causal graph [2] are proposed to learn the mapping between image features and video features via GAN [14]. Kae et al. [19] employs domain adversarial training to learn a spatial model and copy the weights from the spatial model to the spatio-temporal model for reducing the modality gap. Lin et al. [23] proposed a four-stage method where class-agnostic domain alignment is employed in the first stage and the pseudo labels from the first stage is used to train an independent spatio-temporal model in the second stage. They further alternately conduct spatial alignment and spatio-temporal learning with knowledge transfer to each other in the following two stages. In contrast to these methods, we provide a new perspective and propose to synthesize video from the source static image which convert the image-to-video adaptation problem into video-to-video adaptation problem.

Video-to-video adaptation. In our single-stage method, image-to-video adaptation task is converted into a closed-set video unsupervised domain adaptation (VUDA) task. We review the various deep learning-based closed-set VUDA methods that could be mainly categorized into three categories. **Adversarial-based** methods achieve domain adaptation by obtaining domain-invariant features in an

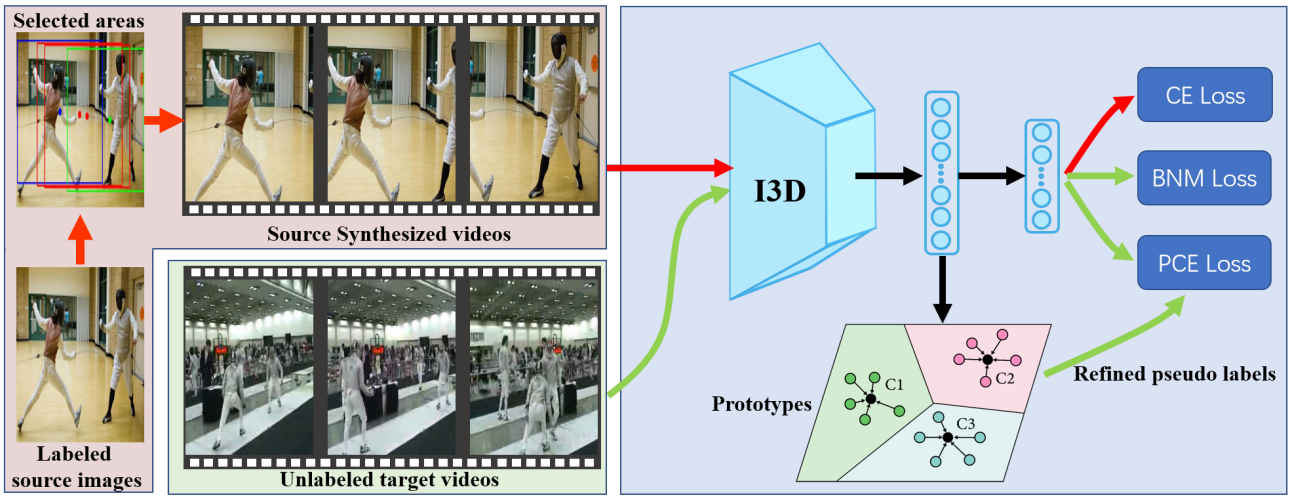


Figure 2: Framework of the proposed method. We first convert labeled source images into videos by selecting several areas of the images. The labeled synthesized source videos and the unlabeled target videos are fed into a I3D model which outputs features and classifier responses. The features of source and target videos are used to construct the class prototype according to the labels of source videos and the pseudo labels of target videos. Then the pseudo labels of target videos are refined according to their distances to the prototypes. The model is trained with cross entropy (CE) loss with source labels, Batch Nuclear norm Maximization (BNM) loss and pseudo cross entropy (PSE) loss with the refined pseudo labels of target videos.

adversarial way, where a feature generator is trained to generate features that the discriminator is not able to distinguish. Based on DANN[12], DAAA [17] applies image extractor to frames sampled from video and achieve good performance. TA^3N [4] focuses on aligning the temporal features of videos obtained by TRN [45]. To avoid the unstable training [30] for adversarial-based methods, **discrepancy-based** methods are proposed to minimize the discrepancy explicitly. For instance, PTC [13] minimizes the MMD [16] loss across both RGB and optical flow modalities to reduce the domain shift and achieves better performances. **Semantic-based** methods aim to extract shared semantics of the two domains. STCDA [32] utilizes a contrastive loss on both the clip and video level such that frames and clips are spatially and temporally associated. STCDA further reduces the domain shift of source and target videos by a video-based contrastive alignment. Recently, DVM [37] leverages MixUp [43] to address the domain-wise gap where the target videos are fused with the source videos progressively on the pixel-level. Different from the video-to-video adaptation task, in our method, the source videos are synthesized from source images whose spatio-temporal information are not that reliable. Therefore, we propose self training according to the refined pseudo labels obtained via constructed prototypes, which can be seen as a semantic-based method.

3 METHODOLOGY

Assuming that there exists a collection of labeled samples $I_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ in the source image domain, and n_t unlabeled videos $V_t = \{v_j^t\}_{j=1}^{n_t}$ in the target video domain. Both domains contain the same C categories, and the aim of the image-to-video adaptation is to utilize the labeled source image domain and the unlabeled target video domain to train a model that can perform effectively in the target domain.

To address the image-to-video adaptation task, as shown in Figure 2, we propose to synthesize video from the source image and convert the image-to-video adaptation problem into video-to-video adaptation problem. Then we train a spatio-temporal model like I3D with cross entropy (CE) loss with source labels, Batch Nuclear norm Maximization (BNM) [6] loss and pseudo cross entropy (PSE) loss. The PSE loss is computed with the refined pseudo labels of target videos according to the constructed class prototypes. The proposed video synthesizing technique, the baseline with BNM loss, and the PSE loss with the refined pseudo labels will be illustrated in the following subsections in detail.

3.1 Synthesizing videos

To mitigate the modality gap between source images and target videos in image-to-video adaptation task, we propose to convert the source images into videos. Specifically, we propose a simple yet effective technique to generate frames from an image by simulating camera movement and camera zooming in the video.

To guarantee that the frames of synthesized videos contain the objects of a certain category, we first utilize the labeled source domain to train a classifier and use Grad-CAM [31] to locate the objects. The position of the highest attention with respect to category y_i^s is regarded as the center c_1 of the start area. We crop an area with the center c_1 whose height and width are set to 2/3 of those for original images. The cropped area is regarded as the start frame of the synthesized video. Then we randomly select a position c_n in the opposite side of the start area, as the center of the end area. Similarly, we crop an area with center c_n as the end frame. We interpolate the path between centers c_1 and c_n and randomly sample $K-2$ centers along the path and crop areas corresponding to these centers as intermediate frames. The generated videos are denoted by $V_s = \{(v_i^s, y_i^s)\}_{i=1}^{n_s}$ whose labels are inherited from $I_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$.

With the synthesized videos from source images, the image-to-video adaptation task can be converted into video-to-video adaptation task. There are two major advantages to synthesizing videos from source images. First, the spatio-temporal model for video recognition pretrained on large scale video dataset like Kinetics [20], can be well utilized for better transfer the knowledge of source labels. Compared with the traditional image-to-video adaptation methods that utilize classification network pretrained on ImageNet, the pretrained spatio-temporal model performs better (see ablation study 4.4) as it is more suitable for video recognition. Second, with the synthesized videos from source images, various well-developed domain adaptation techniques, including image-level domain adaptation techniques like DANN [11], CDAN [25], BNM [6], and MCC [18], can be better explored to improve the performance.

3.2 Simple Baseline

To learn a spatio-temporal model, we choose I3D model Inception v1 [1] pretrained on the Kinetics dataset [20] as our backbone. To ensure the discriminability of the model, we adopt the typical cross-entropy loss for the labeled synthesized source data V_s. Specifically, in I3D model, the cross-entropy loss is applied to both frame-level and video-level. The frame-level classification loss $\mathcal{L}_f$ is defined as:

$$\mathcal{L}_f = \frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K} CE(\hat{p}(v_i^s)_k, y_i^s), \tag{1}$$

where the $CE(\cdot,\cdot)$ is the cross-entropy loss, and $\hat{p}(v_i^s) \in \mathcal{R}^{K\times C}$ is the network prediction over K frames from a synthesized video V_i^s. B denotes the batch size.

The video-level classification loss $\mathcal{L}_c$ is

$$\mathcal{L}_c = \frac{1}{B}\sum_{i=1}^{B} CE(\overline{\hat{p}(v_i^s)}, y_i^s), \tag{2}$$

where $\overline{\hat{p}(v_i^s)} \in \mathcal{R}^{K\times C}$ is an average representation of K frames from video V_i^s.

To encourage the model to acquire discrimination ability over the unlabeled target videos, we resort to Batch Nuclear-Norm Maximization (BNM) [6]. We choose BNM due to its simplicity and effectiveness. Notably, other domain adaptation techniques like DANN [11], CDAN [25], BNM [6], and MCC [18] can also be employed to construct various baselines. By minimizing the BNM loss, the model is encouraged to maximize the rank of the prediction matrix over a batch of data to maintain the prediction discriminability and diversity. Ideally, the full rank of the prediction matrix indicates that all classes are activated, preventing some classes dominate the classification responses. The BNM loss is computed on the matrix of the classification responses for a batch of unlabeled samples, without any supervision as follows:

$$\mathcal{L}_{bnm} = -\frac{1}{B}||p(V)||_\star \tag{3}$$

where the $p(V) \in \mathcal{R}^{B\times C}$ is the output matrix with respect to a batch of target videos V, and B is the batch size. $||\cdot||_\star$ denotes the nuclear-norm, which is the sum of all the singular values of the matrix.

The overall loss for training the baseline model is:

$$\mathcal{L} = \mathcal{L}_c + \lambda_f\mathcal{L}_f + \lambda_b\mathcal{L}_{bnm}, \tag{4}$$

where the λ_f and λ_b are trade-off hyper-parameters.

3.3 Robust self training

As the temporal information of the synthesized video may not be appropriate for distinguishing hard classes, we propose to employ self training of target videos with pseudo labels. The temporal information of the target video is reliable and can be effectively learned for video recognition if the pseudo labels are correct. However, the pseudo labels inevitably contain much noise which leads to performance degradation of the trained model.

To handle the noise in the pseudo label, we propose to extract features of all videos and construct class prototypes. The prototype of each class is defined by taking the average of all features belonging to that class:

$$\mu_c = \frac{1}{|\mathcal{S}_c|}\sum_{v_i\in\mathcal{S}_c} f(v_i), \tag{5}$$

where $\mathcal{S}_c$ denotes the set of $|\mathcal{S}_c|$ samples from class c. Note that, both the source synthesized videos and the target videos are contained in $\mathcal{S}_c$ where the target videos are determined by their pseudo labels.

To refine the pseudo labels of the target videos, we normalize the prototypes and the features of the target videos into unit length. Then the similarities bwtween the target video features and the prototypes can be calculated via matrix multiplication:

$$Sim = F_t * C_p^\top, \tag{6}$$

where $F_t \in \mathcal{R}^{n_t\times d}$ and $C_p \in \mathcal{R}^{C\times d}$ are the normalized features of the target videos and the normalized class prototypes, respectively.

We take the class with the largest similarity as the refined label of the video. We resort to refining the pseudo label of the target video due the robustness of the prototypes. In intuition, even some wrong labeled samples are included to construct the prototype, the average of all samples is still reliable.

With the refined pseudo labels, we adopt standard cross entropy loss to learn a better spatio-temporal model:

$$\mathcal{L}_p = \frac{1}{B}\sum_{i=j}^{B} CE(\overline{\hat{p}(v_j^t)}, \hat{y}_j^s), \tag{7}$$

where $\hat{y}_j^s$ is the refined pseudo label of the target video v_j^t.

To further reduce the negative effect of noisy refined pseudo labels, we only select r% of the data for self training according to small loss criterion. Specifically, we rank the cross entropy losses computed with model predictions and the refined pseudo labels, in ascending order. We then select samples with the least r% losses as reliable samples for self training.

During the training process, the model makes more accurate predictions which in turn leads to more reliable prototypes. Besides, the domain discrepancy is also reduced during the training process so that the constructed prototypes are more suitable for the target videos. In a word, the refined pseudo labels are more accurate and provide better guidance for training a spatio-temporal model.

Overall, the proposed model is trained with the following loss:

$$\mathcal{L} = \mathcal{L}_c + \lambda_f\mathcal{L}_f + \lambda_b\mathcal{L}_{bnm} + \lambda_p\mathcal{L}_p, \tag{8}$$

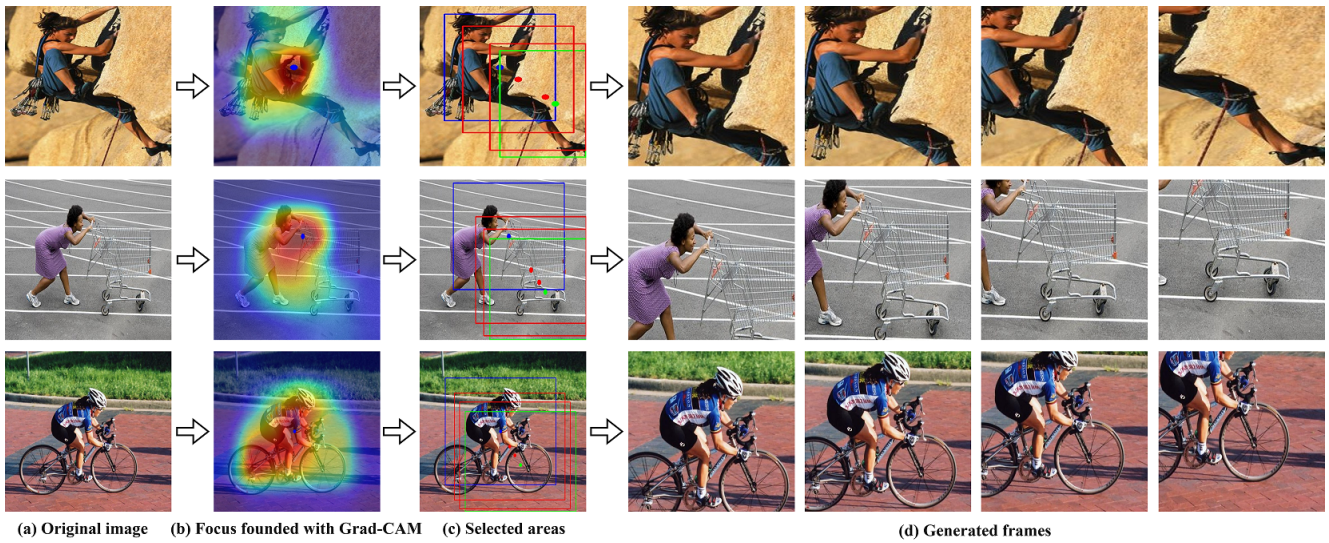


Figure 3: Visualization of synthesizing video from static images. (a) Samples (up to bottom) are from S→U, B→U and E→H, respectively. (b) The attention maps obtained via Grad-CAM, where the highest responses are marked with blue dots, for samples in (a). (c) The blue bounding boxes are the start areas and the green bounding boxes are the end areas. The rest red bounding boxes are intermediate areas along the start and end areas. (d) Generated frames corresponding to the selected areas in (c).

4 EXPERIMENTS

4.1 Datasets and Setup

Our method was evaluated through experiments on three standard image-to-video adaptation benchmarks: S→U, B→U, and E→H. For the S→U benchmark, we utilized the Stanford40 [39] dataset as the labeled source image domain and UCF101 [33] as the unlabeled target video domain. To conduct image-to-video adaptation task, we selected the 12 common classes between Stanford40 and UCF101 for training and evaluation. In the case of the B→U benchmark, we replaced the source image domain in S→U with the BU101 dataset [28]. We use a total of 101 classes for image-to-video adaptation task, as the classes from BU101 entirely correspond to those on UCF101. As for E→H, we utilized the EADs [3] dataset, which consists of Stanford40 [39] and the HII dataset [34], as the source image domain and HMDB51 [21] as the target video domain. There are 13 common classes between EADs and HMDB51 for image-to-video adaptation. The labeled source images and the unlabeled target videos can be used to train a spatio-temporal model that is invariant across the two domains.

4.2 Implementation Details

To train a model for locating objects via Grad-CAM [31], ResNet-50 [15] that was pretrained on ImageNet [7], is utilized as the backbone. We remove the last classifier layer from the original pretrained model and append a fully connected layer that includes C neurons to serve as the classifier layer. We train the ResNet with minibatch stochastic gradient descent (SGD) optimizer where the batch size, momentum and weight decay are set to 36, 0.9 and 0.0003, respectively. The learning rate of the k-th iteration is adjusted by $\eta_k = \eta_0 * (1 + \alpha * p)^{-\gamma}$ ($\alpha = 0.001$ and $\gamma = 0.75$), and η_0 is the initial learning rate. p linearly increases from 0 to 1 as k increases. η_0 and the total number of iterations for E→H and S→U tasks are set to 0.003 and 10000, respectively. For the B→U task , we set $\eta_0 = 0.005$ and train the model with 40000 iterations. We select 16 areas for each static source image and some examples are shown in Figure 3. We can see that the generated frames form reasonable camera movement and camera zooming, which may capture spatio-temporal information beneficial for video recognition.

For training a spatio-temporal model, we use I3D model Inception v1 [1] pretrained on the Kinetics dataset [20], which is also used in CycDA [23]. We train the RGB stream only for a fair comparison and the original fully connected layer is substituted with a new one having C neurons. We fix the first three Unit3D blocks to accelerate the training process. We train the I3D model with SGD optimizer where the momentum, and the weight decay are set to 0.9 and 0.0001, respectively. The initial learning rates and batch sizes are set to (0.025, 0.025, 0.1) and (16, 32, 32) for E→H, S→U and B→U tasks, respectively. We train the model with 20 epochs for E→H and S→U tasks, and with 30 epochs for the B→U task. We adopt multistep decaying learning rate with a 0.1 decay rate where the milestones of multistep learning rate decay are half the total epochs and the 2/3 of the total epochs. We trained the model with only cross entropy loss and BNM loss during the first (5, 5, 10) epochs for E→H, S→U and B→U tasks, as warmup training. After the warmup training phase, self training is implemented since the pseudo labels of model after warmup are much more accurate. We use all 16 frames of the synthesized source videos and randomly select 16 frames for each target video during training. We extract 32 frames uniformly for each target video in inference.

The hyper-parameters of λ_f are set to 1, 1 and 8 for E→H, S→U and B→U tasks, respectively. λ_b are set to 0.3, 0.3 and 0.1 for E→H, S→U and B→U tasks, respectively. λ_p are set to 0.2, 0.2 and 0.1 for

Table 1: Results on S→U (12 classes), B→U (101 classes) and E→H (13 classes), averaged over 3 random splits.

method	S→U	B→U	E→H
Source-only (Img)	76.8	54.8	37.2
DANN [11]	80.3	55.3	39.6
RTN [26]	83.8	-	40.2
JAN [27]	91.4	-	40.9
UnAtt [22]	-	66.4	-
HiGAN [41]	95.4	-	44.6
DAL [29]	97.6	-	45.5
MEDA [36]	94.3	-	43.1
SymGAN [40]	97.7	-	55.0
DANN+I3D	97.9	68.3	53.8
HPDA [2]	40.0	-	38.2
CycDA [23]	**99.1**	72.6	62.0
Ours	98.6	**80.5**	**71.3**
supervised target	99.3	93.1	83.2

Table 2: Ablation study results on S→U (12 classes), B→U (101 classes) and E→H (13 classes), averaged over 3 random splits.

method	S→U	B→U	E→H
Source only (Img)	76.8	54.8	37.2
Source only	96.3	62.2	59.0
Source only+BNM	98.3	79.9	68.9
Source only+BNM+Self-training	**98.6**	**80.5**	**71.3**

E→H, S→U and B→U tasks, respectively. We select 50% data for self training on all the three benchmarks.

4.3 Competitors and results

We compare our method to other image-to-video adaptation approaches as shown in Table 1. We compare against several approaches: DANN [11] is a classical image-level domain adaptation method with domain adversarial training. DAL [29] is also an image-level domain adaptation method that introduces a new domain adaptation layer to reduce the domain discrepancy by aligning source and target distributions to a reference one. RTN [26] jointly learns adaptive classifiers and transferable features where the difference between the source and target classifiers is formulated with a residual function. JAN [27] reduces the image-level domain shift by aligning the joint distributions of multiple domain-specific layers. MEDA [36] employs the principle of minimizing structural risk to train a domain-agnostic classifier on the Grassmann manifold. It also dynamically aligns the distributions of multiple domains while quantitatively evaluates the significance of marginal and conditional distributions. UnAtt [22] transfers the spatial attention map learned from the source domain to the target video frames. HiGAN [41] and SymGAN [40] learn to map the image features to the video features via GAN to mitigate the modality gap. DANN+I3D is a baseline that trains the I3D model with pseudo labels from an image-level adaptation model based on DANN. HPDA [2] is a partial domain adaptation approach and is re-implemented by the authors of CycDA [23] into a closed-set setting for a fair comparison. CycDA [23] is the state-of-the-art method that employs both class-agnostic and class-aware domain alignment for reducing domain discrepancy, and adopts pseudo labels to train a I3D model for bridging the modality gap. The lower bound (Source-only (Img)) and the upper bound (ground truth supervised target) from CycDA [23] are also reported here for reference.

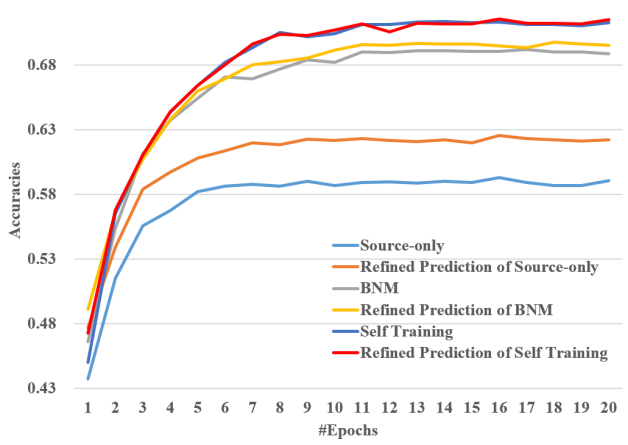
Table 1 presents the comparison results, where most of the results are borrowed from SymGAN [40] and CycDA [23]. Our method achieves new state-of-the-art performances on the B→U and E→H tasks and gets comparable results on the S→U task. Specifically, our method outperforms CycDA by 7.9% and 9.3% on the B→U and E→H tasks, respectively. Notably, the E→H task is much harder than S→U and B→U as the categories in the HMDB51 dataset are challenging to distinguish. The superiority of our method over CycDA shows that the proposed video synthesizing technique and the proposed self training are very effective in image-to-video adaptation task. It's worth noting that CycDA involves four stages and is trained with multiple cycle iterations, making it much more complicated than our method. Though our method performs slightly worse than CycDA on S→U task, our method is much simpler.

4.4 Ablation Study

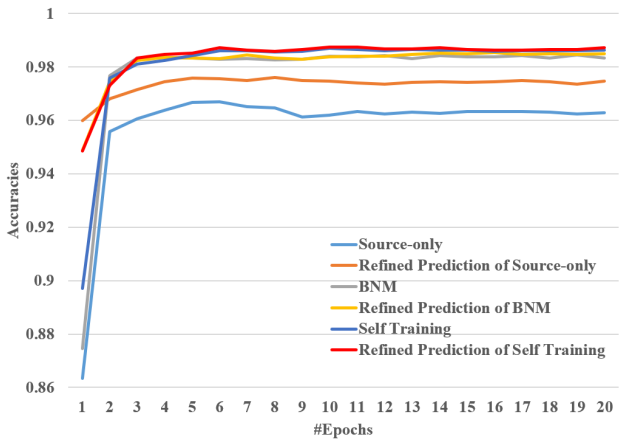
We conduct ablation study to better understand how the video synthesizing technique, the adopted BNM loss and the proposed self training affect the performance. In addition to the lower bound Source-only (Img), we further evaluate several variants of our full model. (1) **Source-only (Img)**, which denotes that we only adopt labeled source images to train a frame-level classifier. We aggregate the predictions of frames from target video as the prediction for that video. (2) **Source-only**, the model trained with synthesized videos from labeled source images by minimizing the cross entropy loss. (3) **Source-only+BNM**, the model is trained by minimizing the cross entropy loss and the BNM loss. (4) **Source-only+BNM+Self-training**, the full model that we add the self training with refined pseudo labels on the basis of Source-only+BNM.

The ablation study results are shown in Table 2. We can observe that the Source-only outperforms Source-only (Img) by large margins with 19.5%, 17.4% and 21.8% accuracies improvements on S→U , B→U and E→H, respectively. Such improvements validate that with the proposed video synthesizing technique, the modality gap can be well bridged which greatly improve the image-to-video adaptation task. As the image-to-video adaptation task is converted into video-to-video adaptation task, existing well-developed domain adaptation technique can be well explored to further improve image-to-video adaptation task. As shown in Table 2, the proposed baseline Source-only+BNM outperforms Source-only by large margins. This confirms that image-level domain adaptation techniques can be better incorporated into image-to-video adaptation task on the video-level domain with the help of the proposed video synthesizing technique. The full model Source-only+BNM+Self-training further improves the baseline Source-only+BNM, which validates the effectiveness of the proposed self training.

We also visualize the prediction accuracies the Source-only, Source-only+BNM and Source-only+BNM+Self-training and the accuracies

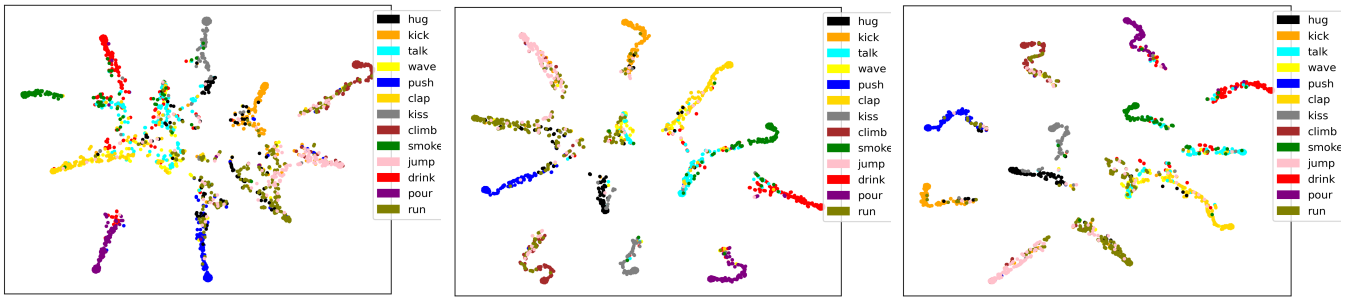


(a) E→H



(b) S→U

Figure 4: The accuracies of the Source-only, Source-only+BNM and Source-only+BNM+Self-training, and the accuracies of their refined predictions w.r.t. the number of trained epochs on E→H and S→U.



(a) Source-only (b) Source-only+BNM (c) Source-only+BNM+Self-training

Figure 5: t-SNE visualizations of target video features (colored w.r.t. ground truth). We plot the representations of Source-only (a) and the representations of constructed baseline Source-only+BNM (b) and the representations of our method (c).

of their refined predictions w.r.t. number of trained epochs on E→H and S→U. As shown in Figure 4, without self training, the accuracies of the refined predictions of Source-only, are much higher than those of Source-only on both E→H and S→U. The refined predictions over Source-only+BNM, perform slightly better than Source-only+BNM on both E→H and S→U. These observations validate that the refined pseudo labels based on the constructed prototypes are more accurate. With self training, the accuracies of Source-only+BNM+Self-training are close to those of refined predictions, indicating that the more accurate pseudo labels are well utilized in self training and further improve the performances.

Feature visualization. We further illustrate the t-SNE [35] feature visualizations of the ablation cases in Figure 5. All target videos of E→H are projected into two-dimensional features. Intuitively, the Source-only (Figure 5 (a)) leads to indiscriminative representations, which is improved by Source-only+BNM (Figure 5 (b)). For example, there is less confusion between "talk" and other categories in the representations obtained by Source-only+BNM. With the proposed self training based on refined pseudo labels, more discriminative representations can be obtained by Source-only+BNM+Self-training (Figure 5 (c)). For example, there is less confusion between "run"and "push" in the representations obtained by Source-only+BNM+Self-training compared with those obtained by Source-only+BNM. Besides, there is also less confusion between "talk" and "smoke" in the representations obtained by Source-only+BNM+Self-training compared with those obtained by Source-only+BNM. For these difficult classes, our method is trained with more accurate refined pseudo labels and is able to better distinguish them.

Hyper-parameter Sensitivity. Hyper-parameters λ_b and λ_p in Eqn. 8 play important roles in affecting the performance of the model. To evaluate the parameter sensitivity of Source-only+BNM+Self-training, we conduct experiments in E→H task. We vary $\lambda_p \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ and $\lambda_b \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ and the results are shown in Figure 6. We can observe that the proposed full method Source-only+BNM+Self-training can achieve competitive results under a wide range of hyper-parameter values. This confirms the effectiveness of the designed baseline based on the proposed

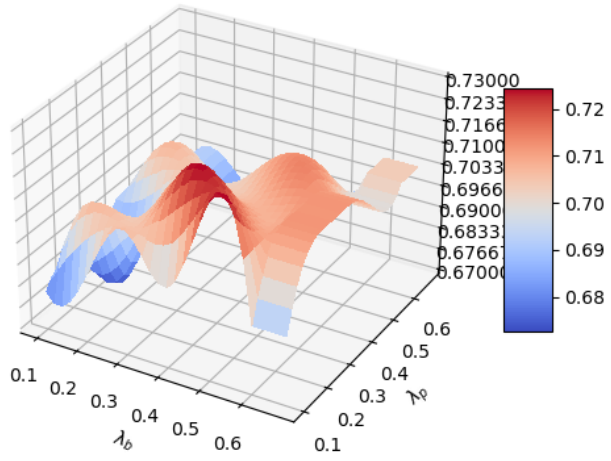


Figure 6: Hyper-parameter sensitivity analysis of Source only+BNM+Self-training with respect to λ_b and λ_p on E→H.

Table 3: Results on E→H (13 classes), averaged over 3 random splits.

method	Accuracy
Source-only	59.0
Source-only+DANN	59.3
Source-only+DANN+Self-training	64.3
Source-only+DAN	59.5
Source-only+DAN+Self-training	65.2
Source-only+CDAN	63.6
Source-only+CDAN+Self-training	66.5
Source-only+MCC	69.6
Source-only+MCC+Self-training	70.1
Source-only+BNM	68.9
Source-only+BNM+Self-training	71.3

Table 4: Results of various Source-only on E→H (13 classes), averaged over 3 random splits.

method	Accuracy
Source-only (4 frames)	40.7
Source-only (8 frames)	57.2
Source-only (16 frames)	59.0
Source-only (32 frames)	58.9

video synthesizing technique and the proposed self training based on the refined pseudo labels.

4.5 Further remarks

Other domain adaptation techniques. We also replace the BNM in the constructed baseline with several typical domain adaptation techniques, including DANN [11], DAN [24], CDAN [25], and MCC [18]. The trade-off parameters of transfer loss are kept as 0.3, the same as λ_b in Source-only+BNM. The trade-off parameters λ_p for pseudo cross entropy loss is set to 0.2, the same as the one adopted in Source-only+BNM+Self-training. The results are shown in Table 3 and we also report the results of Source-only+BNM and Source-only+BNM+Self-training for better demonstration. We can observe that, with the proposed video synthesizing technique, the image-to-video adaptation task is converted into video domain adaptation problem and all methods that adopt domain adaptation techniques outperform Source-only. This indicates that existing well developed domain adaptation techniques can be well explored to improve image-to-video adaptation task. Besides, with the proposed self training based on the refined pseudo labels, the performances of these methods are further improved. This indicates that the proposed self training is effective and can be well corporated with various baselines. Notably, we can also observe that the performances of Source-only+DANN are much better than those of DANN+I3D in Table 1, indicating that the proposed single-stage method are promising for dealing the image-to-video adaptation task.

Effect on number of frames. We also investigate how the number of frames affect the performance of the model. We construct various Source-only model with 4, 8, 16, and 32 frames and the results are shown in Table 4. We can see that with 4 frames, the performances of Source-only drop severely. We also find the accuracy (27.5%) of refined pseudo labels is much worse than that of Source-only (4 frames). This indicates that too few frames can not form appropriate temporal information to distinguish video categories and also lead to inaccurate prototypes. With 8 and 32 frames, the performance of Source-only is slightly worse than the one with 16 frames. Therefore we train our model with 16 frames to save computation resources while achieving comparable performances.

5 CONCLUSION

Our work focuses on addressing the image-to-video adaptation task, which aims to utilize labeled images and unlabeled videos for video recognition. This task poses two major challenges: first, the domain discrepancy between the two domains, which makes it difficult to generalize from images to videos, and second, the modality gap between the image and video modalities, which requires a mechanism to bridge the gap and effectively transfer knowledge across modalities. Most existing methods employ a two-stages paradigm, where frame-level adaptation is first utilized to reduce the domain discrepancy, followed by learning a spatio-temporal model to bridge the modality gap. In contrast, we provide a new perspective and propose a single-stage method that generates videos from a source static image, thereby converting the image-to-video adaptation problem into a video adaptation problem. By synthesizing video, we present a simple baseline approach, where an I3D model is trained with cross-entropy loss and BNM loss to encourage the classification responses of target videos to maintain discriminability and diversity. Furthermore, we propose a novel pseudo label generation method that inherits the robustness of the prototype and the effectiveness of the small loss criterion. With the proposed techniques, we present a method that achieves better or comparable results against the current state-of-the-art method. In future work, various video domain adaptation methods will be investigated in our reformulated image-to-video adaptation task.

REFERENCES

[1] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[2] Jin Chen, Xinxiao Wu, Yao Hu, and Jiebo Luo. 2021. Spatial-temporal causal inference for partial image-to-video adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1027–1035.

[3] Jin Chen, Xinxiao Wu, Yao Hu, and Jiebo Luo. 2021. Spatial-temporal causal inference for partial image-to-video adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1027–1035.

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

[5] Shizhe Chen and Dong Huang. 2021. Elaborative rehearsal for zero-shot action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13638–13647.

[6] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. 2020. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3941–3950.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[8] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. 2020. Omni-sourced webly-supervised learning for video recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 670–688.

[9] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. 2016. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 849–866.

[10] Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. 2016. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 923–932.

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

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

[13] Zan Gao, Leming Guo, Tongwei Ren, An-An Liu, Zhi-Yong Cheng, and Shengyong Chen. 2020. Pairwise two-stream convnets for cross-domain action recognition with small data. *IEEE Transactions on Neural Networks and Learning Systems* 33, 3 (2020), 1147–1161.

[14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 2672–2680. https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

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

[16] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. 2006. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems* 19 (2006).

[17] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. 2018. Deep Domain Adaptation in Action Space.. In *BMVC*, Vol. 2. 5.

[18] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. 2020. Minimum class confusion for versatile domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 464–480.

[19] Andrew Kae and Yale Song. 2020. Image to video domain adaptation using web supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 567–575.

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

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

[22] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2017. Attention transfer from web images for video recognition. In *Proceedings of the 25th ACM international conference on multimedia*. 1–9.

[23] Wei Lin, Anna Kukleva, Kunyang Sun, Horst Possegger, Hilde Kuehne, and Horst Bischof. 2022. CycDA: Unsupervised Cycle Domain Adaptation to Learn from Image to Video. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Springer, 698–715.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning Transferable Features with Deep Adaptation Networks. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 97–105. https://proceedings.mlr.press/v37/long15.html

[25] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems* 31 (2018).

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2016. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems* 29 (2016).

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

[28] Shugao Ma, Sarah Adel Bargal, Jianming Zhang, Leonid Sigal, and Stan Sclaroff. 2017. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition* 68 (2017), 334–345.

[29] Fabio Maria Carlucci, Lorenzo Porzi, Barbara Caputo, Elisa Ricci, and Samuel Rota Bulo. 2017. Autodial: Automatic domain alignment layers. In *Proceedings of the IEEE international conference on computer vision*. 5067–5075.

[30] Divya Saxena and Jiannong Cao. 2021. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–42.

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

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

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[34] Gokhan Tanisik, Cemil Zalluhoglu, and Nazli Ikizler-Cinbis. 2016. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters* 73 (2016), 44–51.

[35] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[36] Jindong Wang, Wenjie Feng, Yiqiang Chen, Han Yu, Meiyu Huang, and Philip S Yu. 2018. Visual domain adaptation with manifold embedded distribution alignment. In *Proceedings of the 26th ACM international conference on Multimedia*. 402–410.

[37] Han Wu, Chunfeng Song, Shaolong Yue, Zhenyu Wang, Jun Xiao, and Yanyang Liu. 2022. Dynamic video mix-up for cross-domain action recognition. *Neurocomputing* 471 (2022), 358–368.

[38] Junfei Xiao, Longlong Jing, Lin Zhang, Ju He, Qi She, Zongwei Zhou, Alan Yuille, and Yingwei Li. 2022. Learning from temporal gradient for semi-supervised action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3252–3262.

[39] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*. IEEE, 1331–1338.

[40] Feiwu Yu, Xinxiao Wu, Jialu Chen, and Lixin Duan. 2019. Exploiting images for video recognition: heterogeneous feature augmentation via symmetric adversarial learning. *IEEE Transactions on Image Processing* 28, 11 (2019), 5308–5321.

[41] Feiwu Yu, Xinxiao Wu, Yuchao Sun, and Lixin Duan. 2018. Exploiting images for video recognition with hierarchical generative adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 1107–1113.

[42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *ICLR 2017 : International Conference on Learning Representations 2017*.

[43] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. [n.d.]. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

[44] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. 2020. Few-shot action recognition with permutation-invariant attention. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 525–542.

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