EXTERN: Leveraging Endo-Temporal Regularization for Black-box Video Domain Adaptation

Yuecong Xu1,*

Jianfei Yang2,*

Min Wu1

Xiaoli Li1

Lihua Xie2

Zhenghua Chen1,†

xuyu0014@e.ntu.edu.sg

yang0478@ntu.edu.sg

wumin@i2r.a-star.edu.sg

xlli@i2r.a-star.edu.sg

elhxie@ntu.edu.sg

chen0832@e.ntu.edu.sg

1Institute for Infocomm Research, A*STAR, Singapore

2School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Abstract

To enable video models to be applied seamlessly across video tasks in different environments, various Video Unsupervised Domain Adaptation (VUDA) methods have been proposed to improve the robustness and transferability of video models. Despite improvements made in model robustness, these VUDA methods require access to both source data and source model parameters for adaptation, raising serious data privacy and model portability issues. To cope with the above concerns, this paper firstly formulates Black-box Video Domain Adaptation (BVDA) as a more realistic yet challenging scenario where the source video model is provided only as a black-box predictor. While a few methods for Black-box Domain Adaptation (BDA) are proposed in image domain, these methods cannot apply to video domain since video modality has more complicated temporal features that are harder to align. To address BVDA, we propose a novel Endo and eXo-TEmporal Regularized Network (EXTERN) by applying mask-to-mix strategies and video-tailored regularizations: endo-temporal regularization and exo-temporal regularization, performed across both clip and temporal features, while distilling knowledge from the predictions obtained from the black-box predictor. Empirical results demonstrate the state-of-the-art performance of EXTERN across various cross-domain closed-set and partial-set action recognition benchmarks, which even surpassed most existing video domain adaptation methods with source data accessibility.

1. Introduction

Video Unsupervised Domain Adaptation (VUDA) [3, 5, 38, 39] aims to transfer knowledge from a labeled source video domain to an unlabeled target video domain, and has wide applications in scenarios where massive labeled video data may not be available. Despite their effectiveness in improving the robustness of video models [39], current VUDA methods require access to the source video data, raising serious concerns about data privacy and model portability issues [19, 20]. The Source-Free Video Domain Adaptation (SFVDA) [41] is subsequently formulated to learn a target model without the access of source data, but it still relies on the well-trained source model parameters which allows generative models to recover source videos [9].

To further cope with the video privacy issue, we formulate and study a more realistic and challenging video domain adaptation scenario termed the Black-box Video Domain Adaptation (BVDA) where the source video model is provided for adaptation only as a black-box predictor (e.g., API service). For example, in hospitals or banks, the anomaly action recognition is required but the videos are not allowed to be shared across different organizations. In this case, BVDA helps to derive an accurate model without access to both the parameters and data in the source domain.

Recently, there have been a few research efforts [21, 44] aiming at Black-box Domain Adaptation (BDA) for images. One representative work is DINE [20], where target features are extracted by obtaining pseudo-labels from the black-box predictors while applying structural regularizations [34, 35] that incorporates structural information in the target domain and encourage better model discriminability and generalization ability. However, structural regularizations of DINE are tailored for images that only contain spatial features. Different from images, characterized by the multi-modality na-
Figure 1. Clip features of target videos may be scattered, where videos contain indistinct semantic information. We augment target video domain with virtual temporal features through a novel mask-to-mix strategy, and apply endo-temporal regularization. The resulting temporal features are more discriminative and complies with the cluster assumption.

2. Related Work

(Video) Unsupervised Domain Adaptation ((V)UDA). UDA and VUDA aims to distill shared knowledge across a labeled source domain and an unlabeled target domain, which improves the robustness and transferability of deep learning models. (V)UDA methods could be broadly categorized into three categories: i) reconstruction-based methods [8, 43], where domain-invariant features are extracted by encoders trained with data-reconstruction objectives, commonly constructed based on a encoder-decoder structure; ii) discrepancy-based methods [29, 42, 48], where domain alignment is achieved by applying metric learning approaches, optimized with metric-based objectives such as MDD [48], CORAL [31], and MMD [24]; and iii) adversarial-based methods [3, 32, 39], where methods leverage additional domain discriminators along with feature generators, trained jointly in an adversarial manner [13] by minimizing adversarial losses [7]. Compared to UDA research which is primarily focused on image-based tasks, VUDA research lags behind owing to the challenges
brought by aligning temporal features. Despite the challenges, there has been a substantial increase in research for VUDA, backed by the introduction of various cross-domain closed-set or partial set video datasets [3, 38, 42]. Regardless of the improvements in video model robustness and transferability, VUDA approaches all require access to both source data and source model parameters during adaptation, which could raise serious privacy concerns given the amount of private information of the subjects and scenes contained in videos.

Black-box Domain Adaptation (BDA). With the increased importance of data privacy with concerns of possible data leakage through white-box attack given model parameters [49], there have been a few research that explores BDA with images. BDA enables image models to be adapted to the unlabeled target domain without either the source data or the source model parameters, with the source model provided only as a black-box predictor. Under such settings, BBSE [23] focused on black-box label shift, and requires a hold-out source set for class confusion matrix estimation, which is commonly unavailable. More recently, LNL [47] proposed to tackle BDA by an iterative noisy learning approach via pseudo labels that are refined with KL divergence, while DINE [20] leveraged knowledge distillation with information maximization and structural regularization. Despite the above recent advances, BVDA has never been explored, which is more challenging as temporal features must also be aligned. We propose to tackle BVDA by applying a temporal feature tailored endo-temporal regularization leveraging a mask-to-mix strategy, along with exo-temporal regularization.

3. Methodology

For Black-box Video Domain Adaptation (BVDA), we only have access to a black-box video predictor $H_S$ (i.e., the constraint Source API) which is trained from the labeled source video domain $D_S = \{(V_{iS}, y_{iS})\}_{i=1}^{n_S}$ containing $n_S$ videos, with $V_{iS} \in \mathcal{V}_S, y_{iS} \in \mathcal{Y}_S$. We are also given an unlabeled target video domain $D_T = \{V_{iT}\}_{i=1}^{n_T}$ with $n_T$ i.i.d. videos $V_{iT} \in \mathcal{V}_T$. Additionally, we assume that source and target video domains share the same label space with $C$ classes, i.e., $\mathcal{Y}_S = \mathcal{Y}_T$ while the source and target videos follow different data distributions. Therefore, there exists a domain shift [1] between $D_S$ and $D_T$. The objective of BVDA is to learn a mapping model $\mathcal{Y}_T \rightarrow \mathcal{Y}_T$ to perform the action recognition task on the unlabeled $D_T$ while both $D_S$ and the parameters of $H_S$ are not accessible.

Constrained by the absence of both the source data and the parameters of the source model, neither VUDA methods nor SFVDA methods could be directly applied for BVDA. To tackle BVDA, we resort to an alternative strategy where we adapt target models to the embedded semantic information of the source data resorting to the hard or soft predic-

3.1. Backbone Network

Videos differ from images greatly due to the existence of temporal features. A key prerequisite for the target network to be adapted to the embedded source semantic information is that its backbone could extract temporal features explicitly. An efficient and popular approach constructs the temporal features explicitly with a series of local temporal features (also termed clip features), obtained through clips sampled from the corresponding videos [36, 50]. One notable example is the Temporal Relation Network (TRN) [50]. TRN has been widely adopted in previous video domain adaptation tasks, such as VUDA [3, 39], PVDA [38] and SFVDA [41] bringing state-of-the-art results, thanks to its ability in extracting accurate temporal features by reasoning over correlations between spatial representations which coincides with the process of humans recognizing actions.

Formally, we define an input target video with $k$ frames as $V_i = \{f_i^{(1)}, f_i^{(2)}, ..., f_i^{(k)}\}$, with $f_i^{(j)}$ being the spatial feature of the $j$-th frame in the $i$-th source video. The subscript for the target domain $T$ is omitted for clarity. The spatial features are extracted from the source spatial feature generator $G_{sp}$ which is formulated as a 2D-CNN (e.g., ResNet-50 or ResNet-101 [11]). Subsequently, the temporal feature of $V_i$ is constructed by a combination of multiple clip features, obtained from the temporal feature generator $G_t$. More specifically, $G_t$ builds each clip feature with $r$ temporal-ordered sampled frames where $r \in [2, k]$. Formally, a clip feature for $V_i$, denoted as $c_i^{(r)}$, is defined by:

$$c_i^{(r)} = \sum_m g^{(r)}((V_i^{(r)})_m).$$  \hspace{1cm} (1)

Here $(V_i^{(r)})_m = \{f_i^{(a)}, f_i^{(b)}, ..., \}$ is defined as the $m$-th clip with $r$ temporal-ordered frames, where $a$ and $b$ are the frame indices. $a$ and $b$ may not be consecutive as the clip could be extracted with nonconsecutive frames, but should be both constrained within the range of $[1, k]$ with $b > a$. Eventually, the clip feature $c_i^{(r)}$ is obtained by fusing all the $r$ time ordered frame-level spatial features through an integration function $g^{(r)}$, usually formulated as a Multi-Layer Perceptron (MLP). The above computation would result in a total of $(k - 1)$ clips. Finally, the temporal feature of $V_i$.
denoted as $t_i$, is computed through the mean aggregation applied across all clip features, defined as:

$$t_i = \frac{1}{k-1} \sum_r c_{cl}^{(r)}.$$  

(2)

### 3.2. Endo and Exo-Temporal Regularizations

Current methods for BDA [20] attempt to obtain features from the unlabeled target data with high discriminability via self-supervised learning by applying strong structural regularizations. However, such regularizations are only tailored over the image-based BDA tasks. For temporal features that are constructed with clip features, applying such regularizations may not guarantee discriminative clip features that meet the cluster assumption. In such cases, the temporal features contain indistinct semantic information as depicted in Fig. 1. To ensure that temporal features are effective, their corresponding clip features should also be discriminative. In other words, the cluster assumption should not only be applied to the final temporal features $t_i$, but also all the corresponding clip features $c_{cl}^{(r)}$. The ultimate objective is to improve the discriminability of both the temporal features as well as the clip features.

Previous studies [46] suggest that the discriminability of extracted data features can be improved if the feature generator behaves linearly in-between training samples. This inspires us to propose the endo-temporal regularization over the clip features leveraging the MixUp [46] operation such that the target model behaves linearly in-between clip features. The proposed regularization is designed to be simple yet effective in enhancing the discriminability of target domain temporal features.

To achieve this, we first augment the target domain with virtual temporal features with a mask-to-mix strategy as depicted in Fig. 2. We define the virtual temporal feature of $V_i$ as $\tilde{t}_i$. It is constructed by mixing a set of clip features $c_{cl}^{(r)}$, $r \in [2, k]$ selected by a random masking process. Specifically, if the temporal feature $t_i$ is built upon a set of $(k-1)$ clips, there will be exactly $(k - 3)$ clips masked randomly, leaving two randomly unmasked clips: $c_{cl}^{(r_1)}$ and $c_{cl}^{(r_2)}$ where $r_1, r_2 \in [2, k]$. For each mini-batch, the selection of masked clips is random across all input videos within the mini-batch and across each epoch. This is to ensure that the virtual temporal feature is built upon all possible clip pairs across the whole training process. Different from the temporal feature which is constructed through a simple mean aggregation where all clip features are combined with equal weights, we compute the virtual temporal feature via MixUp [46]. Essentially, the virtual temporal feature is computed as a linear combination of the two unmasked clips assigned with random weights. Formally, $\tilde{t}_i$ is computed as:

$$\tilde{t}_i = \text{MixUp}_{\lambda_v}(c_{cl}^{(r_1)}, c_{cl}^{(r_2)}) = \lambda_v c_{cl}^{(r_1)} + (1 - \lambda_v)c_{cl}^{(r_2)},$$  

(3)

where $\lambda_v \in Beta(\alpha_v, \alpha_v)$ is the weight assigned to $c_{cl}^{(r_1)}$ sampled from a Beta distribution with $\alpha_v$ as the hyper-
parameter. The virtual temporal feature constructed is essentially the linear intermediate representation between the unmasked clip features.

With the virtual temporal feature constructed, we drive the linearity behaviour of the target model by employing the interpolation consistency training (ICT) technique [34]. Specifically, such technique encourages the virtual temporal prediction to be consistent with the mixed unmasked clip prediction. Formally, the virtual temporal prediction is obtained from the target predictor directly, i.e., \( \hat{y}_i = H_T(t_i) \). The mixed unmasked clip prediction is computed as the linear combination of the target predictions from both unmasked clips, defined as

\[
y_{\text{mix},i} = \text{MixUp}_{\lambda_y}(H_T(cl_i^{(r_1)}), H_T(cl_i^{(r_2)})) = \lambda_y H_T(cl_i^{(r_1)}) + (1 - \lambda_y) H_T(cl_i^{(r_2)}).
\]

Subsequently, we aim to optimize the loss function:

\[
L_{\text{vir}} = l_{ce}(\hat{y}_i, y_{\text{mix},i}),
\]

where \( l_{ce} \) denotes the cross-entropy loss.

To further ensure that the temporal feature contains distinctive semantic information, it is observed that the clip features should be clustered towards the temporal feature. This is in line with the intuition that since both the virtual temporal feature and the temporal feature originate from the same target video (i.e., the same set of clip features), the prediction of the virtual temporal feature should be consistent with the prediction of the corresponding target video (or equivalently, the corresponding target temporal feature). To achieve such prediction consistency, we aim to minimize the divergence between the virtual temporal prediction (computed as \( \hat{y}_i \)) and the target prediction. The latter is obtained by applying the target predictor to the target temporal feature, i.e., \( y_i = H_T(t_i) \). The minimization of prediction divergence is formulated as:

\[
L_{\text{pre}} = D_{KL} (\hat{y}_i \| y_i),
\]

where \( D_{KL}(\cdot \| \cdot) \) denotes the Kullback-Leibler divergence.

Overall, the proposed endo-temporal regularization drives the target model to obtain discriminable temporal features by extracting clip features with higher discriminability that complies with the cluster assumption from two perspectives: i) extracting clip features that are linear in-between; and ii) aligning clip features towards the corresponding temporal feature. The endo-temporal regularization is applied by jointly optimizing Eq. 5 and Eq. 6 as:

\[
L_{\text{endo}} = L_{\text{vir}} + L_{\text{pre}}.
\]

The implementation of the endo-temporal regularization can be observed to be very simple, yet it brings significant improvements towards tackling BVDA, as would be presented in Sec. 4.

To further enhance the discriminability of the temporal feature, we extend the promotion of linear behavior in-between training samples towards the temporal feature. Given a pair of videos \( V_i, V_j \), we employ the ICT [34] across their corresponding temporal features \( t_i, t_j \), formulating the exo-temporal regularization term. Such operation is equivalent to augmenting the target video domain with interpolated temporal features which would drive the model towards better generalization. Formally, similar to how virtual temporal features are constructed, the interpolated temporal features of \( t_i, t_j \) are obtained by applying MixUp [46]. With \( y_i = H_T(t_i) \) and \( y_j = H_T(t_j) \) denoting the target predictions of \( t_i \) and \( t_j \), the exo-temporal regularization aims to optimize the loss function:

\[
L_{\text{exo}} = l_{ce}(\text{MixUp}_{\lambda_y}(t_i, t_j), \text{MixUp}_{\lambda_y}(y_i, y_j)),
\]

where \( l_{ce} \) is the cross-entropy loss. MixUp\(_{\lambda_y} \) is defined as shown in Eq. 3 and 4, and \( \lambda_y \in \text{Beta}(\alpha_t, \alpha_t) \) is the weight assigned to \( t_i \) with \( \alpha_t \) as the hyper-parameter.

### 3.3. Endo and eXo-TEmporal Regularized Network

With the endo-temporal regularization and exo-temporal regularization terms defined, we propose the EXTERN to address BVDA leveraging on both regularizations, as depicted in Fig. 2. EXTERN builds upon the TRN backbone structure as specified in Sec. 3.1.

**Extracting Knowledge via Knowledge Distillation.** To extract knowledge from the black-box predictor \( H_S \), knowledge distillation (KD) [12] has proven to be an effective solution. The target model is seen as the student, and is trained to learn predictions analogous to that produced by the source model, which is seen as the teacher. However, due to domain shift between the source domain \( D_S \) and target domain \( D_T \), the output from the source model could be noisy and inaccurate. To cope with such drawback, the adaptive label smoothing (AdaLS) technique [20] with self-distillation [18] exponential moving average (EMA) [10] update is recently proposed. Here, only the predictions of the top-\( c \) classes are directly utilized while predictions of other classes are forced to a uniform distribution as:

\[
\hat{y}_i^p = \text{AdaLS}_c(\hat{y}_i) = \begin{cases} 
\frac{\hat{y}_i^p}{1 - \sum_{c' \in T_c} \hat{y}_{i}^{p}} & p \in T_c \\
\frac{p - \sum_{c' \in T_c} \hat{y}_{i}^{c}}{c - c} & \text{otherwise}
\end{cases},
\]

where \( \hat{y}_i \in \hat{Y} \) is the prediction of the target video \( V_i \) obtained from the black-box source predictor \( H_S \) (i.e., the teacher prediction), while \( \hat{y}_i^p \) denotes the prediction of the \( p \)-th class and \( T_c \) is the class index set of the top-\( c \) predictions for input \( V_i \). The teacher prediction is further dynamically updated per training epoch to maintain a EMA prediction.
We apply AdaLS with EMA to reduce noisy information by focusing only the top-predicted classes. Extracting source knowledge is eventually achieved by optimizing:

$$L_{kd} = \mathbb{E}_{V \in \mathcal{D}_s} D_{KL}(y'_i \parallel y_i).$$

\[ (10) \]

**Learning Adaptive Clip Weights.** Previous research [3, 26] shows that features with lower prediction uncertainty would possess higher discriminability. Therefore, to better aggregate the clip features for the temporal feature, we assign a *clip weight* to each clip feature by attending to clip features with lower prediction uncertainty. Specifically, the clip weight is defined as the additive inverse of the target predictions of the corresponding clip, computed as:

$$w_{cl}(c) = 1 - h(H_T(c_{cl}^{(c)})), \quad (11)$$

where the constant 1 is added as a residual connection for more stable optimization. Subsequently, the temporal feature is obtained as the weighted aggregation of all clip features, with Eq. 2 modified as:

$$t_i = \frac{1}{k-1} \sum_r w_r cl_i c_r^{(r)}. \quad (12)$$

**Information Maximization.** Inspired by prior works in BDA [20, 44], we maximize the mutual information (MI) to encourage diversity among target predictions and to promote their individual certainty:

$$L_{mi} = h(\mathbb{E}_{V_i \in \mathcal{D}_s} y_i) - \mathbb{E}_{V_i \in \mathcal{D}_s} h(y_i), \quad (13)$$

where \( y_i = H_T(G_{sp}(G_t(V_i))) = H_T(t_i) \) is the target prediction for input \( V_i \) and \( h(y_i) = -\sum_{c=1}^{C} y_i^c \log y_i^c \) is the conditional entropy function. Maximizing MI could marginally improve the performances for BVDA, as would be shown later in the ablation studies (Sec. 4.3).

**Overall Objective.** Summarizing all the loss functions as presented in Eqs. (7, 8, 10, 13), the overall optimization objective of EXTERN is expressed as:

$$L = L_{kd} + \beta_{reg}(L_{endo} + L_{exo}) - L_{mi}, \quad (14)$$

where \( \beta_{reg} \) is the hyper-parameter that controls the strength of the regularizations and is empirically set to 1. We refer to the settings of DINE [20] by setting \( \alpha_{t} \) as 0.3 and \( c \) as 3.

4. Experiments

In this section, we evaluate our proposed EXTERN across a variety of cross-domain action recognition benchmarks, covering a wide range of cross-domain scenarios. We demonstrate exceptional performances on all benchmarks. Moreover, thorough ablation studies and analysis of EXTERN are performed to further justify the design of EXTERN.
Table 1. Results for BDVA on UCF-HMDB$_{full}$ and Sports-DA for closed-set video domain adaptation.

| Methods       | Privacy | UCF-HMDB$_{full}$ | Sports-DA |
|---------------|---------|-------------------|-----------|
|               | TRN [50]| ECCV-18           |           |
|               |         | -                 | -         |
| LNL [47]      | ✓ ✓     | -                 | -         |
| HD-SHOT [23]  | ✓ ✓     | -                 | -         |
| SD-SHOT [21]  | ✓ ✓     | -                 | -         |
| DINE [13]     | ✓ ✓     | -                 | -         |
| EXTERN        | ✓ ✓     | -                 | -         |
| TA N [3]      | ✓ ✓     | -                 | -         |
| DANN [7]      | ✓ ✓     | -                 | -         |
| MK-MMD [24]   | ✓ ✓     | -                 | -         |
| SAVA [5]      | ✓ ✓     | -                 | -         |
| SHOT [1]      | ✓ ✓     | -                 | -         |
| ACAN [9]      | ✓ ✓     | -                 | -         |
| ATOCn [61]    | ✓ ✓     | -                 | -         |

Table 2. Results for BDVA on Daily-DA for closed-set video domain adaptation.

| Methods       | Privacy | Daily-DA |
|---------------|---------|----------|
|               | TRN [50]| ECCV-18  |
| LNL [47]      | ✓ ✓     |        |
| HD-SHOT [23]  | ✓ ✓     |        |
| SAVA [5]      | ✓ ✓     |        |
| SHOT [1]      | ✓ ✓     |        |
| ACAN [9]      | ✓ ✓     |        |
| ATOCn [61]    | ✓ ✓     |        |

and Daily-DA benchmarks showing inferior performance to that of the source-only model. Prior BDA approaches focused solely on spatial features, and may not obtain temporal feature with distinct semantic information, resulting in negative impacts compared to the source-only baseline. Notably, EXTERN even outperforms various VUDA approaches [3,5,39] with source data accessibility. Since EXTERN performs adaptation based solely on prediction results, it is found that EXTERN would not be affected by noise contained within the source data or source model, generating superior adaptation results. This shows that training a target model from scratch with strong regularization while adapting solely with source predictions can be as effective as data-based domain alignment techniques.

Partial-set domain adaptation. Apart from closed-set video domain adaptation, we further demonstrate the generalization ability of our proposed EXTERN by evaluating on partial-set video domain adaptation (PVDA) tasks. To achieve this, we follow [38] and leverage on three other benchmarks: UCF-HMDB$_{partial}$, HMDB-ARID$_{partial}$ and MiniKinetics-UCF.

UCF-HMDB$_{partial}$ is built from UCF101 and HMDB51 from 14 overlapping categories with 2,780 videos. The first 7 categories in alphabetic order are chosen as target categories, resulting in two PVDA tasks: H-10$\rightarrow$A-5 and A-10$\rightarrow$H-5. MiniKinetics-UCF is a large-scale dataset built from MiniKinetics [37] and UCF101 containing 45 overlapping categories with 22,102 videos. The first 18 categories in alphabetic order are chosen as target categories, resulting in two PVDA tasks: M-45$\rightarrow$U-18 and U-45$\rightarrow$M-18.

The results for partial-set cross-domain action recognition is presented in Tab. 3. Partial-set is more challenging due to the asymmetric label spaces with the existence of “source-only” classes, causing negative transfer [2,38]. Negative transfer affects all previous BDA approaches, where all approaches will under-perform against the source-only baseline in at least one benchmark. Despite such challenge, EXTERN still achieves outstanding results, outperforming the best BDA approach by a relative 8.0%, 23.1% and 6.2% on the three benchmarks respectively. EXTERN also surpasses several PDA approaches even though EXTERN is not specifically catered for label shift.

4.3. Analysis

Ablation studies. To gain deeper understanding of the effectiveness of EXTERN while justifying its design, we perform detailed ablation studies as shown in Tab. 4. Specifically, the ablations study explore EXTERN from three perspectives: firstly, the effects of key learning objectives that constitute the regularization terms; secondly the effects of assigning clip weights; and lastly the effects of applying information maximization. Ablation studies are con-
Table 3. Results for BVDA on UCF-HMDB\textsubscript{partial}, HMDB-ARID\textsubscript{partial} and MiniKinetics-UCF for partial-set video domain adaptation.

| Methods       | UCF-HMDB\textsubscript{full}   | UCF-HMDB\textsubscript{partial} | Avg.   |
|---------------|-------------------------------|---------------------------------|--------|
| EXTERN (w/o $L_{\text{endo}}$) | $\times$                     | 78.25                           | 85.64  | 60.72                           | 81.09                           | 76.42                           |
| EXTERN (w/o $L_{\text{ex}}$)    | $\checkmark$                 | 80.08                           | 86.34  | 61.91                           | 82.71                           | 77.76                           |
| EXTERN (w/o $L_{\text{vir}}$)   | $\times$                     | 84.65                           | 90.11  | 66.62                           | 87.59                           | 82.24                           |
| EXTERN (w/o $L_{\text{pre}}$)   | $\checkmark$                 | 85.83                           | 90.57  | 67.71                           | 88.91                           | 83.21                           |
| EXTERN (w/o $L_{\text{mi}}$)    | $\times$                     | 82.66                           | 89.49  | 64.52                           | 86.47                           | 80.78                           |
| EXTERN (w/o $L_{\text{reg}}$)   | $\checkmark$                 | 83.89                           | 89.93  | 65.24                           | 88.16                           | 81.80                           |
| EXTERN (w/o $L_{\text{mi}}$)    | $\times$                     | 86.47                           | 90.63  | 68.11                           | 88.23                           | 83.36                           |
| EXTERN (w/o $L_{\text{mi}}$)    | $\checkmark$                 | 87.22                           | 90.89  | 69.05                           | 89.47                           | 84.16                           |
| EXTERN (w/o $L_{\text{mi}}$)    | $\times$                     | 87.56                           | 91.02  | 70.07                           | 89.69                           | 84.75                           |
| EXTERN (w/o $L_{\text{mi}}$)    | $\checkmark$                 | 88.43                           | 91.64  | 71.06                           | 90.34                           | 85.36                           |
| EXTERN         | $\times$                     | 87.92                           | 91.33  | 70.38                           | 90.08                           | 84.93                           |
| EXTERN         | $\checkmark$                 | 88.89                           | 91.95  | 71.43                           | 90.60                           | 85.72                           |

Table 4. Ablation studies of learning objectives and clip weights on UCF-HMDB\textsubscript{full} and UCF-HMDB\textsubscript{partial}.

![t-SNE visualizations of clip features and temporal features extracted by different models](image)

Figure 3. t-SNE [33] visualizations of clip features and temporal features extracted by (a) source-only model, (b) DINE [20], and (c) EXTERN, with class information. Different colors denote different classes.

![Sensitivity to $\beta_{\text{reg}}$ on U101→H51](image)

![Sensitivity to $\alpha_v$ on U101→H51](image)

Figure 4. Sensitivity of $\beta_{\text{reg}}$ and $\alpha_v$ on UCF-HMDB\textsubscript{full}.

The trained models were further conducted on UCF-HMDB\textsubscript{full} and UCF-HMDB\textsubscript{partial} with the same TRN backbone as previous experiments. With any of the five learning objectives (i.e., $L_{\text{endo}}$, $L_{\text{ex}}$, $L_{\text{vir}}$, $L_{\text{pre}}$ and $L_{\text{mi}}$) removed, the respective variants of EXTERN perform inferior to the original EXTERN, thus justifying that the designed learning objectives complement each other. It could be observed that by applying the *endo-temporal regularization* alone (variant EXTERN (w/o $L_{\text{ex}}$)), EXTERN could outperform all prior BDA approaches. This further prove the effectiveness of *endo-temporal regularization*.
ization since this regularization is tailored to temporal features. Meanwhile, the improvements brought by constructing temporal features attentively with clip weight is consistent among all EXTERN variants. However, such improvement is marginal compared to that brought by the temporal feature-tailored regularization. Similarly, the improvements from applying information maximization is also consistent, yet even more negligible than that brought by the application of clip weight.

**Hyper-parameter Sensitivity.** We focus on studying the hyper-parameter sensitivity of $\beta_{\text{reg}}$ which controls the strength of the regularizations and $\alpha_v$ which relates to the construction of virtual temporal feature $\hat{t}_v$. Here $\beta_{\text{reg}}$ is in the range of 0.2 to 1.2 and $\alpha_v$ is in the range of 0.1 to 0.9. As shown in Fig. 4, the results of EXTERN falls within a marginal 0.83%, ranging from 88.06% to 88.89%, with the best results obtained at $\alpha_v = 0.3$ and $\beta_{\text{reg}} = 1.0$. The minimal variations shows that the performance of EXTERN is robust to both hyper-parameters. Meanwhile, despite the slight variations, EXTERN maintains the best results with all the hyper-parameter settings.

**Feature Visualization.** We further understand the characteristics of EXTERN by plotting the t-SNE embeddings [33] of both the clip features and temporal features extracted by the source-only model, DINE and EXTERN for the HS1→U101 task, as displayed in Fig. 3. It is clearly observed that both the clip features and temporal features from EXTERN are more clustered and discriminable, justifying that the applied regularizations promotes higher discriminability and better compliance to the cluster assumption. We can also observe that the distribution of clip features is more similar to the distribution of temporal features with EXTERN. This intuitively proves that EXTERN aligns clip features towards the temporal features which further ensures that the temporal features contain distinct semantic information with high discriminability.

**5. Conclusion**

In this work, we pioneer in formulating and exploring the realistic yet more challenging task of Black-box Video Domain Adaptation (BVDA) for privacy-preserving and portable video model transfer. To tackle BVDA, we propose EXTERN which obtains effective and discriminative temporal features by applying a novel endo-temporal regularization following a mask-to-mix strategy, along with an exo-temporal regularization. EXTERN drives the clip features to better compliance of the cluster assumption while ensuring the temporal features contain distinct semantic information. Extensive empirical results across cross-domain action recognition benchmarks under both closed-set and partial-set domain adaptation settings with detailed analysis justifies the efficacy of EXTERN in tackling BVDA. We believe such the superior performance of EXTERN whose adaptation is based solely on source prediction without access to source data and model could pave a new way for tackling video domain adaptation.

**References**

[1] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144, 2006.

[2] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. Partial adversarial domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.

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

[4] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019.

[5] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 678–695. Springer, 2020.

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

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

[8] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pages 597–613. Springer, 2016.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[10] Denis S Grebenkov and Jeremy Serror. Following a trend with an exponential moving average: Analytical results for a gaussian model. *Physica A: Statistical Mechanics and its Applications*, 394:288–303, 2014.

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

[12] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[13] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning.
In Proceedings of the 4th ACM workshop on Security and artificial intelligence, pages 43–58, 2011.

[14] Andrey Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1725–1732, 2014.

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In 2011 International Conference on Computer Vision, pages 2556–2563. IEEE, 2011.

[17] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaanhs Bhandri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, and Venkatesh Babu Radhakrishnan. Balancing discriminability and transferability for source-free domain adaptation. In International Conference on Machine Learning, pages 11710–11728. PMLR, 2022.

[18] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017.

[19] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In International Conference on Machine Learning, pages 6028–6039. PMLR, 2020.

[20] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[21] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[22] Jian Liang, Yunbo Wang, Dapeng Hu, Ran He, and Jiashi Feng. A balanced and uncertainty-aware approach for partial domain adaptation. In European Conference on Computer Vision, pages 123–140. Springer, 2020.

[23] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In International conference on machine learning, pages 3122–3130. PMLR, 2018.

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

[25] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. IEEE transactions on pattern analysis and machine intelligence, 42(2):502–508, 2019.

[26] Dang-Khoa Nguyen, Wei-Lun Tseng, and Hong-Han Shuai. Domain-adaptive object detection via uncertainty-aware distribution alignment. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2499–2507, 2020.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In Advances in neural information processing systems, pages 8026–8037, 2019.

[28] Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. Journal of Machine Learning Research, 8(7), 2007.

[29] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3723–3732, 2018.

[30] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[31] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 30, 2016.

[32] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 7167–7176, 2017.

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

[34] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In International Joint Conference on Artificial Intelligence, pages 3635–3641, 2019.

[35] Paul Viola and William M Wells III. Alignment by maximization of mutual information. International journal of computer vision, 24(2):137–154, 1997.

[36] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence, 41(11):2740–2755, 2018.

[37] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. arXiv preprint arXiv:1712.04851, 1(2):5, 2017.

[38] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Aligning correlation information for video understanding. IEEE transactions on pattern analysis and machine intelligence, 42(2):502–508, 2019.
for domain adaptation in action recognition, 2021. 1, 2, 3, 6, 7

[40] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiang Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In International Workshop on Deep Learning for Human Activity Recognition, pages 70–84. Springer, 2021. 6

[41] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Wu Min, and Zhenghua Chen. Learning temporal consistency for source-free video domain adaptation. arXiv preprint arXiv:2203.04559, 2022. 1, 2, 3, 6, 7

[42] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, Rui Zhao, and Zhenghua Chen. Multi-source video domain adaptation with temporal attentive moment alignment. arXiv preprint arXiv:2109.09964, 2021. 2, 3, 6

[43] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In European Conference on Computer Vision, pages 480–498. Springer, 2020. 2

[44] Jianfei Yang, Xiangyu Peng, Kai Wang, Zheng Zhu, Jiashi Feng, Lihua Xie, and Yang You. Divide to adapt: Mitigating confirmation bias for domain adaptation of black-box predictors. arXiv preprint arXiv:2205.14467, 2022. 1, 6

[45] Jianfei Yang, Han Zou, Yuxun Zhou, Zhaoyang Zeng, and Lihua Xie. Mind the discriminability: Asymmetric adversarial domain adaptation. In European Conference on Computer Vision, pages 589–606. Springer, 2020. 2

[46] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. International Conference on Learning Representations, 2018. 4, 5

[47] Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models. arXiv preprint arXiv:2101.02839, 2021. 3, 6, 7, 8

[48] Yuchen Zhang, Tianle Liu, Mingheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In International Conference on Machine Learning, pages 7404–7413. PMLR, 2019. 2

[49] Yinghui Zhang, Yangqiu Song, Jian Liang, Kun Bai, and Qiang Yang. Two sides of the same coin: White-box and black-box attacks for transfer learning. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2989–2997, 2020. 3

[50] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In Proceedings of the European Conference on Computer Vision (ECCV), pages 803–818, 2018. 2, 3, 6, 7, 8