Contrast and Mix: Temporal Contrastive Video Domain Adaptation with Background Mixing

Anonymous Author(s)
Affiliation
Address
email

Abstract

Unsupervised domain adaptation which aims to adapt models trained on a labeled source domain to a completely unlabeled target domain has attracted much attention in recent years. While many domain adaptation techniques have been proposed for images, the problem of unsupervised domain adaptation in videos remains largely underexplored. In this paper, we introduce Contrast and Mix (CoMix), a new contrastive learning framework that aims to learn discriminative invariant feature representations for unsupervised video domain adaptation. First, unlike existing methods that rely on adversarial learning for feature alignment, we utilize temporal contrastive learning to bridge the domain gap by maximizing the similarity between encoded representations of an unlabeled video at two different speeds as well as minimizing the similarity between different videos played at different speeds. Second, we propose a novel extension to the temporal contrastive loss by using background mixing that allows additional positives per anchor, thus adapting contrastive learning to leverage action semantics shared across both domains. Moreover, we also integrate a supervised contrastive learning objective using target pseudo-labels to enhance discriminability of the latent space for video domain adaptation. Extensive experiments on several benchmark datasets demonstrate the superiority of our proposed approach over state-of-the-art methods.

1 Introduction

Unsupervised domain adaptation (UDA), which alleviates the requirement of large amounts of annotated data by adapting a model learned on a labelled source domain to an unlabelled target domain, has drawn a great deal of attention in the last few years [13, 80]. Much progress has been made in developing deep UDA methods by minimizing the cross-domain divergence [40, 70], adding adversarial domain discriminators [21, 74], and image-to-image translation techniques [27, 52]. However, despite impressive results on commonly used benchmark datasets (e.g., [61, 75, 57]), most of the methods have been developed only for images and not for videos, where the annotation task is often more complicated requiring tedious human labor as compared to images.

More recently, very few works have attempted deep UDA for video action recognition by directly matching segment-level features [8, 28, 51, 43] or with attention weights [12, 54]. However, (1) trivially matching segment-level feature distributions by extending the image-specific approaches, without considering the rich temporal information may not alone be sufficient in learning domain invariant features for video domain adaptation; (2) prior methods often focus on aligning target features with source, rather than exploiting any action semantics shared across both domains (e.g., difference in background with the same action: videos in the top row of Figure 1 are from the source and target domain respectively, but both capture the same action *walking*); (3) existing methods often rely on complex adversarial learning which is unwieldy to train, resulting in very fragile convergence.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Meanwhile, self-supervised pretext tasks like predicting rotation and translation have recently emerged as an alternative to adversarial learning for unsupervised domain adaptation in images [39, 71]. While these works show the promising potential of self-supervised learning in aligning source and target domains, the more recent very successful contrastive representation learning [9, 24, 53] has never been used to adapt video action recognition models to target domains. Motivated by this, in this paper, we explore the following natural, yet important question: *whether and how contrastive learning could be exploited for the challenging and practically important task of unsupervised video domain adaptation for human action recognition?*

To this end, we introduce Contrast and Mix (`CoMix`), a simple yet effective approach based on contrastive learning to adapt video action recognition models trained on a labeled source domain to unlabelled target domains. First, we propose to represent video as a graph and then utilize temporal contrastive self-supervised learning over the graph representations as a nexus between source and target domains to align features, without requiring any additional adversarial learning, as most prior works do in video domain adaptation [8, 12, 54]. Specifically, we maximize the similarity between encoded representations of the same video at two different speeds as well as minimize the similarity between different videos played at different speeds, leveraging the fact that changing video speed does not change an action on both domains. While minimization of contrastive self-supervised losses in both domains simultaneously helps in domain alignment, it ignores action semantics shared across them as the loss treats each domain individually. To alleviate this, we incorporate new synthetic videos into the temporal contrastive objective, which are obtained by mixing background of a video from one domain to a video from another domain, as shown in Figure 1 (bottom). Importantly, since mixing background doesn't change the temporal dynamics, we introduce pseudo-labels for the mixed videos to be same as the label of the original videos and consider additional positives per anchor (see Figure 2), which encourages the model to generalize to new samples that may not be covered by temporal contrastive learning in hand. In other words, mixed background video of an input sample in the embedding space act as small semantic perturbations that are not imaginary, i.e., they are representative of the action semantics shared across source and target domains. Finally, rather than relying only on the supervision of source categories to learn a discriminative representation, we generate pseudo-labels for the target samples in every batch and then harness the label information using a temporal supervised contrastive term, that pushes the examples from the same class close and the examples from different classes further apart (Figure 2: right). While our modified contrastive losses are motivated by the supervised contrastive learning [31], we use pseudo labels for exploiting shared action semantics and discriminative information from target domain, instead of using true labels as an alternative to supervised cross-entropy loss (which is not present for target samples). To the best of our knowledge, ours is the first work that successfully leverage contrastive learning in an unified framework to align cross-domain features while enhancing discriminabilty of the latent space for unsupervised video domain adaptation.

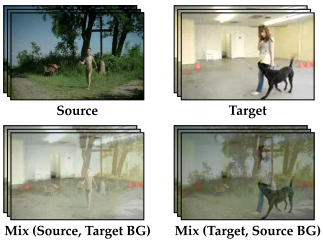

Figure 1: **Background Mixing.** Top row shows two representative videos from the source and target domain respectively. Both videos capture the same action "*walking*" with different backgrounds. Bottom row shows videos obtained after mixing target background with source video and vice versa.

To summarize, the main **contributions** of our work are as follows:

- We introduce Contrast and Mix (`CoMix`), a new contrastive learning framework to learn discriminative invariant feature representations for unsupervised video domain adaptation. Overall, `CoMix` is simple and easy to implement which perfectly fits into modern mini-batch end-to-end training.
- We propose a novel extension to temporal contrastive loss by using background mixing that allows additional positives per anchor, thus adapting contrastive learning to leverage action semantics shared across both domains. We also integrate a supervised contrastive learning objective using pseudo label information from the target domain to enhance discriminabilty of the latent space.
- We conduct extensive experiments on several challenging benchmarks (UCF-HMDB [8], Jester [54], and Epic-Kitchens [51]) for video domain adaptation to demonstrate the superiority of our approach over state-of-the-art methods. Our experiments show that `CoMix` delivers a significant performance increase over the compared methods, e.g., `CoMix` outperforms SAVA [12] (ECCV'20) by 3.6% on UCF-HMDB [8] and TA^{3}N [8] (ICCV'19) by 9.2% on Jester [46] benchmark respectively).

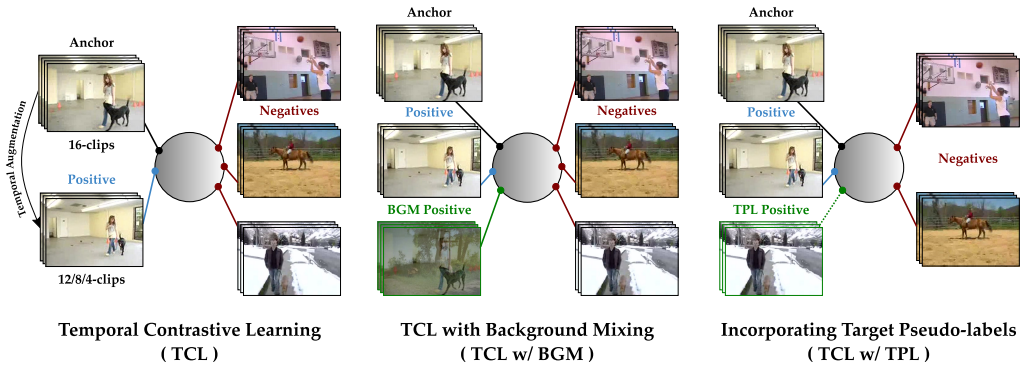

Figure 2: **Temporal Contrastive Learning with Background Mixing and Target Pseudo-labels.** Temporal contrastive loss (left) contrasts a single temporally augmented positive (same video, different speed) per anchor against rest of the videos in a mini-batch as negatives. Incorporating background mixing (middle) provides additional positives per anchor possessing same action semantics with a different background alleviating background shift across domains. Incorporating target pseudo-labels (right) additionally enhances the discriminabilty by contrasting the target videos with the same pseudo-label as positives against rest of the videos as negatives.

2 Related Work

Action Recognition. Much progress has been made in developing a variety of ways to recognize video actions, by either applying 2D-CNNs [6, 38, 48, 79] or 3D-CNNs [4, 18, 23, 73]. Many successful architectures are usually based on the two-stream model [67], processing RGB frames and optical-flow in two separate CNNs with a late fusion in the upper layers [30]. SlowFast network [19] employs two pathways for recognizing actions by processing a video at different frame rates. Mitigating background bias in action recognition has also been presented in [11, 37]. Despite remarkable progress, these models critically depend on large labeled datasets which impose challenges for cross-domain action recognition. In contrast, our work focuses on unsupervised domain adaptation for action recognition, with labeled data in source domain, but only unlabeled data in target domain.

Unsupervised Domain Adaptation. Unsupervised domain adaptation has been studied from multiple perspectives (see reviews [13, 80]). Representative works minimize some measurement of distributional discrepancy [22, 40, 65, 70] or adopt adversarial learning [5, 21, 41, 56, 74] to generate domain-invariant features. Leveraging image translation [26, 27, 52] or style transfer [16, 91] is also another popular trend in domain adaptation. Deep self-training that focus on iteratively training the model using both labeled source data and generated target pseudo labels have been proposed in [47, 90]. Semi-supervised domain adaptation leveraging a few labeled samples from the target domain has also been proposed for many applications [15, 34, 63]. A very few methods have recently attempted video domain adaptation, using adversarial learning combined with temporal attention [8, 43, 54], multi-modal cues [51], and clip order prediction [12]. While existing video DA methods mainly rely on adversarial learning (which is often complicated and hard to train) in some form or other, they do not take any action semantics shared across domains into consideration. Our approach on the other hand, successfully leverages temporal contrastive learning to learn domain-invariant features while exploiting shared action semantics through background mixing for video domain adaptation. Recently, self-supervised tasks like predicting rotation and translation have been used for unsupervised domain adaptation and generalization, mainly for images [3, 39, 71]. By contrast, we focus on the more challenging problem of domain adaptation for human action recognition, where our goal is to align domains by learning consistent features representing different speeds of unlabeled videos. We further propose a temporal supervised contrastive loss to ensure discriminabilty by considering pseudo-labeling in an unified framework for video domain adaptation.

Contrastive Learning. Contrastive representation learning is becoming increasingly attractive due to its great potential to leverage large amount of unlabeled images [9, 17, 24, 49, 25, 53] and videos [20, 32, 55, 60, 59, 78]. Speed of a video is investigated for self-supervised [1, 29, 77, 86] and semi-supervised learning [68, 92] unlike the problem we consider in this paper. Recent works [84, 87] utilize contrastive learning with different augmentations for learning unsupervised representations of graph data. Contrastive learning has also been recently used in supervised settings, where labels are used to guide the choice of positive and negative pairs [31]. While our approach is inspired by these, we propose a novel temporal contrastive learning framework with background mixing for video domain adaptation, which to our best knowledge has not been explored in the literature.

Image Mixtures. Mixup regularization [89] and its variants [2, 76, 88] that train models on virtual examples constructed as convex combinations of pairs of images and labels have been used to improve the generalization of neural networks. Very few methods apply Mixup in domain adaptation, but mainly to stabilize the domain discriminator [62, 83, 85] or to smoothen the predictions [45]. Several works have recently leveraged the idea of different image mixtures [35, 66] for improving contrastive representation learning. Our proposed background mixing can be regarded as an extension of this line of research by adding background of a video from one domain to a video from another domain, to explore shared semantics while learning domain-invariant features for action recognition.

3 Proposed Method

Unsupervised video domain adaptation aims to improve the model generalization performance by transferring knowledge from a labeled source domain to an unlabeled target domain. Formally, we have a set of labelled source videos $\mathcal{D}_{source} = \{(\mathbf{V}^{i\{s\}}, y^i)\}_{i=1}^{N_S}$ and a set of unlabelled target videos $\mathcal{D}_{target} = \{\mathbf{V}^{i\{t\}}\}_{i=1}^{N_T}$, with a common label space $\mathcal{L}$. Given these data sets, our goal is to learn a single model for action recognition that performs well on previously unseen target domain videos.

Approach Overview. Figure 3 illustrates an overview of `CoMix`. Our action recognition model consists of a feature encoder $\mathcal{F}$ with a temporal graph encoder $\mathcal{G}$. Given a video, the feature encoder $\mathcal{F}$ first extracts clip-level features, and then a graph encoder $\mathcal{G}$ utilizes those features to model intrinsic temporal relations for providing a robust encoded representation for action recognition. `CoMix` adopts supervised learning on the source videos, as the labels are available, jointly with two novel temporal contrastive learning loss terms to align the features for domain adaptation. Specifically, we maximize the similarity of the encoded representation of the fast version of a video (represented by f clips) with that of the slow version of the same video (represented by s clips, where $s < f$) as well as minimize the similarity of the representations of different videos within each of the two domains. However, as temporal contrastive loss treats each domain individually, we further add two new sets of synthetic videos that contain source videos mixed with target background and vice versa, respectively for introducing the background variations among the videos while keeping the action semantics intact. Finally, we generate pseudo-labels for the target videos in every mini-batch and utilize them using another temporal supervised contrastive term that contrasts target videos with the same pseudo-label as positives to learn features discriminative for the target domain. We now describe each of our proposed components individually in detail in the following subsections.

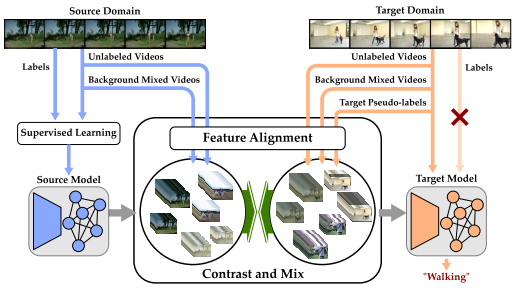

Figure 3: **An Overview of our Approach.** Given labeled videos in source domain and only unlabeled videos in target domain, `CoMix` adopts supervised learning on source videos, jointly with temporal contrastive learning on both domains to align features. Additional cross-domain contrastive supervision is obtained using background mixing across domains and using target pseudo-labels for enhancing discriminability of the latent space. `CoMix` provides a more simpler yet effective approach than adversarial learning for aligning both domains.

Video Representation. Capturing long-range temporal structure in videos is crucial for action recognition, which in turn affects the overall generalization performance of a model when adapting across domains. Thus, we adopt a graph convolutional neural network ($\mathcal{G}$) on top of a 3D convolutional neural network ($\mathcal{F}$) as our video feature encoder. Specifically, for a video $\mathbf{V}$ with n clips, the feature extractor $\mathcal{F}$ maps the clips into the corresponding sequence of features, which alone do not incorporate the rich temporal structure of the video. Therefore, we use the temporal graph encoder which constructs a fully connected graph on top of the clip-level features, with learnable edge weights through a parameterized adjacency matrix, as in [81]. With these graph representations, we apply a graph convolutional neural network with three layers and finally perform average pooling over all the node features to output the encoded representation of the video $\mathbf{V}$. In summary, the end-to-end network $\mathcal{G}(\mathcal{F}(.)) : \mathbf{V} \rightarrow \mathbb{R}^c$ takes a sequence of clips from a video as input and outputs confidence scores (logits) over the number of classes c for recognizing actions.

Temporal Contrastive Learning. Given video representations, our goal is to leverage contrastive self-supervised learning in both domains for unsupervised domain adaptation. To this end, we use temporal speed invariance in videos as a proxy task and enforce this with a pairwise contrastive loss.

Specifically, our key idea is to represent videos in two different temporal speeds (fast and slow) to obtain their encoded representations and then consider the fast and slow version representations of the same video to constitute positive pairs, while versions from different videos to constitute negative pairs. Formally, let us consider a mini-batch of B videos $\{\mathbf{V}_n^1, \mathbf{V}_n^2, ..., \mathbf{V}_n^B\}$ with corresponding feature representations $\{\mathbf{z}_n^1, \mathbf{z}_n^2, ..., \mathbf{z}_n^B\}$, where each of the videos $\mathbf{V}_n^i$ is represented using n number of sampled clips. Let f be the number of clips used to represent the fast version of the videos (forwarded through the base branch), and s be that used for the slow version (forwarded through the auxiliary branch), with $s < f$, as shown in Figure 4. Given positive and negative pairs, the model is trained such that it learns to maximize agreement between positive pairs, while minimizing agreement between negative pairs. This is achieved by employing a temporal contrastive loss ($\mathcal{L}_{tcl}$) as

$$\mathcal{L}_{tcl}(\mathbf{V}_f^i, \mathbf{V}_s^i) = -\log \frac{h(\mathbf{z}_f^i, \mathbf{z}_s^i)}{h(\mathbf{z}_f^i, \mathbf{z}_s^i) + \sum_{\substack{j=1, j\neq i \\ v\in\{s,f\}}}^{B} h(\mathbf{z}_f^i, \mathbf{z}_v^j)} \tag{1}$$

where, $h(\mathbf{u}, \mathbf{v}) = \exp(\frac{\mathbf{u}^\intercal \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}/\tau)$ is the exponential of cosine similarity measure and τ is the temperature hyperparameter [9]. We use $f = 16$, and choose s from $\{12, 8, 4\}$ following a random uniform distribution in every training iteration where randomness encourages the model to learn from a variety of temporal speed variations to learn robust representations.

Background Mixing. As temporal contrastive loss treats each domain individually, it ignores shared action semantics which is vital for domain alignment. Thus, we propose a new perspective of temporal contrastive loss through background mixing, specifically to alleviate the cross-domain background shift, as seen in Figure 1. The basic idea is to obtain the background frames for the videos in one domain and mix it with the frames of the videos from the other domain. This introduces variation in each of the domains by adding new synthetic videos with the same action semantics as earlier, but possessing background from the other domain. Given two videos $\mathbf{V}^{i\{s\}} \in \mathcal{D}_{source}$ and $\mathbf{V}^{i\{t\}} \in \mathcal{D}_{target}$ with corresponding background frames (single image per video) as $\mathbf{BG}^{i\{s\}}$ and $\mathbf{BG}^{i\{t\}}$, we obtain the synthetic videos in both domains by a convex combination of the background with each of the frames in the videos as follows.

$$\begin{aligned} \hat{\mathbf{V}}^{i\{s\}} &= (1-\lambda)\cdot \mathbf{V}^{i\{s\}} + \lambda \cdot \mathbf{BG}^{i\{t\}} \\ \hat{\mathbf{V}}^{i\{t\}} &= (1-\lambda)\cdot \mathbf{V}^{i\{t\}} + \lambda \cdot \mathbf{BG}^{i\{s\}} \end{aligned} \tag{2}$$

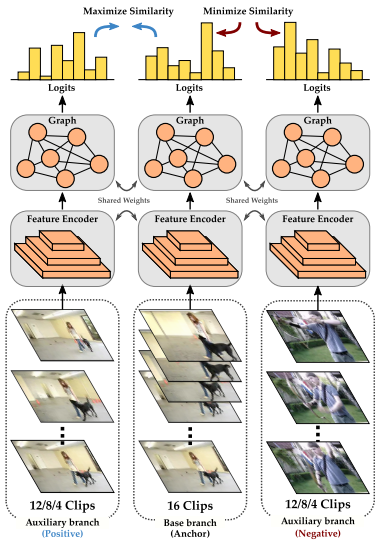

Figure 4: **Temporal Contrastive Loss.** Given unlabeled videos, we maximize the similarity between encoded representations of the same video at two different speeds (fast and slow) as well as minimize the similarity between different videos played at different speed.

where, λ is sampled from the uniform distribution $[0, \gamma]$, $\hat{\mathbf{V}}^{i\{s\}}$ and $\hat{\mathbf{V}}^{i\{t\}}$ correspond to the video from source domain with target background and vice versa, respectively. Since mixing background doesn't change the motion pattern of a video which actually defines an action, we assume both the original and mixed video to be of the same action class and go beyond single instance positives in Eq. 1 by adding additional positives per anchor, as in supervised contrastive learning [31] (see Figure 2). The modified temporal contrastive loss with background mixing ($\mathcal{L}_{bgm}$) is defined as below:

$$\mathcal{L}_{bgm}(\mathbf{V}_f^i, \mathbf{V}_s^i) = -\frac{1}{|\mathbf{P}(\mathbf{z}_f^i)|} \sum_{\mathbf{p}\in\mathbf{P}(\mathbf{z}_f^i)} \log \frac{h(\mathbf{z}_f^i, \mathbf{p})}{\sum\limits_{\mathbf{p}\in\mathbf{P}(\mathbf{z}_f^i)} h(\mathbf{z}_f^i, \mathbf{p}) + \sum\limits_{\substack{j=1, j\neq i \\ v\in\{s,f\}}}^{B} \left\{h(\mathbf{z}_f^i, \mathbf{z}_v^j) + h(\mathbf{z}_f^i, \hat{\mathbf{z}}_v^j)\right\}} \tag{3}$$

where, $\mathbf{P}(\mathbf{z}_f^i) \equiv \{\mathbf{z}_s^i, \hat{\mathbf{z}}_s^i, \hat{\mathbf{z}}_f^i\}$ is the set of positives for the anchor $\mathbf{z}_f^i$, and $\hat{\mathbf{z}}_{s/f}^i$ represent the feature representation of the corresponding background-mixed video depending on the domain to which $\mathbf{V}^i$ belongs. Note that for anchor $\mathbf{z}_f^i$, there are 3 positive pairs: (a) slow version of the mixed video ($\hat{\mathbf{z}}_s^i$), (b) fast version of the mixed video ($\hat{\mathbf{z}}_f^i$), and (c) slow version of the original video ($\mathbf{z}_s^i$). Also, the

loss is computed for all positive pairs in the mini-batch, i.e., $(\mathbf{V}_f^i, \mathbf{V}_s^i)$, $(\mathbf{V}_s^i, \mathbf{V}_f^i)$, $(\hat{\mathbf{V}}_f^i, \hat{\mathbf{V}}_s^i)$, and $(\hat{\mathbf{V}}_s^i, \hat{\mathbf{V}}_f^i)$. Simultaneous minimization of $\mathcal{L}_{bgm}$ in both source and target domains not only learns temporal dynamics and but also helps to better align the features for video domain adaptation by leveraging action semantics shared across both domains.

Incorporating Target Pseudo Labels. While temporal contrastive loss with background mixing helps in aligning the learned representations across the two domains, we cannot fully rely on source categories to learn features discriminative for target domain. Therefore, we propose to use a supervised contrastive loss [31] over pseudo-labeled target samples, an extended version of temporal contrastive loss in Eqn. 1 to enhance discriminabilty by allowing many samples per anchor to be positive, so that videos of the same pseudo-label can be attracted to each other in the embedding space. Let A be the subset of videos assigned pseudo-labels using a confidence threshold, from a mini-batch of B videos, the supervised temporal contrastive loss for incorporating target pseudo-labels ($\mathcal{L}_{tpl}$) is defined as

$$\mathcal{L}_{tpl}(\mathbf{V}_f^i, \mathbf{V}_s^i) = -\frac{1}{|\mathbf{P}(\mathbf{z}_f^i)|} \sum_{\mathbf{p}\in\mathbf{P}(\mathbf{z}_f^i)} \log \frac{h(\mathbf{z}_f^i, \mathbf{p})}{\sum\limits_{\mathbf{p}\in\mathbf{P}(\mathbf{z}_f^i)} h(\mathbf{z}_f^i, \mathbf{p}) + \sum\limits_{\substack{a\in A, a\neq i \\ v\in\{s,f\}}} h(\mathbf{z}_f^i, \mathbf{z}_v^a)} \tag{4}$$

where, $\mathbf{P}(\mathbf{z}_f^i) \equiv \{\mathbf{z}_s^p, \mathbf{z}_f^p : p \in A \;\&\; \tilde{y}^p = \tilde{y}^i\} \setminus \{\mathbf{z}_f^i\}$ is the set of all positives for video $\mathbf{V}_f^i$ and $\tilde{y}^i$ represent the pseudo-label for target video $\mathbf{V}^i$. Note that the set of positives ($\mathbf{P}(.)$) includes all the target domain samples (fast and slow) classified as the same action class as that of the anchor ($\mathbf{z}_f^i$) through the pseudo labels. Following [93], we leverage a temporal ensemble prediction for a given video $\mathbf{V}^i$ from the target domain to produce robust and better-calibrated version of pseudo-labels. Specifically, we obtain the encoded (logits) representations $\mathbf{z}_f^i$ and $\mathbf{z}_s^i$ from the base and auxiliary branch respectively and then compute the pseudo-label as $\tilde{y}^i = \arg\max_k \text{ softmax}(\mathbf{z}_{f\,used}^i)$, where $\mathbf{z}_{f\,used}^i$ represents the mean of both logits. We consider the class index k on which the model is most confident among c classes, provided it is higher than a confidence threshold.

Optimization. Besides the losses $\mathcal{L}_{bgm}$ and $\mathcal{L}_{tpl}$, we minimize the standard supervised cross-entropy loss ($\mathcal{L}_{ce}$) on the labelled source videos as follows.

$$\mathcal{L}_{ce}(\mathbf{V}^{i\{s\}}, y^i) = -\sum_{k=1}^{c} (y^i)_k \log(\mathcal{G}(\mathcal{F}(\mathbf{V}^{i\{s\}})))_k \tag{5}$$

Overall, the loss function for training our model involving both source and target domain data is,

$$\mathcal{L}_{CoMix} = \mathcal{L}_{ce}^{\{s\}} + \lambda_{bgm}(\mathcal{L}_{bgm}^{\{s\}} + \mathcal{L}_{bgm}^{\{t\}}) + \lambda_{tpl}\mathcal{L}_{tpl}^{\{t\}} \tag{6}$$

where λ_{bgm} and λ_{tpl} are weights to balance the impact of individual loss terms. To reduce the number of hyper-parameters, we use the same weight λ_{bgm} for both $\mathcal{L}_{bgm}^{\{s\}}$ and $\mathcal{L}_{bgm}^{\{t\}}$. Notably, for the semi-supervised domain adaptation setting, we also use supervised cross-entropy loss for the few labeled target domain videos in addition to the source domain videos.

4 Experiments

Datasets. We evaluate the performance of our approach using several publicly available benchmark datasets for video domain adaptation, namely UCF-HMDB [7], Jester [54], and Epic-Kitchens [51]. UCF-HMDB (assembled by [7]) is an overlapped subset of the original UCF [69] and HMDB datasets [33], containing $3,209$ videos across 12 action classes. Jester (assembled by [54]) is a large-scale cross-domain dataset that contains videos of humans performing pre-definded hand gestures [46] from two domains, namely Source and Target that contain $51,498$ and $51,415$ video clips respectively across 7 classes. Epic-Kitchens (assembled by [51]) is a challenging egocentric dataset that consists of videos across 8 largest action classes from three domains, namely D1, D2 and D3, corresponding to P08, P01 and P22 kitchens on the full Epic-Kitchens dataset [14]. We use the standard training and testing splits provided by the authors in [7, 54, 51] to conduct our experiments on each datasets. More details about the datasets can be found in the supplementary material.

Baselines. We compare our approach with the following baselines. (1) source only (a lower bound) and supervised target only (an upper bound) baselines that trains the network using supervised source data and supervised target data respectively, (2) popular UDA methods based on adversarial

learning (e.g., DANN [21], and ADDA [74]), (3) existing video domain adaptation methods, including SAVA [10], TA^{3}N [8], ABG [43] and TCoN [54]. We also compare with Source + Target (which simply uses all labelled data available to it to train the network) and ENT [63] in semi-supervised domain adaptation experiments. We directly quote the numbers reported in published papers when possible and use author's provided source codes for TA^{3}N [8] on both Jester and Epic-Kitchens.

Implementation Details. Following [12], we use I3D [4] as the backbone feature encoder network, initialized with Kinetics pre-trained weights. For the temporal graph encoder, we use a 3-layer GCN similar to [82]. We follow the standard 'pre-train then adapt' procedure used in prior works [74, 12] and train the model with only source data to provide a warmstart before the proposed approach is trained. The dimension of the feature extracted from the I3D encoder is 1024 which is the same as the node-feature dimension of the initial layer of the GCN. The final layer of the GCN has its node-feature dimension same as the number of action classes in a dataset and uses a mean aggregation strategy to output the logits. We use a clip-length of 8-frames and train all the models end-to-end using SGD with a momentum of 0.9 and a weight decay of 1e-7. We use an initial learning rate of 0.001 for the I3D and 0.01 for the GCN in all our experiments. We use a batch size of 40 equally split over the two domains, where each batch consists of n clips from the same video, where n is 16 for the fast version (f) and 12, 8, or 4 for the slow version (s). For inference, we use 16 uniformly sampled clips per video and use the base branch of the model to recognize the action. The temperature parameter is set to $\tau = 0.5$. We extract backgrounds from videos using temporal median filtering [58] and empirically set $\gamma = 0.5$ for background mixing. We use a pseudo-label threshold of 0.7 in all our experiments and smooth the cross-entropy loss with $\epsilon = 0.1$, as in [72, 50]. We set λ_{bgm} and λ_{tpl} from $\{0.01, 0.1\}$ depending on the dataset. We report the average action recognition accuracy over 3 random trials. We use 6 NVIDIA Tesla V100 GPUs for training all our models. More implementation details and source codes are included in supplementary material.

Results and Analysis. Table 1 shows results of our method and other competing approaches on UCF-HMDB dataset. Our `CoMix` framework achieves the best average performance of $\mathbf{90.3\%}$, which is about $\mathbf{2.2\%}$ more than the previous state-of-the-art performance on this dataset. While comparing the recent method, SAVA [10] with the same I3D backbone, `CoMix` obtains $\mathbf{4.5\%}$ and $\mathbf{2.7\%}$ improvement on UCF→HMDB and HMDB→UCF task respectively, without relying on frame attention or adversarial learning. These improvements clearly show that our temporal graph contrastive learning with background mixing is not only able to better leverage the temporal information but also shared action semantics, essential for effective video domain adaptation. In summary, `CoMix` outperforms all the existing video DA methods on UCF-HMDB, showing the efficacy of our approach in learning more transferable features for cross-domain action recognition without using any target labels.

Table 1: **Results on UCF-HMDB Dataset.** `CoMix` establishes new state-of-the-art for unsupervised video domain adaptation on UCF-HMDB, by significantly outperforming existing methods.

Method	Backbone	UCF→HMDB	HMDB→UCF	Average
DANN [21]	ResNet-101	75.3	76.4	75.8
JAN [42]	ResNet-101	74.7	79.3	77.0
AdaBN [36]	ResNet-101	75.5	77.4	76.4
MCD [64]	ResNet-101	74.4	79.3	76.8
TA^{3}N [8]	ResNet-101	78.3	81.8	80.1
ABG [43]	ResNet-101	79.1	85.1	82.1
TCoN [54]	ResNet-101	87.2	89.1	88.1
Source Only	I3D	80.3	88.8	84.5
DANN [21]	I3D	80.7	88.0	84.3
ADDA [74]	I3D	79.1	88.4	83.7
TA^{3}N [8]	I3D	81.4	90.5	85.9
SAVA [12]	I3D	82.2	91.2	86.7
`CoMix`	I3D	**86.7**	**93.9**	**90.3**
Supervised Target	I3D	95.0	96.8	95.9

On the large-scale Jester dataset, `CoMix` also outperforms other DA approaches by increasing the Source Only (no adaptation) accuracy from $\mathbf{51.5\%}$ to $\mathbf{64.7\%}$, as shown in Table 2 (left). In particular, our approach achieves an absolute improvement of $\mathbf{9.2\%}$ over TA^{3}N [8], which corroborates that `CoMix` can well handle not only the appearance gap but also the action gap present on this dataset (e.g., for the action class "rolling hand", source domain contains videos of "rolling hand forward", while the target domain only consists of videos of "rolling hand backward"). Table 2 (right) summarizes the results on Epic-Kitchens, which is another challenging dataset consisting of total 6 transfer tasks with a large imbalance across different action classes. Overall, `CoMix` obtains the best on 5 tasks including the best average performance of $\mathbf{43.2\%}$, compared to only $\mathbf{35.3\%}$ and $\mathbf{39.9\%}$ achieved by the source only and TA^{3}N [8] respectively. While the improvements achieved by our approach are encouraging on both Jester and Epic-Kitchens, the accuracy gap between `CoMix` and supervised target is still significant ($\mathbf{30.9\%}$ on Jester and $\mathbf{18.3\%}$ on Epic-Kitchens), which highlights the great potential for improvement in future for unsupervised video domain adaptation.

Table 2: **Results on Jester and Epic-Kitchens Datasets.** CoMix outperforms TA^{3}N [8] by 9.2% on the challenging Jester dataset. On Epic-Kitchens, CoMix achieves the best performance on 5 out of 6 transfer tasks including the best average performance among all compared methods.

Method	Backbone	Jester	Epic-Kitchens						Average
		Source→Target	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	
Source Only	I3D	51.5	35.4	34.6	32.8	35.8	34.1	39.1	35.3
DANN [21]	I3D	55.4	38.3	38.8	37.7	42.1	36.6	41.9	39.2
ADDA [74]	I3D	52.3	36.3	36.1	35.4	41.4	34.9	40.8	37.4
TA^{3}N [8]	I3D	55.5	**40.9**	39.9	34.2	44.2	37.4	42.8	39.9
CoMix	I3D	**64.7**	38.6	**42.3**	**42.9**	**49.2**	**40.9**	**45.2**	**43.2**
Supervised Target	I3D	95.6	57.0	57.0	64.0	64.0	63.7	63.7	61.5

To further study the robustness of our proposed approach, we extend the unsupervised domain adaptation to a semi-supervised setting, where one (1-shot) and three target labels (3-shot) per class are available for training. Table 3 shows that our simple approach consistently outperforms the adversarial DA methods (DANN [21], and ADDA [74]) including the semi-supervised method, ENT [63], on both UCF-HMDB and Jester datasets. Remarkably, CoMix with three target labels per class improves the performance of Source + Target baseline from $\mathbf{93.7\%}$ to $\mathbf{96.6\%}$, which is only $\mathbf{0.2\%}$ lower than the the supervised target upper bound on HMDB→UCF task ($\mathbf{96.6\%}$ vs $\mathbf{96.8\%}$). These results well demonstrate the utility of our proposed approach in many practical applications where annotating *a few* videos per class is typically possible and therefore worth doing given the boost it provides.

Table 3: **Semi-Supervised Domain Adaptation on UCF-HMDB and Jester Datasets.** CoMix significantly outperforms all the compared methods in both one-shot and three-shot settings.

Method	UCF→HMDB		HMDB→UCF		Jester(S) → Jester(T)	
	1-shot	3-shot	1-shot	3-shot	1-shot	3-shot
Source + Target	83.2	85.8	90.3	93.7	53.8	55.0
DANN [21]	85.4	86.9	92.1	93.1	55.1	59.9
ADDA [74]	83.6	86.3	91.2	93.0	59.5	61.3
ENT [63]	85.6	88.6	92.8	95.8	58.6	61.5
CoMix	**88.4**	**93.1**	**95.4**	**96.6**	**65.3**	**69.6**

Ablation Studies. We present following ablation experiments on UCF-HMDB to show effectiveness of different components in our method.

Effectiveness of Individual Components. As seen from Table 4, the vanilla temporal contrastive learning (TCL) achieves an average accuracy of 85.8% (1st row), which is already better than DANN [21], and ADDA [74] (ref. Table 1), showing its effectiveness over adversarial learning in aligning features. While both background mixing (BGM) and incorporation of target pseudo-labels (TPL) individually improves the performance over TCL ($+\mathbf{2.9\%}$ using BGM and $+\mathbf{1.9\%}$ using TPL), addition of both of them leads to the best average performance of 90.3% on UCF-HMDB dataset. This corroborates the fact that both cross-domain action semantics (through BGM) and discriminabilty (through TPL) of the latent space plays a crucial role in video domain adaptation in addition to the vanilla contrastive learning for aligning features.

Table 4: **Ablation Study on UCF-HMDB.** TCL: Temporal Contrastive Learning, BGM: Background Mixing, TPL: Target Pseudo-Labels.

TCL	BGM	TPL	U→H	H→U	Average
✓	✗	✗	83.3	88.4	85.8
✓	✓	✗	86.2	91.2	88.7
✓	✗	✓	83.5	91.9	87.7
✓	✓	✓	**86.7**	**93.9**	**90.3**

Table 5: **Comparison with MixUp Strategies**. Background mixing outperforms other alternatives in leveraging shared action semantics on UCF-HMDB.

Method	U→H	H→U	Average
Gaussian Noise	84.7	90.6	87.6
Video MixUp	85.1	91.7	88.4
Video CutMix	84.6	92.1	88.3
Background Mixing	**86.7**	**93.9**	**90.3**

Comparison with Different MixUp Strategies. We explore the effectiveness of background mixing by comparing with different MixUp strategies (Table 5): (a) Gaussian Noise: adding White Gaussian Noise to videos in both domains; (b) Video MixUp [89]: directly mixing one video with another from a different domain, as in images; (c) Video CutMix [88]: randomly replacing a region of a video with another region from the other domain. The proposed way of generating synthetic videos by mixing background of a video from one domain to a video from another domain, outperforms all three alternatives on UCF-HMDB. Note that while both MixUp and CutMix destroy motion pattern of original video, background mixing keeps semantic consistency without changing temporal dynamics.

Effect of Background Pseudo-labels. We investigate the effect of pseudo-labels on background mixed videos (i.e., both videos considered to be of same action class while creating positives) by simply adding them as unlabeled videos without any modification to the contrastive objective in Eq. 1. CoMix without background pseudo-labels decreases the performance from 90.3% to 89.0% ($-\mathbf{1.3\%}$: Table 6), showing its effectiveness in leveraging action semantics shared across both domains.

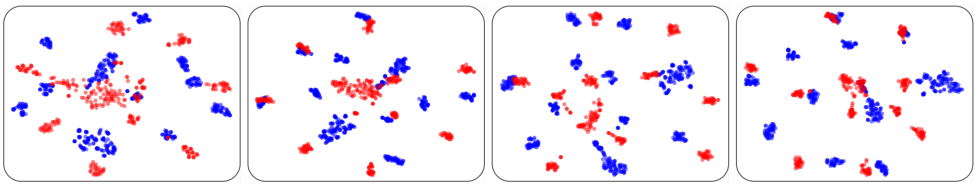

Figure 5: **Feature Visualizations using t-SNE.** Plots show visualization of our approach with different components on UCF→HMDB task. Blue and red dots represent source and target data respectively. Features for both target and source domain become progressively discriminative and improve from left to right by adoption of our novel components within a temporal contrastive learning framework. Best viewed in color.

Effect of Source Contrastive Learning. `CoMix` adopts contrastive learning on both source and target domains, although we already have supervised cross-entropy loss on source videos. We observe that applying contrastive learning on target domain only, by removing source contrastive objective $\mathcal{L}_{bgm}^{\{s\}}$ from Eq. 6, lowers down the performance from 90.3% to 88.4% ($\mathbf{-1.9\%}$) on UCF-HMDB (Table 6). This shows the importance of training the model using the same temporal invariance objective on both domains simultaneously to achieve effective alignment across domains.

Effect of Random Speed Invariance. We remove randomness in video speed from the auxiliary branch of our temporal contrastive learning framework and observe that `CoMix` (with 16 clips in the base branch and only 8 clips in the auxiliary branch) leads to an average top-1 accuracy of 89.6% compared to 90.3% ($\mathbf{-0.7\%}$: Table 6), showing the importance of random speed invariance in learning robust features.

Table 6: **Ablation Study on Contrastive Learning.**

Method	U→H	H→U	Average
`CoMix`	**86.7**	**93.7**	**90.3**
– w/o Background Pseudo-labels	85.8	92.2	89.0
– w/o Source Contrastive Learning	85.1	91.8	88.4
– w/o Random Speed Invariance	86.4	92.8	89.6

Self-Training vs Supervised Contrastive Learning. We directly use self-training that uses cross-entropy loss on target pseudo labels instead of $\mathcal{L}_{tpl}^{\{t\}}$ and find that the average performance drops to 88.7% on UCF-HMDB, indicating the advantage of supervised contrastive objective in enhancing discriminability of the latent space by successfully leveraging label information from target domain.

Effect of Graph Representation. We examine the effect of graph representation for videos and find that by removing GCN from our framework lowers down the performance from 90.3% to 88.1% on UCF-HMDB dataset, which shows that graph contrastive learning is more useful in capturing the temporal dependencies, essential for video domain adaptation.

Feature Visualizations. We use t-SNE [44] to visualize the features learned using different components of our `CoMix` framework. As seen from Figure 5, alignment of domains including discriminability improves as we adopt "TCL" and "BGM" to the vanilla Source only model. The best results are obtained when all the three components "TCL", "BGM" and "TPL" i.e., `CoMix` are added and trained using an unified framework (Eq. 6) for unsupervised video domain adaptation.

5 Conclusions

In this paper, we introduce a new end-to-end temporal contrastive learning framework to bridge the domain gap by learning consistent features representing two different speeds of the unlabeled videos. We also propose two novel extension to temporal contrastive loss by using background mixing and target pseudo-labels, that allows additional positive(s) per anchor, thus adapting contrastive learning to leverage cross-domain action semantics and label information from the target domain respectively in an unified framework, for learning discriminative invariant features. We demonstrate the effectiveness of our approach on three standard datasets, outperforming several competing methods.

Broader Impact. Our research can help reduce burden of collecting large-scale supervised data in many real-world applications of human action recognition by transferring knowledge from auxiliary datasets. The positive impact that our work could have on society is in making technology more accessible for institutions and individuals that do not have rich resources for collecting and annotating large-scale video datasets. Negative impacts of our research are difficult to predict, however, it shares many of the pitfalls associated with standard deep learning models such as susceptibility to adversarial attacks and lack of interpretablity. Other adverse effects could be potential attrition in jobs in certain sectors of economy where fewer employees (security guards, nurses, etc.) are needed to monitor human activities as a result of wider adoption of automated video recognition systems.

References

[1] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020.

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019.

[3] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2229–2238, 2019.

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

[5] Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3296–3303, 2019.

[6] Chun-Fu Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. *arXiv preprint arXiv:2010.11757*, 2020.

[7] Jin Chen, Xinxiao Wu, Lixin Duan, and Shenghua Gao. Domain adversarial reinforcement learning for partial domain adaptation. *arXiv preprint arXiv:1905.04094*, 2019.

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

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607, 2020.

[10] Zhihong Chen, Chao Chen, Zhaowei Cheng, Boyuan Jiang, Ke Fang, and Xinyu Jin. Selective transfer with reinforced transfer network for partial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12706–12714, 2020.

[11] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *arXiv preprint arXiv:1912.05534*, 2019.

[12] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695, 2020.

[13] Gabriela Csurka. Domain adaptation for visual applications: A comprehensive survey. *arXiv preprint arXiv:1702.05374*, 2017.

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

[15] Hal Daumé III, Abhishek Kumar, and Avishek Saha. Frustratingly easy semi-supervised domain adaptation. In *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, pages 53–59, 2010.

[16] Aysegul Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation. *arXiv preprint arXiv:1807.09384*, 2018.

[17] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. *arXiv preprint arXiv:2104.14548*, 2021.

[18] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019.

[20] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. *arXiv preprint arXiv:2104.14558*, 2021.

[21] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[22] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[23] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3154–3160, 2017.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[25] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[26] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[27] Lanqing Hu, Meina Kan, Shiguang Shan, and Xilin Chen. Duplex generative adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1498–1507, 2018.

[28] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, volume 2, page 4, 2018.

[29] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. *arXiv preprint arXiv:2007.10730*, 2020.

[30] Andrej 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.

[31] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[32] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018.

[33] 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.

[34] Abhishek Kumar, Avishek Saha, and Hal Daume. Co-regularization based semi-supervised domain adaptation. *Advances in neural information processing systems*, 23:478–486, 2010.

[35] Kibok Lee, Yian Zhu, Kihyuk Sohn, Chun-Liang Li, Jinwoo Shin, and Honglak Lee. i-mix: A domain-agnostic strategy for contrastive representation learning. 2021.

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

[37] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018.

[38] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. 2019 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 7082–7092, 2019.

[39] Weizhe Liu, David Ferstl, Samuel Schulter, Lukas Zebedin, Pascal Fua, and Christian Leistner. Domain adaptation for semantic segmentation via patch-wise contrastive learning. *arXiv preprint arXiv:2104.11056*, 2021.

[40] 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, 2015.

[41] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional Adversarial Domain Adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

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

[43] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial bipartite graph learning for video domain adaptation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 19–27, 2020.

[44] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[45] Xudong Mao, Yun Ma, Zhenguo Yang, Yangbin Chen, and Qing Li. Virtual mixup training for unsupervised domain adaptation. *arXiv preprint arXiv:1905.04215*, 2019.

[46] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[47] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. *arXiv preprint arXiv:2008.12197*, 2020.

[48] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *European Conference on Computer Vision*, pages 86–104, 2020.

[49] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[50] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703, 2019.

[51] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020.

[52] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[53] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[54] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11815–11822, 2020.

[55] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. *arXiv preprint arXiv:2103.05905*, 2021.

[56] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. *arXiv preprint arXiv:1809.02176*, 2018.

[57] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[58] Massimo Piccardi. Background subtraction techniques: a review. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 4, pages 3099–3104. IEEE, 2004.

[59] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *arXiv preprint arXiv:2008.03800*, 2020.

[60] Adrià Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Patraucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. *arXiv preprint arXiv:2103.16559*, 2021.

[61] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[62] Aadarsh Sahoo, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Select, label, and mix: Learning discriminative invariant feature representations for partial domain adaptation. *arXiv preprint arXiv:2012.03358*, 2020.

[63] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8050–8058, 2019.

[64] 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.

[65] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. *arXiv preprint arXiv:1707.01217*, 2017.

[66] Zhiqiang Shen, Zechun Liu, Zhuang Liu, Marios Savvides, Trevor Darrell, and Eric Xing. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. *arXiv preprint arXiv:2003.05438*, 3(7), 2020.

[67] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[68] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. *arXiv preprint arXiv:2102.02751*, 2021.

[69] 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.

[70] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016.

[71] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

[72] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[73] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[74] 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.

[75] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017.

[76] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

[77] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*, pages 504–521, 2020.

[78] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Rongrong Ji, and Xing Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. *arXiv preprint arXiv:2009.05769*, 2020.

[79] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36, 2016.

[80] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[81] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018.

[82] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[83] Yuan Wu, Diana Inkpen, and Ahmed El-Roby. Dual mixup regularized learning for adversarial domain adaptation. In *European Conference on Computer Vision*, pages 540–555. Springer, 2020.

[84] Yaochen Xie, Zhao Xu, Zhengyang Wang, and Shuiwang Ji. Self-supervised learning of graph neural networks: A unified review. *arXiv preprint arXiv:2102.10757*, 2021.

[85] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. *arXiv preprint arXiv:1912.01805*, 2019.

[86] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6548–6557, 2020.

[87] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems*, 33, 2020.

[88] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019.

[89] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[90] Yabin Zhang, Bin Deng, Kui Jia, and Lei Zhang. Label propagation with augmented anchors: A simple semi-supervised learning baseline for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 781–797. Springer, 2020.

[91] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6810–6818, 2018.

[92] Yuliang Zou, Jinwoo Choi, Qitong Wang, and Jia-Bin Huang. Learning representational invariances for data-efficient action recognition. *arXiv preprint arXiv:2103.16565*, 2021.

[93] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Broader impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code is included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Code is included in the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]