Image-to-video Adaptation with Outlier Modeling and Robust Self-learning

Junbao Zhuo[1], Shuhui Wang[2*], Zhenghan Chen[3], Li Shen[4], Qingming Huang[5], Huimin Ma[1*]

[1]University of Science and Technology Beijing, Beijing, China
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[3]Microsoft (China) Co., Ltd., Beijing, China
[4]Sun Yat-Sen University, Guangzhou, China
[5]University of Chinese Academy of Sciences, Beijing, China
junbaozhuo@ustb.edu.cn, wangshuhui@ict.ac.cn, zhenghan.chen@alumni.pku.edu.cn
shenli6@mail.sysu.edu.cn, qmhuang@ucas.ac.cn, mhm@ustb.edu.cn

Abstract

The image-to-video adaptation task seeks to effectively harness both labeled images and unlabeled videos for achieving effective video recognition. The modality gap of the image and video modalities and the domain discrepancy across the two domains are the two essential challenges in this task. Existing methods reduce the domain discrepancy via close-set domain adaptation techniques, resulting in inaccurate domain alignment as there exist outlier target frames. To tackle this issue, we extend the vanilla classifier with outlier classes, where each outlier class responsible for capturing outlier frames for a specific class via batch nuclear norm maximization loss. We further propose a new loss by treating the source images apart from class c as instances from outlier class specific for c. As for the modality gap, existing methods usually utilize the pseudo labels obtained from an image-level adapted model to learn a video-level model. Rare efforts are dedicated to handling the noise in pseudo labels. We proposed a new metric based on label propagation consistency to select samples for training a better video-level model. Experiments on 3 benchmarks validating the effectiveness of our method.

Introduction

Video recognition plays a multifaceted role in various applications like security and surveillance, and personalized content delivery, across different domains. Training deep video models with good performance typically requires sufficient labeled videos (Duan et al. 2022), which is usually impractical as collecting and annotating videos are quite time-consuming and labor-intensive. To alleviate the dependence on large amounts of labeled video, semi-supervised video classification methods (Hu et al. 2023), few-shot video classification methods and zero-shot video classification methods (Gao, Chen, and Xu 2023; Gao, Zhang, and Xu 2020), have drawn much attention from the community and industry. However, these settings still obtain inferior performance or require substantial quantities of labeled videos of support set and seen categories. In the meanwhile, as it is easier to collect and label images, and also there are many well annotated image datasets, image-to-video adaptation (Chen et al. 2021) has been explored to leverage both the labeled images and unlabeled video for effective video recognition.

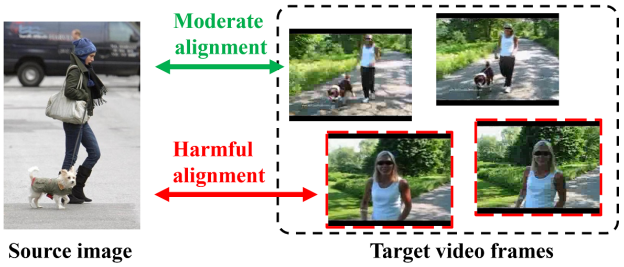

Figure 1: The outlier issue of image-to-video adaptation. There exist frames lack semantic information (with red dash boundaries) which is not appropriate for domain alignment as they may cause the negative transfer.

In this paper, we tackle image-to-video adaptation task which faces two essential issues, including the domain discrepancy across the source images and the target video frames, and the modality gap that the temporal information exists in videos is absent in source images. The domain discrepancy originates from the differences in background, light, image styles, camera perspectives and so on. The domain discrepancy results in performance degradation of high-performance source model over the target video frames and causes inaccurate guidance to further bridge the modality gap. As for the modality gap, the temporal information is essential for discriminating categories like "walk" and "run" where these two categories contain similar visual content. Without temporal information, the labels from the source domain can not be well transferred to the target video domain leading to an inaccurate video classifier.

To reduce the domain discrepancy, existing approaches (Gan et al. 2016; Yu et al. 2019; Kae and Song 2020) adopt statistical discrepancy metric like Maximum Mean Discrepancy, joint distributions (Long et al. 2017) or introduce domain adversarial learning. However, existing methods focus on global level domain-invariant feature learning based on the close set domain shift assumption. But there are many frames in the video lack of semantic information which can be seen as outlier samples.

*Corresponding authors

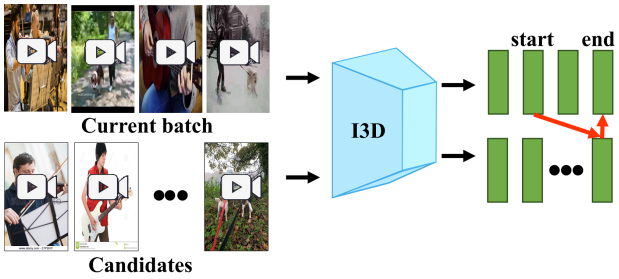

Figure 2: Label propagation consistency. The current batch of the unlabeled videos and all the candidates are fed into an I3D network to produce embeddings. An imaginary walker is sent from an embedding (start) from current batch to the candidates and back (end). The label consistency between the start and the end is used to judge the pseudo label quality.

For instance, as depicted in Fig. 1, "dog" in some target frames is absent for the class of "walk with dog". Directly reducing the domain discrepancy with the outlier samples leads to improper domain-invariant model. Such model may wrongly classify target video frames into other category and resulting in incorrect aggregated video prediction.

As for the modality gap, existing methods resort to generative model for temporal information completion (Yu et al. 2018, 2019) or self-learning with pseudo label (Kae and Song 2020; Lin et al. 2022). Using the generative model to learn the mapping network from the key frame features to the overall video features may suffer from the instability of generative model and the unstable quality of the selected key frames as there exists outlier frames. In the meanwhile, self-learning with pseudo labels can well model the temporal information with I3D or C3D, but the noise of pseudo label is not well addressed in existing methods.

To tackle with the domain discrepancy with outliers and modality gap, we propose a two-stage method including outlier aware image-level adaptation and noise-tolerant video-level self-learning. The first-stage model is explicitly designed to fit the domain adaptation with outlier for providing better pseudo labels for the second stage. The model in the second stage is optimized with selected samples according to a proposed metric.

Specifically, in the first stage, we construct a simple baseline in which we extend the original classifier with outlier classes. We utilize additional batch nuclear norm maximization loss (Cui et al. 2020) to encourage those outlier frames to be categorized into the outlier classes. Besides, to guide the outlier classes to better capture outlier frames, we design a pseudo outlier-class loss by treating the labeled source images apart from class y_i^s as instances from outlier class specific for y_i^s. Such design is based on the intuition that the objects exist in source images of class y_i^s could exist in all outlier frames for videos of all classes except class y_i^s (Saito, Kim, and Saenko 2021).

For the second stage, we propose a new metric to identify noise in the predicted pseudo labels obtained from the first stage, which is inspired by K-reciprocal nearest neighbours (Zhong et al. 2017) and CycleGAN. As shown in Fig. 2, we extract features of all videos as candidates and use feature of a video to find the most similar video from the rest videos. Then in turn, the retrieved most similar videos are used as queries to search for the most similar videos from the batch of videos. If the cycle ended at the same class from which it was started, then the label of the started sample is more likely to be correct. We use the label consistency of the retrieval cycle as a metric to select videos as correctly labeled data and keep the rest ones unlabeled. Then the semi-supervised methods are utilized to learn a video classifier.

We present theoretical analysis of our method and conduct extensive experiments on 3 image-to-video action recognition benchmarks and the experimental results verify that our two-stage method achieves promising results. The codes are available at https://github.com/junbaoZHUO/OMRS-I2V. To summarize:

- We reveal the outlier issue in the image-to-video adaptation task, and present a simple but effective method. Specifically, we explicitly model the outlier frames in videos and proposed a pseudo outlier-class loss to guide the outlier classes to capture "outlier" frames.
- We proposed a new metric based on the proposed label propagation consistency. The clean labeled videos can be selected and we use semi-supervised learning methods to train the video classification model.
- Based on the proposed techniques mentioned above, we construct a two-stage method that achieves promising performances. We provide ablation analysis to validate the contributions of the proposed techniques.

Related Work

Webly-supervised action recognition. Leveraging web-sourced images and videos as a label-free resource to enhance the video classification model, has garnered significant interest, primarily due to its potential to mitigate reliance on extensively annotated video datasets (Gan et al. 2016; Duan et al. 2020). Some techniques rely upon a small amount of labeled target video and further utilize collected web data for improving the trained model. For instance, in addition to the labeled video frames, Ma et al. (Ma et al. 2017) trained spatial CNNs using the collected web action image dataset as auxiliary training data and achieve comparable performance. However, these approaches do not explicitly handle the domain discrepancy between the web-sourced images and the target videos, potentially hindering effective knowledge transfer. Other strategies solely utilize collected web data for training video classification models (Gan et al. 2016). Gana *et al.* (Gan et al. 2016) proposed a mutual filtering technique to match distributions between the filtered web images and video frames through minimizing maximum mean discrepancy. While these methods are label-free and flexible, they generally underperform compared with supervised learning approaches.

Image-to-video domain adaptation. Current image-to-video domain adaptation (DA) techniques typically rely on the assumption that all the source images are accurately annotated and we can not access to the unlabeled target

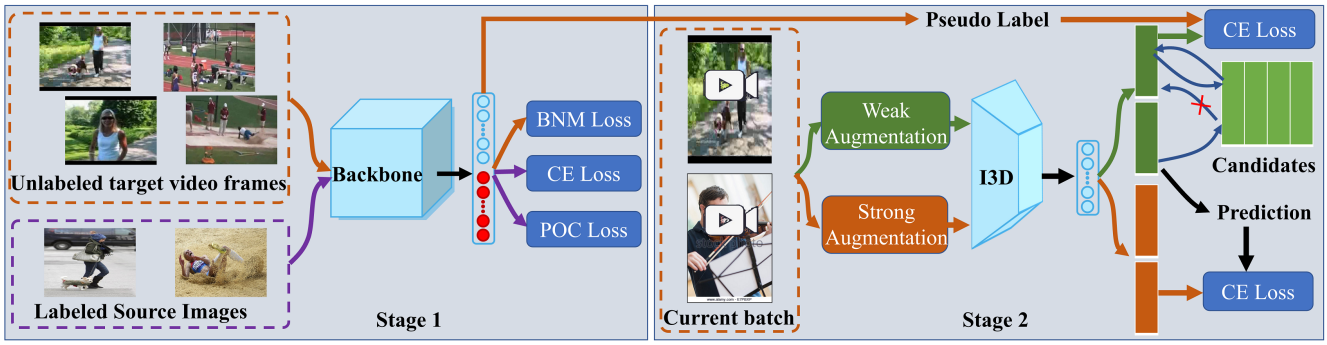

Figure 3: Framework of the proposed two-stage method. In the first stage, we add additional neurons for the outlier classes on the basis of an original classifier. The whole model is optimized via minimizing cross entropy (CE) loss and pseudo outlier loss with labeled source images, and batch nuclear norm maximization (BNM) loss over frames of unlabeled target video. In the next stage, we use the proposed label propagation consistency metric to select accurate pseudo labels and the rest samples are treated as unlabeled ones and adopt a semi-supervised learning paradigm to train the spatio-temporal model.

videos. The primary challenge in these approaches is to minimize the modality gap and the domain discrepancy (Li et al. 2017; Yu et al. 2019; Liu et al. 2019; Kae and Song 2020). Some approaches (Liu et al. 2019) assume that part of labeled target videos can be used for training which are usually called semi-supervised methods. For instance, Liu et al. (Liu et al. 2019) proposed a method to learn neural networks that map image, key frame and video modalities into a domain-invariant representations space via maintaining the cross-modal similarities and fuse features of different modalities. Other methods handle the more difficult unsupervised domain adaptation setting that the target videos are unlabeled. In (Li et al. 2017), 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 category is inferred as the one with the largest responses among all classes. Furthermore, hierarchical GAN (Yu et al. 2018), symmetric GAN (Yu et al. 2019) and spatio-temporal causal graph (Chen et al. 2021) are proposed to learn the mapping network from the key frame features to the unlabeled video features with GAN.

The two related works close to ours are the ones of Kae et al. (Kae and Song 2020) and Lin et al. (Lin et al. 2022) which train a frame-level domain-invariant model and a video-level model to narrow the modality gap. Kae *et al.* (Kae and Song 2020) applied domain adversarial training to train a frame-level network and transferred the learned weights to a spatio-temporal network to narrow the modality gap. Lin *et al.* (Lin et al. 2022) presented a four-stage method, starting with class-agnostic frame-level domain alignment to produce pseudo labels, followed by training an independent video-level model. The latter stages alternately conduct spatial alignment and spatio-temporal learning, including class-aware domain alignment. In contrast to these methods, we address the domain adaptation with outlier in the first stage, and handle the noise in the pseudo label in the second stage.

Methodology

Suppose that there are a source image domain $I_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ of n_s labeled samples and a target video domain $V_t = \{v_j^t\}_{j=1}^{n_t}$ of n_t unlabeled videos. Each video v_j contains M_j frames and the set of frames is denoted by $V_t^F = \{\{v_{j,k}^t|_{k=1}^{K_j}\}\}_{j=1}^{n_t}$. Assume that $(x_i^s, y_i^s) \sim P_S(x, y)$ and $v_{j,k}^t \sim P_T(x)$. Both two domains share the same C categories. The image-to-video adaptation task seeks to effectively harness both labeled images and unlabeled videos for learning a model that performs well in the target domain.

As shown in Fig. 3, we present a two-stage method where the first stage addresses the domain adaptation with outlier problem in image-to-video adaptation task, and the second stage tackles the noisy pseudo labels obtained in stage 1 to train spatio-temporal model for mitigating the modality gap.

Stage 1

To handle the outlier issue in image-to-video adaptation, we explicitly model the outlier classes to capture the outlier frames that lack semantic information. Specifically, as shown in the left part of Fig. 3, we add C additional neurons into the original C neurons in the classifier layer. The first C neurons for the designed classifier stand for the C categories. The remaining C neurons represent class-specific "outliers". The intuition is that we expect the $(k+C)$-th neuron to capture the existence of the outlier frames for videos from the k-th category.

To train a discriminative model, we employ the standard cross-entropy (CE) loss over the labeled source data I_s. The classification loss $\mathcal{L}_c$ is expressed as:

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

where the $CE(\cdot,\cdot)$ denotes the cross-entropy loss, and $\hat{p}(x_i^s) \in \mathcal{R}^{2C}$ is the network prediction over an image x_i^s.

To encourage the model to push the responses of outlier frames into the outlier classes, we resort to maximizing the rank of the category prediction matrix over a batch

of target video frames via Batch Nuclear-Norm Maximization (abbreviated as BNM). Ideally, full rank prediction matrix means that all categories including the outlier classes are activated. Adopting BNM enforces the model automatically assigns some frames to the outlier classes and frames similar to source images are more likely to be assigned to the original categories. Besides, the discrimination of the model can also be improved by adopting BNM. BNM is performed on the category predictions matrix over a batch of unlabeled video frames, with no supervision as follows:

$$\mathcal{L}_{bnm} = -\frac{1}{B}||p(V)||_\star \tag{2}$$

where the $p(V)$ is the category prediction matrix towards a batch of unlabeled target frames V with B samples. $||\cdot||_\star$ is the matrix nuclear-norm.

To guide the outlier-class neurons to better capture the "outlier" frames, we leverage the annotations of source images and propose a pseudo outlier-class loss as follows:

$$\mathcal{L}_{poc} = \frac{1}{B}\sum_{i=1}^{B}\phi(\hat{p}(x_i^s[:, C:])) * \log(one\_hot(y_i^s)), \tag{3}$$

where $x_i^s[:, C:]$ are the responses of outlier-class neurons for the sample x_i^s and ϕ is the softmax function. one_hot function first convert the layer y_i^s into one-hot embedding and then the values of the embedding are clipped into $[1e-4, 1-1e-4]$. Minimizing the pseudo outlier-class loss encourages the outlier-class neurons except the y_i^s-th one to be activated. The reason is that the objects exist in labeled source images could exist in outlier frames for videos of all classes except class y_i^s.

The total loss for training the image-level domain adaptation model in stage 1 is:

$$\mathcal{L} = \mathcal{L}_c + \lambda_b\mathcal{L}_{bnm} + \lambda_{poc}\mathcal{L}_{poc}, \tag{4}$$

where λ_b and λ_{poc} denotes the trade-off parameters.

Once trained, we extract the classifier responses for all frames of a target video and the in-class responses are averaged as the final prediction. The class with the largest response in the final prediction is adopted as the pseudo label for a target video to train a spatio-temporal model in stage 2.

Theoretical Analysis: We present a generalization bound of unsupervised domain adaptation:

$$\begin{aligned}
\mathcal{L}_{\text{test}} =& \mathbb{E}_{p_T(x,y)}[-\log\hat{p}(y|x)] && (5)\\
=& \mathbb{E}_{p_T(x,y)}[-\log\mathbb{E}_{p(z|x)}[\hat{p}(y|z)]] && (6)\\
\leq& \mathbb{E}_{p_T(x,y)}[\mathbb{E}_{p(z|x)}[-\log\hat{p}(y|z)]] && (7)\\
=& \mathbb{E}_{p_T(z,y)}[-\log\hat{p}(y|z)] && (8)\\
=& \int -\log\hat{p}(y|z)p_T(z,y)dzdy && (9)\\
=& \int -\log\hat{p}(y|z)p_S(z,y)dzdy + && (10)\\
& \int -\log\hat{p}(y|z)[p_T(z,y) - p_S(z,y)]dzdy && (11)\\
=& \mathcal{L}_{\text{train}} + \int -\log\hat{p}(y|z)[p_T(z,y) - p_S(z,y)]dzdy && (12)
\end{aligned}$$

Assume that $-\log\hat{p}(y|z)$ is bounded by N, for $\forall z \in \mathcal{Z}, y \in \mathcal{Y}$, we have:

$$\begin{aligned}
\mathcal{L}_{\text{test}} \leq& \mathcal{L}_{\text{train}} + \frac{N}{2}\sqrt{2\int p_T(z,y)\log\frac{p_T(z,y)}{p_S(z,y)}dzdy} && (13)\\
=& \mathcal{L}_{\text{train}} + \frac{N}{\sqrt{2}}\sqrt{\text{KL}[p_T(z,y)||p_S(z,y)]} && (14)
\end{aligned}$$

$$\begin{aligned}
&\text{KL}[p_T(z,y)||p_S(z,y)] && (15)\\
=&\text{KL}[p_T(z)|p_S(z)] + \mathbb{E}_{p_T(z)}[\text{KL}[p_T(y|z)|p_S(y|z)]] && (16)
\end{aligned}$$

According to the popular covarite shift assumption, the conditional misalignment $\mathbb{E}_{p_T(z)}[\text{KL}[p_T(y|z)|p_S(y|z)]]$ is often small (and fixed, not dependent on z).

So the bound can be reduced via minimizing $\text{KL}[p_T(z)|p_S(z)]$, with the objective:

$$\begin{aligned}
&\mathcal{L}_{\text{train}} + \alpha\text{KL}[p_T(z)|p_S(z)] && (17)\\
=&\mathcal{L}_{\text{train}} + \mathbb{E}_{p_T(z)}[\log p_T(z) - \log p_S(z)] && (18)\\
=&\mathcal{L}_{\text{train}} - H(p_T(z)) - \mathbb{E}_{p_T(z)}[\log p_S(z)] && (19)
\end{aligned}$$

where α is a hyper-parameter.

This is why existing methods (Vu et al. 2019) works via minimizing entropy $-H(p_T(z))$ over target samples.

In practical implementation, the loss computation and the learning process are implemented in mini-batches. There is no exact parametric model for $p_S(z)$. To make the computations feasible, within each minibatch $\{(x_i^t, y_i^t)\}_{i=1}^B$, the source and the target distributions are approximated using a mixture of multivariate Gaussian distributions, similar to the approach described in (Nguyen et al. 2021), but with a fixed variance. Specifically, $p(z|x) = \mathcal{N}(z; \mu(x), \sigma^2 I)$, for each input x. $\mu(x)$ is the network representation for a input x. For a mini-batch $\{(x_i^t, y_i^t)\}_{i=1}^B$, the source and the target distributions can be approximated as: $p_S(z) \approx \frac{1}{B}\sum_{i=0}^{B} p(z|x_i^s) = \mathcal{N}(z; \mu(\bar{x_s}); \sigma^2 I)$ and $p_T(z) \approx \frac{1}{B}\sum_{i=0}^{B} p(z|x_i^t)$, respectively. $p_S(z)$ is multivariate Gaussian distributions with $\mu(\bar{x_s})$, which is the mean embedding of source domain examples. Minimizing $-\mathbb{E}_{p_T(z)}[\log p_S(z)]$, the last term in Eqn. (19), means maximum likelihood estimation of $p_T(z)$ towards $p_S(z)$. That is the reason why the information maximization (IM) loss ((Shi and Sha 2012)) works in unsupervised domain adaptation. In IM loss, the mean embedding of target domain is encouraged to match the mean embedding of source domain, which is assumed to be C-dimensional vector with all $1/C$.

The F-norm $||P(V)||_F$ and entropy of matrix $H(p(V))$ have strict opposite monotonicity. So minimizing $-H(p_T(z))$ can be achieved by maximizing $||P(V)||_F$. And we can further show that $\frac{1}{\sqrt{D}}||P(V)||_* \leq ||P(V)||_F$. So $-||P(V)||_F \leq -\frac{1}{\sqrt{D}}||P(V)||_*$. Therefore, minimizing $\mathcal{L}_c + \lambda_b\mathcal{L}_{bnm}$ leads to minimizing the $\mathcal{L}_{\text{train}} - H(p_T(z))$ in Eqn. (19). Proof can be found in Supplementary material.

As for minimizing $-\mathbb{E}_{p_T(z)}[\log p_S(z)]$, as illustrated above, we need to match the mean embedding of the target domain to the one of source domain. However, as mentioned in (Nguyen et al. 2021), the approximations $p(z) \approx$

$\frac{1}{B}\sum_{i=0}^{B} p(z|x)$ is biased. Besides, the category uniformly distributed assumption is improper the target domain as their exists outlier frames. So we release the above assumption to the one that the element of mean embedding is larger than 0. That is, there is no category without responses. Such constraint means the rank of $p(V)$ is full and can be well satisfied with maximization of $||p(V)||_*$ as that the convex envelope of rank(p(V)) is the $||p(V)||_*$ when $||p(V)||_F \leq 1$.

Stage 2

To tackle the modality gap between the source image domain and the target video domain, we use the pseudo labels of the video obtained in the first stage to supervise the training of a spatio-temporal model using merely video modality. However, the pseudo label contains much noise resulting in severe performance degradation of the trained model.

We propose a new metric based on label propagation consistency, which is inspired by K-reciprocal nearest neighbours (Zhong et al. 2017) and CycleGAN. To facilitate label propagation consistency, as shown in Fig. 3, the current sample batch and all the candidate videos are mapped into vectors. As the candidates set usually contains large amounts of samples, feeding all samples into the network is impractical. Therefore we maintain a memory that stores the representations of all samples. As the representations changes during the training process, we adopt the temporal ensembling technique. In detail, let $\vec{t}_i(k)$ and $\vec{\mu}_i(k)$ be the stored representation and current representation generated by the model, respectively, for a sample i in iteration k during training. The stored representations in iteration $k+1$ are computed as:

$$\vec{t}_i(k+1) = \beta\vec{t}_i(k) + (1-\beta)\vec{\mu}_i(k), \tag{20}$$

where momentum satisfies $0 \leq \beta < 1$ (we directly set $\beta = 0.9$ by following (Liu et al. 2020)). The stored memory can be represented by $M_m \in \mathcal{R}^{n_t \times C}$ and m is the epoch index.

We feed a current batch of samples with pseudo labels L_m into the I3D model and obtain their representations $F_m \in \mathcal{R}^{B\times C}$, where C is the length of the representation vector, and B is the batch size. We treat them as queries to search for the most similar items in the candidate memory with cosine similarity. By normalizing the memory and the representations of the current batch into unit length, we can calculate the similarities via inner product:

$$Sim = F_m * M_m'^{\top}, \tag{21}$$

where M'_m is nearly identical to M_m whose representations corresponding to the current batch is set to zero vectors. These zero vectors prevent the "imaginary walker" visit the sample itself and return to itself. We sort Sim and obtain the most similar items for the current batch of samples, which form a matrix $Z_m \in \mathcal{R}^{B\times C}$. Then we use Z_m as queries to search for the most similar items in F_m via inner product $Z_m * F_m^{\top}$. Let L'_m denote the pseudo labels of the obtained most similar items and the label propagation consistency can be represented as:

$$R_m = (L_m == L'_m). \tag{22}$$

Such consistency can to some extent indicate whether the pseudo label is correct or not, as it is hard for a wrongly labeled sample to walk to a sample with the same wrong label via two successive nearest neighbor searches. We preserve the label propagation consistency for all sample in each training epoch. The temporal average aggregation of the label propagation consistency (R_m is converted into float number) for sample i is used as a metric to select pseudo labeled samples. Specifically, the metric is computed as $\hat{R}_i = \frac{1}{E}\sum_{m=1}^{E} R_m[i]$, in which E denotes the total number of training epochs.

By ranking the proposed metric by class in an ascending order, we use the $\hat{R}_i$ of the $(m_c \times p)$-th example as a threshold of class c for choosing samples. p is the selection ratio and m_c is the number of samples in class c. The classwise selection prevents heavily imbalanced selected data.

Then we can obtain $p \times n_t$ of pseudo labeled videos and treat the rest $(1-p) \times n_t$ of videos as unlabeled samples and perform semi-supervised learning. Specifically, we leverage FixMatch (Sohn et al. 2020), which integrates pseudo-labeling and consistency regularization. FixMatch employs distinct strong and weak augmentations during the consistency regularization process. For every unlabeled sample $u \in \mathcal{D}^u$ in the target domain, strong augmentation $\mathcal{A}^s$ and the weak augmentation $\mathcal{A}^w$ are applied as $u_s = \mathcal{A}^s(u); u_w = \mathcal{A}^w(u)$. The strong data augmentation $\mathcal{A}^s$ utilizes the technique introduced in (Cubuk et al. 2020). Meanwhile, the weak data augmentation $\mathcal{A}^w$ comprises image flipping and image translation across video frames. The consistency regularization, incorporated with pseudo-labeling, is performed by treating the prediction of weakly augmented images u_w as pseudo-labels and encouraging the predictions of strongly augmented images to align with these pseudo-labels. However, the pseudo-labels may include inaccuracies, leading to error propagation. To address this issue and minimize the impact of wrong pseudo-labels, only those samples with highly confident predictions are chosen for calculating consistency regularization. The consistency regularization loss on unlabeled video frames is defined as:

$$\mathcal{L}_u = \mathbb{E}_{u\sim\mathcal{D}^u}\ \mathbb{1}(max(p(u_w)) > \tau) CE(p(u_s), \hat{y}(u_w)) \tag{23}$$

where $\hat{y}(u_w) = argmax(p(u_w))$ and τ is the threshold. By minimizing the consistency loss $\mathcal{L}_u$, the obtained decision boundary is driven away from the labeled samples. This ensures that the decision boundary compels the model to be robust against perturbations in the video data and effective in classifying unlabeled videos.

To train a discriminative model, we employ the cross-entropy (CE) loss for the selected $p \times n_t$ of pseudo labeled videos. The loss $\mathcal{L}_s$ with pseudo labels $\hat{y}$ is defined as:

$$\mathcal{L}_s = \frac{1}{B}\sum_{i=1}^{B} CE(\hat{p}(v^t), \hat{y}). \tag{24}$$

The loss minimized by FixMatch (Sohn et al. 2020) is

$$\mathcal{L}' = \mathcal{L}_s + \lambda_u \mathcal{L}_u, \tag{25}$$

where λ_u is a hyper-parameter to weight the unlabeled loss.

Both outlier-class modeling and pseudo label selection play important roles. The high quality pseudo labels obtained from stage 1 help training a effective video-level

model. With the proposed label propagation consistency metric, more accurate pseudo labels can be chosen to train a robust model according to the theoretical guarantee obtained in (Wei et al. 2020).

Experiments

Datasets and Setup

We conduct experiments on 3 image-to-video benchmarks such as Stanford40→UCF101 (S→U), BU101→UCF101 (B→U) and EADs→HMDB51 (E→H) for evaluation. Specifically, in S→U, the source image domain is Stanford40 (Yao et al. 2011) and the target video domain is UCF101 (Soomro, Zamir, and Shah 2012). The 12 common categories across Stanford40 and UCF101 are adopted for evaluation. For B→U, the source image domain is replaced by BU101 dataset (Ma et al. 2017) and the total 101 classes of both BU101 and UCF101 are selected for evaluation, since the categories from BU101 are completely the same as those on UCF101. For the E→H benchmark, the source image domain is EADs (Chen et al. 2021), which consists of Stanford40 and the HII dataset (Tanisik, Zalluhoglu, and Ikizler-Cinbis 2016), and the target video domain is HMDB51 (Kuehne et al. 2011). There are 13 common categories across HMDB51 and EADs for evaluation.

Implementation Details

For the frame-level domain adaptation stage, we employ ResNet-50 pretrained on ImageNet as the backbone, which could obtain source-only performance similar to CycDA.The final classifier layer of the pretrained ResNet-50 is removed and we add a bottleneck fully connected (FC) layer with 256 neurons and a following FC layer with $2 \times C$ neurons as the classifier layer. For Stage 2, we employ the I3D model based on Inception v1, pre-trained on the Kinetics dataset (Yu et al. 2019; Lin et al. 2022). Specifically, we train only the RGB stream and modify the original final FC layer by replacing it with one that consists of C neurons.

We train the ResNet using the Stochastic Gradient Descent (SGD) optimizer. The weight decay, batch size, and momentum are set to $3e-4$, 36 and 0.9. We implement a learning rate annealing strategy for iteration k-th as $\xi_k = \xi_0 * (1 + 0.001 * p)^{-0.75}$, and ξ_0 denotes the initial learning rate. p is a parameter that linearly increases from 0 to 1 as k increases. $\xi_0 = 3e-3$ for S→U and E→H and $\xi_0 = 5e-3$ for B→U. We train the model in stage 1 with 10000 iterations for S→U and E→H tasks, and with 40000 iterations for the B→U task. λ_b and the λ_{poc} are set to 0.5 and 0.1.

We train the modified I3D network using SGD optimizer, setting the weight decay, momentum, and batch size to 0.0001, 0.9 and 16. We train the model with 30 epochs for B→U task, and with 20 epochs for others. The initial learning rate for S→U, E→H and B→U are 0.05, 0.05 and 0.1. We utilize a multistep decaying learning rate scheme with a decay factor of 0.1. The milestones for the learning rate decay are set at the midpoint and at 75% epochs. $\lambda_u = 1$. We randomly choose 32 frames over target video for training and uniformly extract 32 frames for inference. τ in Eqn. (23) is set to 0.9, following (Sohn et al. 2020).

method	S→U	B→U	E→H
source only†	89.1	68.2	41.7
UnAtt (Li et al. 2017)	-	66.4	-
SymGAN (Yu et al. 2019)	97.7	-	55.0
DANN+I3D	97.9	68.3	53.8
HPDA (Chen et al. 2021)	40.0	-	38.2
CycDA (Lin et al. 2022)	**99.1**	72.6	62.0
Ours	99.0	**78.1**	**66.3**

Table 1: Results on S→U, B→U and E→H, averaged over 3 random experiments. † denotes our evaluation.

method	S→U	B→U	E→H
SO	89.1	68.2	41.7
SO+BNM	96.5	77.0	54.6
SO+BNM+POC	**96.9**	**77.1**	**55.3**

Table 2: Ablation study results for stage 1 on S→U, B→U and E→H, averaged over 3 random experiments.

Results

We compare our method with other image-to-video adaptation approaches as illustrated in Table 1. We compare against several methods: UnAtt (Li et al. 2017) transfers the spatial attention map learned from the source image domain to the unlabeled target video frames. SymGAN (Yu et al. 2019) learns the mapping network between the image features and the unalbeled video features using GAN. CycDA (Lin et al. 2022), a 4-stage method that utilizes class-agnostic domain alignment and class-aware frame-level domain alignment, and adopt pseudo labels to train a video-level model.

As shown in Table 1, we can clearly observe that our method surpasses all the compared methods on E→H and B→U, and get comparable results on S→U. In detail, our method surpass CycDA (Lin et al. 2022) with 4.3% and 5.5% improvements in E→H and B→U tasks. Noting that E→H is much more difficult than B→U and S→U since the actions in HMDB51 dataset are hard to distinguish. There are many frames that lack semantic information and aligning such frames does not help to distinguish the difficult class. As for B→U, there are much more diverse categories and the "outlier" issue is severe. Nevertheless, by explicitly modeling the "outlier", our method could improve the quality of the pseudo label and further improve the performance. Note that CycDA (Lin et al. 2022) is much more complicated as it involves four stages and is optimized with several cycles, which is much complicated than ours. Our method is much simpler and effective, validating its superiority over CycDA.

Ablation Study

Stage 1: We perform an ablation study to gain deeper insights into the effects of batch nuclear norm maximization (BNM) loss and the proposed pseudo outlier-class loss work in stage 1. We evaluate 3 variants: (1) **SO** (Source only), means that only the labeled source images are uti-

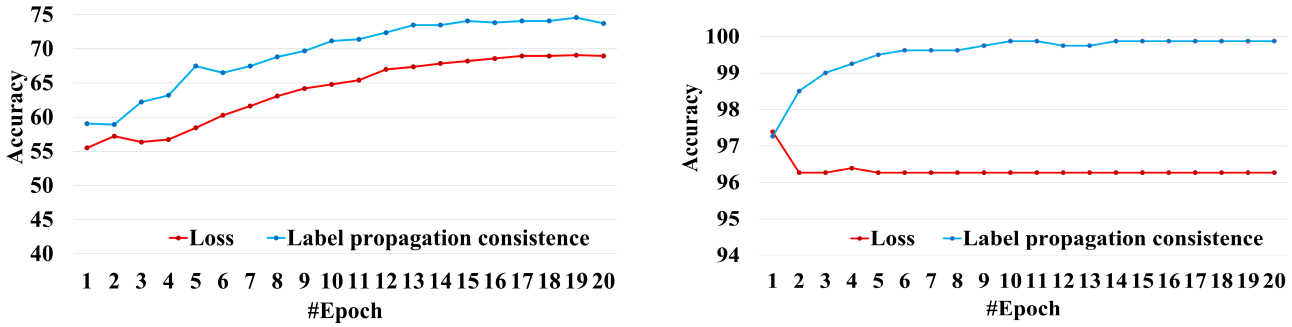

Figure 4: The accuracies of the selected pseudo label according to CE loss (red) or Label propagation consistency (blue) w.r.t. the trained epochs on E→H (left) and S→U (right), respectively.

method	S→U	B→U	E→H
Baseline	98.5	77.5	64.0
Ours	**99.0**	**78.1**	**66.3**

Table 3: Ablation study results for stage 2 on S→U, B→U and E→H, averaged over 3 random experiments.

lized to train the model. (2) **SO+BNM**, the model with outlier classes and trained with both CE loss and BNM loss. (3) **SO+BNM+POC**, the full model that we utilize the proposed pseudo outlier-class loss on the basis of SO+BNM. As shown in Table 2, we can see that the performance of **SO+BNM** surpasses **SO** by a large margin. This confirms that by explicitly modeling the outlier classes, the domain discrepancy can be well reduced resulting in a model with good generalization ability to the target video frames. With the proposed pseudo outlier-class loss, **SO+BNM+POC** further improves SO+BNM, which validates its effectiveness.

Stage 2: We perform ablation study to see how the proposed metric affect the spatio-temporal model in stage 2. We evaluate 2 variants: (1) **Baseline**, that trained with the all pseudo label from stage 1. (2) **Ours**, that trained with Fixmatch that the labeled samples are selected with the proposed metric. We can clearly see that our method surpasses Baseline, especially in the E→H task. As the categories in E→H are much harder to distinguish, the temporal information is more valuable for classification. With the proposed metric, our method could select more accurate labeled videos and guide the spatio-temporal model to distinguish difficult categories.

We also visualize the accuracies of select pseudo labels according to the proposed metric in E→H and S→U tasks. The accuracies of select pseudo labels according to small loss criterion are also visualized for a comparison. As depicted in Fig. 4, our method could select more accurate pseudo labels against those selected via CE loss.

Hyper-parameter Sensitivity Analysis. For λ_b and λ_u presented in Eqn. (4) and Eqn. (25), we follow the standard practice (Sohn et al. 2020; Cui et al. 2020). Hyper-parameter λ_{poc} plays a essential role in affecting the performance of the image-level domain adaptation model. For evaluating the sensitivity of our model SO+BNM+POC over different

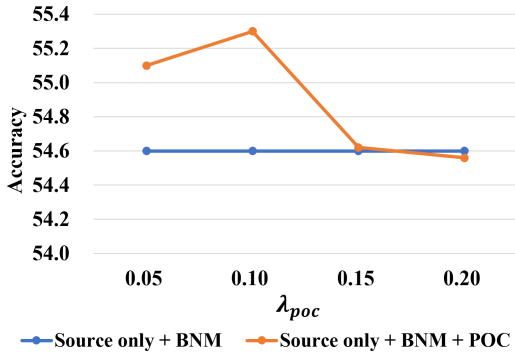

Figure 5: Hyper-parameter sensitivity analysis of our full model SO+BNM+POC on E→H.

parameters, we implement several experiments with E→H benchmark by varying $\lambda_{poc} \in \{0.05, 0.10, 0.15, 0.20\}$. Fig. 5 shows that the accuracy of SO+BNM+POC first increases and then decreases as λ_{poc} varies. This confirms the effectiveness of our pseudo outlier-class loss in guiding the outlier-class neurons to capture the "outlier" and improve the transferability of image-level adaptation model.

Conclusion

We present a two-stage method for image-to-video adaptation to handle the domain shift issue and the modality gap challenge. We focus on the outlier issue ignored in existing methods by explicitly modeling the outlier classes for tolerating the outlier frames that lack semantic information. We extend an original classifier with outlier classes and adopt additional BNM loss for training outlier classes. We further proposed a pseudo outlier-class loss with labeled source images to imitate pseudo outlier classes. As for the modality gap, pseudo labels obtained in the first stage are utilized to train a video-level model. We proposed a new metric based on label propagation consistency to identify noise in pseudo labels to train a better video classification model. Experimental results show that our method achieves promising performances on 3 image-to-video benchmarks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China: 62402033, U20B2062, 62236008, U21B2038 and 61931008.

References

Chen, J.; Wu, X.; Hu, Y.; and Luo, J. 2021. Spatial-temporal causal inference for partial image-to-video adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1027–1035.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 702–703.

Cui, S.; Wang, S.; Zhuo, J.; Li, L.; Huang, Q.; and Tian, Q. 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.

Duan, H.; Zhao, Y.; Chen, K.; Lin, D.; and Dai, B. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2969–2978.

Duan, H.; Zhao, Y.; Xiong, Y.; Liu, W.; and Lin, D. 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*, 670–688. Springer.

Gan, C.; Sun, C.; Duan, L.; and Gong, B. 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*, 849–866. Springer.

Gao, J.; Chen, M.; and Xu, C. 2023. Vectorized Evidential Learning for Weakly-supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15949 – 15963.

Gao, J.; Zhang, T.; and Xu, C. 2020. Learning to model relationships for zero-shot video classification. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3476–3491.

Hu, Y.; Gao, J.; Dong, J.; Fan, B.; and Liu, H. 2023. Exploring rich semantics for open-set action recognition. *IEEE Transactions on Multimedia*.

Kae, A.; and Song, Y. 2020. Image to video domain adaptation using web supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 567–575.

Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; and Serre, T. 2011. HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision*, 2556–2563. IEEE.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2017. Attention transfer from web images for video recognition. In *Proceedings of the 25th ACM international conference on multimedia*, 1–9.

Lin, W.; Kukleva, A.; Sun, K.; Possegger, H.; Kuehne, H.; and Bischof, H. 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*, 698–715. Springer.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342.

Liu, Y.; Lu, Z.; Li, J.; Yang, T.; and Yao, C. 2019. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29: 3168–3182.

Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2018. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Long, M.; Zhu, H.; Wang, J.; and Jordan, M. I. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, 2208–2217. PMLR.

Ma, S.; Bargal, S. A.; Zhang, J.; Sigal, L.; and Sclaroff, S. 2017. Do less and achieve more: Training cnns for action recognition utilizing action images from the web. *Pattern Recognition*, 68: 334–345.

Nguyen, A. T.; Tran, T.; Gal, Y.; Torr, P.; and Baydin, A. G. 2021. KL Guided Domain Adaptation. In *International Conference on Learning Representations*.

Saito, K.; Kim, D.; and Saenko, K. 2021. Openmatch: Open-set semi-supervised learning with open-set consistency regularization. *Advances in Neural Information Processing Systems*, 34: 25956–25967.

Shi, Y.; and Sha, F. 2012. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 1275–1282.

Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Tanisik, G.; Zalluhoglu, C.; and Ikizler-Cinbis, N. 2016. Facial descriptors for human interaction recognition in still images. *Pattern Recognition Letters*, 73: 44–51.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2517–2526.

Wei, C.; Shen, K.; Chen, Y.; and Ma, T. 2020. Theoretical Analysis of Self-Training with Deep Networks on Unlabeled Data. In *International Conference on Learning Representations*.

Yao, B.; Jiang, X.; Khosla, A.; Lin, A. L.; Guibas, L.; and Fei-Fei, L. 2011. Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*, 1331–1338. IEEE.

Yu, F.; Wu, X.; Chen, J.; and Duan, L. 2019. Exploiting images for video recognition: heterogeneous feature augmentation via symmetric adversarial learning. *IEEE Transactions on Image Processing*, 28(11): 5308–5321.

Yu, F.; Wu, X.; Sun, Y.; and Duan, L. 2018. Exploiting images for video recognition with hierarchical generative adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 1107–1113.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1318–1327.