DISCOVERING INFORMATIVE AND ROBUST POSITIVES FOR VIDEO DOMAIN ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised domain adaptation for video recognition is challenging where the domain shift includes both spatial variations and temporal dynamics. Previous works have focused on exploring contrastive learning for cross-domain alignment. However, limited variations in intra-domain positives, false cross-domain positives, and false negatives hinder contrastive learning from fulfilling intra-domain discrimination and cross-domain closeness. This paper presents a non-contrastive learning framework without relying on negative samples for unsupervised video domain adaptation. To address the limited variations in intra-domain positives, we set unlabeled target videos as anchors and explored to mine "informative intra-domain positives" in the form of spatial/temporal augmentations and target nearest neighbors (NNs). To tackle the false cross-domain positives led by noisy pseudo-labels, we reversely set source videos as anchors and sample the synthesized target videos as "robust cross-domain positives" from an estimated target distribution, which are naturally more robust to the pseudo-label noise. Extensive experiments on several cross-domain action recognition benchmarks demonstrate the superiority of our approach over state-of-the-art methods.

1 INTRODUCTION

Recent advances in deep neural networks have revolutionized many computer vision tasks, such as image and video recognition He et al. (2016), semantic segmentation Chen et al. (2017), etc. However, these impressive results heavily rely on human annotations, which are expensive and time-consuming to obtain. To address this issue, semi-supervised learning Miyato et al. (2018) and self-supervised learning (SSL) He et al. (2020) have been studied to utilize the knowledge from a label-rich dataset to a label-scarce dataset. However, the domain shift problem between the source and target datasets usually exists in real-world scenarios, leading to performance degradation. Unsupervised domain adaption (UDA) has been exploited to transfer knowledge across datasets with domain discrepancies to mitigate this problem. While most of the methods have been developed only for images, the problem of UDA in videos remains largely underexplored.

Recently, a few works have attempted UDA for video action recognition by directly matching frame/clip-level features Chen et al. (2019a); Pan et al. (2020a); Choi et al. (2020). However, these methods usually extend the image-based UDA methods without considering long-term temporal information or action semantics. Song and Kim et al. Song et al. (2021); Kim et al. (2021) alleviate these issues with contrastive learning to learn such long-term spatial-temporal representations by instance discrimination. To further understand how contrastive learning helps UDA, we firstly recall that domain-wise discrimination and class-wise closeness are the two main criteria to solve UDA problems Shi & Sha (2012); Tang et al. (2020). Considering an unlabeled video from the target domain as an anchor, we explain that *Song et al. (2021) introduces intra-domain positives (in the form of cross-modal correspondence, e.g, optical flow) to help UDA by learning discriminative representation in the target domain.* On the other hand, *Kim et al. (2021) utilizes cross(source)-domain positives with the help of pseudo-labels and benefits the class-wise closeness by pushing the samples of the same class but different domains closer.*

By thoroughly studying the effect of the positives and negatives selection from the existing contrastive-based methods in Figure 1 (a), we empirically find that **limited variations** in intra-domain positives, **pseudo-label noise** in cross-domain positives and **false negatives** are the three issues that largely hinder the performance. Explicitly, when selecting the negatives based on the ground-truth, intra-

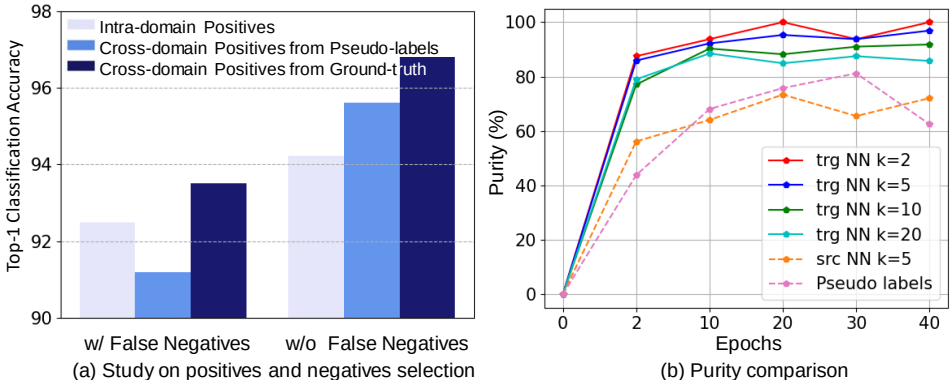


Figure 1: On HMDB→UCF: (a) **Effect of false negatives and intra/cross-domain positives.** We study two contrastive methods in video DA: STCDA Song et al. (2021) with intra-domain positives and LCMC Kim et al. (2021) with cross-domain positives. We observe that *limited variations in intra-domain positives*, *pseudo-label noise in cross-domain positives* and *false negatives* are the three issues that largely hinder the performance. (b) **Purity comparison among nearest neighbors (NNs) and pseudo label.** We found the NNs in target domain are clean and fit as informative and robust intra-domain positives. Note that purity is the cleanness of pseudo-supervision compared to the ground truth.

domain positives provide the least performance gains, indicating their limitation in variations. Further, selecting the cross-domain positives based on pseudo-labels leads to 2% drop compared to the ground-truth. Importantly, when selecting the negatives from either different instance Song et al. (2021) or pseudo-labels Kim et al. (2021), the performance of the three baselines in Figure 1 (a) drops by $2 \sim 4\%$. Besides, the purity analysis in Figure 1 (b) also indicates that the pseudo-labels of target video are noisy and thus not reliable for selecting cross-domain positives and negatives. Based on the observations, several straightforward questions might be raised:

1. Are there any unexplored intra-domain positives to enrich intra-domain variations?
2. How to alleviate the pseudo-label noise issue when selecting cross-domain positives?
3. How to address the adverse effect from false negatives?

To answer the first question, we propose to leverage the spatial and temporal augmentations of the unlabeled target video as intra-domain positives. The rationale is that partially modifying the spatial/temporal information of videos could alter the samples without changing the whole action semantics. Incorporating those samples as positives could help the model be invariant to spatial/temporal shift. Take a step further, we also explore the anchor's nearest neighbors (NNs) in target feature space, which capture rich target-domain variations. By analyzing the purity of the NNs in both source and target domain in Figure 1 (b), we observe that *target-domain NNs are clean and thus fit as intra-domain positives*.

To address the second question, we are motivated from Xie et al. (2018) that *though the pseudo-labels of target samples are noisy, the class-conditional centroids μ_c^t of target features would weaken this noise by averaging all the target features of same pseudo class*. To incorporate it into our contrastive learning framework, we re-formulate the optimization reversely by setting source videos as anchors and then mining target videos as cross-domain positives. We then estimate the gaussian target distribution $\mathcal{N}(\mu_c^t, \Sigma_c^t)$ of target features based on pseudo-labels. Given a source video as anchor, synthesized target features can be drawn from this distribution that shares the same class as the source anchor. Consequently, we could leverage those synthesized features as cross-domain positives as the estimated target distribution is robust to pseudo-label noise.

To tackle the last question, we present a simple yet effective non-contrastive learning framework without relying on negative samples for video domain adaptation. Specifically, we are motivated by the recent advances in SSL Grill et al. (2020); Chen & He (2021) to discard the negatives and maximize the similarity between the anchor and positives in feature space with the help of an extra MLP head and a stop-gradient operation. Beyond adopting this non-contrastive optimization into our setting, we highlight our compelling finding that *the MLP head is crucial for intra-domain positives*

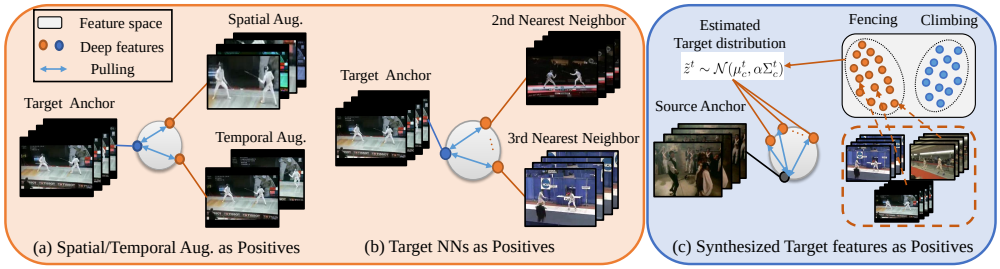


Figure 2: **Overview of our intra/cross-domain positives.** (a) spatial/temporal augmentations as intra-domain positives. (b) target nearest neighbors as robust intra-domain positives. Note that we consider self-augmentation as the 1st NN. (c) synthesizing target features from an estimated target distribution as cross-domain positives. In short, intra-domain positives help target-domain discrimination while cross-domain positives support class closeness across domains.

optimization but largely hampers the convergence for cross-domain positives. Consequently, we remove the MLP head for the latter case. To summarize our main contributions:

- By studying the effects of the positives and the negatives from previous contrastive-based video DA methods, we make the first attempt to emphasize that limited variations in intra-domain positives, pseudo-label noise in inter-domain positives and false negatives are the three under-explored bottlenecks
- We contribute a unified solution to address the above issues by introducing more informative and robust intra-domain and cross-domain positives without replying on negative samples for video DA problem.
- We conduct extensive experiments and ablation studies on challenging cross-domain action recognition benchmarks (UCF-HMDB and Epic-Kitchens) and thoroughly verify our effectiveness over state-of-the-art methods.

2 Related Work

2.1 Action Recognition

Action recognition is a challenging problem with great application potential. Many previous approaches are deployed in both 2D Karpathy et al. (2014a), and 3D CNN-based framework (such as I3D Carreira & Zisserman (2017a), ResNet3D Hara et al. (2017)), which achieves significant progress on RGB modality. To incorporate multiple modalities or temporal dynamics, two-stream architecture Simonyan & Zisserman (2014) is widely used to process RGB and optical-flow in two separate CNNs with a late fusion in the upper layers Karpathy et al. (2014b). SlowFast network Feichtenhofer et al. (2019) employs two pathways for recognizing actions by processing a video at different frame rates. Despite promising results on supervised benchmarks, these models are critically dependent on large datasets that require careful human annotation effort. In contrast, our focus is on unsupervised domain adaptation settings for action recognition where target domain data is unlabeled.

2.2 Self-supervised Learning

Self-supervised representation has witnessed tremendous progress on images and video by using different pretext tasks (i.e., rotations Gidaris et al. (2018), jigsaw puzzle Noroozi & Favaro (2016)) and contrastive learning Chen et al. (2020); He et al. (2020); Han et al. (2020). Recently, non-contrastive learning methods (i.e., BYOL Grill et al. (2020) and SimSiam Chen & He (2021)) dominate the SSL performance by learning meaningful representation by maximizing the similarity between two augmented positive samples without considering negative pairs. While our method is motivated by the non-contrastive learning above, we highlight our difference in mining "informative" and "robust" intra-domain and cross-domain positives beyond instance-level variations for better target-domain discrimination and class-wise closeness. Importantly, we observe that the predictor in SimSiam Chen & He (2021), which is designed for preventing model collapse, would lead to

convergence issue for cross-domain positives. In contrast, we remove the predictor for cross-domain positives in optimization while maintain it for intra-domain positives. We will explain the details in the experiment section.

2.3 UNSUPERVISED DOMAIN ADAPTATION

There are several directions to make UDA effective. Specifically, mainstream methods seek for learning domain-invariant feature representations by minimizing some measurement of distribution discrepancy Tzeng et al. (2014); Chung et al. (2022); Qiu et al. (2021), utilizing adversarial learning Ganin et al. (2016); Long et al. (2018); Saito et al. (2018); Pan et al. (2020b) or cross-domain augmentation Li et al. (2021). Self-training Xie et al. (2018); Zou et al. (2018; 2019) also thrives on UDA by training the model with labeled source data and pseudo-labeled target data. However, less attention has been paid to video domain adaptation until recently. Existing video UDA methods can be summarized into two folds: temporal attention-based adversarial learning Chen et al. (2019b); Luo et al. (2020); Pan et al. (2020a) and SSL-based cross-domain alignmentMunro & Damen (2020); Choi et al. (2020); Song et al. (2021); Kim et al. (2021). Explicitly, self-supervisions such as predicting video clips order Choi et al. (2020), multi-modality correspondence Munro & Damen (2020) and contrastive learning Song et al. (2021); Kim et al. (2021); Sahoo et al. (2021), have been exploited. Compared to the image-based DA methods, our method provides a new perspective for domain discrimination and class-wise closeness via exploring positive samples in contrastive learning. Compared to the contrastive-based video DA methods, (1) we provide empirical insights on the limitations of existing contrastive methods by analyzing the effect of positives ans negatives; (2) we address the issues by proposing more informative and robust intra-domain positives and cross-domain positives without replaying on negative samples.

3 PROPOSED METHOD

In unsupervised video domain adaptation problem, we are given a source domain $\mathbb{D}_s = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)|_{i=1}^{N_s}\}$ of N_s labeled source videos and a target domain $\mathbb{D}_t = \{(\boldsymbol{x}_i^t)|_{i=1}^{N_t}\}$ of N_t unlabeled target videos with a shared label space $\mathcal{L}$. The joint distributions of source and target domain are not identically and independently distributed, specifically defined as $P(\boldsymbol{x}^s, \boldsymbol{y}^s) \neq Q(\boldsymbol{x}^t, \boldsymbol{y}^t)$. Given these datasets, the goal is to train a deep neural network for action recognition on the labeled source videos $\mathbb{D}_s$ and the unlabeled target videos $\mathbb{D}_t$ such that the model can generalize well on unseen target domain videos. To specify, the model is comprised of a feature extractor F followed by a classifier G.

3.1 INTRA-DOMAIN POSITIVES

The goal of intra-domain positives is to ensure domain-wise discrimination, which means that data in either source or target domains are discriminatively clustered.

Intra-source Discrimination. As source domain videos have labels, we simply minimize the standard supervised cross-entropy loss ($\mathcal{L}_{ce}$) to achieve source-domain discrimination:

$$\mathcal{L}_{ce}(x_i^s, y_i^s) = -\frac{1}{N_s}\sum_{i=1}^{N_s} \log p_{i,y_i^s}, \tag{1}$$

where $p_{i,y_i^s} = G(F(x_i^s))_{y_i^s}$ is the y_i^s-th element of a K-dimensional prediction and K indicates the number of class.

Self-supervision as Intra-domain positives. Without proper regularization on unlabeled target data, target-domain discrimination cannot be realized for free. Motivated by the self-supervised learning methods Chen & He (2021); Bai et al. (2021), we firstly propose to learn discriminative representation in target domain by maximizing the similarity between different augmented views of unlabeled target data. Specifically, we consider 1) spatial augmentation, including standard image-based augmentation such as random flip, crop and color distortion. 2) temporal augmentation, including video play speed Benaim et al. (2020), and shuffled clips Fernando et al. (2017). Precisely, we randomly double the sampling rate ($\times 2$) for video play speed or randomly shuffle a clip ($1/3$ consecutive frames) of a video.

The rationale behind is that *the effectiveness of the intra-domain positives are highly coupled with how "informative" and "clean" the positives are given an anchor video.* "Informative" positives usually capture rich spatial variations (e.g., background/viewpoint changes) or temporal dynamics (e.g., video speed/action length). Our choices of spatial and temporal augmentations above could partially change the anchor video along these variations without changing the whole action semantics. Consequently, incorporating those samples as intra-domain positives could help the model be invariant to the variations in spatial and temporal axis.

Explicitly, our model takes two randomly augmented views x_1^t and x_2^t from a target video x^t as input: one view is strongly augmented by one of the augmentations above while the other view is weakly augmented. A feature extractor F processes the two views into feature space and a two-layers prediction MLP head, denoted as H, further transforms the feature of one view and matches it to the other view. As shown in Figure 2 (a), our intra-target domain alignment loss can be formulated as maximizing the cosine similarity between the two views:

$$\mathcal{L}_{intra}(x_1^t, x_2^t) = \mathcal{D}(h_1^t, z_2^t), \tag{2}$$

$$\mathcal{D}(h_1^t, z_2^t) = -\frac{h_1^t}{\|h_1^t\|_2} \frac{z_2^t}{\|z_2^t\|_2}, \tag{3}$$

where $h_1^t = H(F(x_1^t))$ is the predictor output, $z_2^t = F(x_2^t)$ is the spatial-temporal feature, $\|\cdot\|$ is l_2-norm and $\mathcal{D}(\cdot,\cdot)$ is a cosine distance function. Specifically, a stop-gradient operation is applied on the z_2^t and we implement it by modifying Eqn. 3 as $\mathcal{D}(h_1^t, stopgrad(z_2^t))$.

Explanation on optimization without negatives. Unlike traditional contrastive learning objective Song et al. (2021); Kim et al. (2021), we only involve positive pairs (e.g, (x_1^t, x_2^t)) into the optimization without negative samples. The insight Chen & He (2021) behind is that the negative samples in contrastive-learning play a role in preventing models from collapsing solutions (e.g, constant outputs). Alternatively, it is found that applying the predictor $H(\cdot)$ on one view x_1^t and the stop-gradient operation $stopgrad(\cdot)$ on the other view x_2^t could also avoid model collapse.

As the existing contrastive video DA methods are largely constrained by the false negatives issue (Figure 1), we leverage this optimization to cut off the reliance on the negatives and thus avoid the error accumulation led by false negatives.

Target nearest neighbors as Intra-domain positives. The empirical finding in Figure 1(b) shows that intra-domain NNs (target NNs) have higher purity than cross-domain NNs (source NNs) and pseudo-labels. It motivates us to further explore the intra-domain NNs as intra-domain positives which capture rich spatial and temporal variations in target domain.

Following the notation of Eqn. 3, we add the detached feature z_2^t of the weakly-augmented view and its class pseudo label $\hat{y^t} = \arg\max G(z_2^t)$ to a memory bank, denoted as $\mathbb{M}_t = \{(z_i^t, \hat{y_i}^t)|_{i=1}^{N_t}\}$ in an online fashion where the size equals to the number of target dataset N_t. $\mathbb{M}_t$ is updated with the newest features and pseudo labels given a current batch. Then we find the k nearest neighbors of z_2^t in the memory bank to obtain k target features $\{\hat{z_j}^t|_{j=1}^k\}$. Finally, we extend the Eqn. 3 to incorporate target NNs as intra-domain positives:

$$\mathcal{L}_{t-intra}(x^t) = \frac{1}{k}\sum_{j=1}^{k} \mathcal{D}(h_1^t, \hat{z_j}^t). \tag{4}$$

The final loss is the summation of the above loss for all target videos $\mathbb{D}_t$. By pulling unlabeled target videos close to their NNs, the model could achieve the better target-domain discrimination. Importantly, the Eqn. 4 is equivalent to the Eqn. 3 when $k = 1$. As k increases, the intra-domain positives could cover richer target-domain variations while under the risk of bringing false positives. The trade-off between the "informativeness" and "robustness" of intra-domain positives is discussed in experiment section in Figure 4.

3.2 Robust Cross-domain Positives

While the intra-domain positive focuses on domain-wise discrimination, it ignores the cross-domain alignment in the shared action semantic space. A natural solution is to extend Eqn. 4 by incorporating

source domain positives which cover cross-domain spatial-temporal variations. As target videos do not have labels, we resort to use the pseudo-labels for choosing source samples or directly using source NNs as positives. However, as Figure 1 (b) suggests, both source NNs and pseudo labels have low purity. Incorporating cross-domain positives based on such noisy supervisions would result in false domain alignment.

To tackle this issue, we formulate it reversely to utilize source videos as anchors and mine for more informative target domain positives. Previous work Xie et al. (2018) motivates us that, *though the pseudo labels of target samples are noisy, the class-conditional centroids μ_c^t of target features would weaken this noise by averaging all the target features of same pseudo class*. Consequently, the target centroid of each class μ_c^t is more robust to pseudo-label noise. To further explore target domain variations, we also estimate the corresponding covariance matrix Σ_c^t from the target features of the same pseudo-class.

Instead of directly mining real target videos as positives based on noisy NNs or PL, we randomly sample a synthesized target feature from a normal distribution $\mathcal{N}(\mu_c^t, \Sigma_c^t)$ of the same action class c to the source anchor video in the form of

$$\tilde{z}^t \sim \mathcal{N}(\mu_c^t, \alpha\Sigma_c^t), \tag{5}$$

where α is a positive coefficient to control the strength of target data variations. In the implementation, the target class centroids and the covariance matrix are computed based on the target features of the same pseudo class from the memory bank $\mathbb{M}_t$ in Eqn. 4. As the estimation in the first few epochs is not so accurate, we set $\alpha = (t/T) \times \alpha_0$ where t and T are the current epoch and maximum epochs. As a result, the negative impact from the estimated covariance in the early training stage will be reduced.

Similarly to Eqn. 4, our model takes a source video x^s as input (anchor) and processes it with the feature extractor F. Then we match it to the synthesized positives sampled from the class-conditional target distribution of class y^s from Eqn. 5 in feature space. As shown in Figure 2 (c), when the number of sampled positives is M, our cross-domain alignment loss can be formulated as:

$$\mathcal{L}_{s-cross}(x^s) = \frac{1}{M}\sum_{i=1}^{M}\mathcal{D}(z^s, \tilde{z_i}^t), \tag{6}$$

where $z^s = F(x^s)$ is the feature of the source anchor and $\tilde{z_i}^t$ is the target positive sampling from $\mathcal{N}(\mu_{y_s}^t, \alpha\Sigma_{y_s}^t)$. We further derive a close-form upper-bound for the expectation of Eqn. 6 under all possible target features as $\mathcal{L}_{s-cross}^{\infty}$ in appendix B.1.

Explanation on cross-domain optimization. Unlike Eqn. 4, we do not transform the z^s by the predictor H which is originally designed for avoiding model collapse. Interestingly, we find that *discarding the predictor H helps better convergence for cross-domain positives in optimization.* Our insight is that, as source feature z^s and the synthesized target feature $\tilde{z}^t$ are from different distribution, pulling them close requires non-trivial optimization efforts and thus models are less prone to collapse. In contrast, matching different views of a target video z^t from the same distribution is relatively "easy". Without regularization from the predictor H, models will easily match all the views (e.g, augmentations or intra-domain NNs) of the target video into a constant output.

Discussion on Xie et al. (2018). It is worth noting some fundamental differences between our algorithm and Xie et al. (2018). From the **optimization view**, they propose centroids alignment across domains where target centroids are updated with the current target batch based on pseudo-labels and thus will receive gradients. If the pseudo labels are noisy, wrong gradients will hurt the training. In comparison, we align the cross-domain samples within constrastive framework where our synthesized target features are detached from gradients; thus, only the source batch will receive the robust gradients. **Algorithmically**, besides the estimation of target centroids, we further integrate target covariance to estimate the whole gaussian distribution of target domain. Table 1 shows that we have a significant improvement over Xie et al. (2018).

3.3 Overall Optimization

In summary, the overall optimization includes the intra-domain positives in section 3.1 and the cross-domain positives in section 3.2 as follows:

Table 2: **Results on Epic-Kitchens Datasets.** All the methods use I3D two-stream backbone including both RGB and flow modality.

Method	Backbone	D2 → D1	D3 → D1	D1 → D2	D3 → D2	D1 → D3	D2 → D3	Mean
Source-only	I3D	42.5	44.3	42.0	56.3	41.2	46.5	45.5
AdaBN Li et al. (2018)	I3D	44.6	47.8	47.0	54.7	40.3	48.8	47.2
MMD Long et al. (2015)	I3D	43.1	48.3	46.6	55.2	39.2	48.5	46.8
MCD Saito et al. (2018)	I3D	42.1	47.9	46.5	52.7	43.5	51.0	47.3
MM-SADA Munro & Damen (2020)	I3D	48.2	50.9	49.5	56.1	44.1	52.7	50.3
LCMC Kim et al. (2021)	I3D	49.5	51.5	50.3	56.3	46.3	52.0	51.0
STCDA Song et al. (2021)	I3D	49.0	52.6	52.0	55.6	45.5	52.5	51.2
Ours (RGB)	I3D	44.8	41.8	48.1	46.8	41.9	48.2	45.2
Ours (RGB + Flow)	I3D	**49.9**	**53.5**	**52.7**	**57.5**	**46.9**	**53.4**	**52.3**
Supervised-target	I3D	62.8	62.8	71.7	71.7	74.0	74.0	69.5

$$\mathcal{L}_{all}^{rgb} = \mathcal{L}_{ce}^{s} + \mathcal{L}_{s-cross}^{\infty} + \lambda \mathcal{L}_{t-intra}, \tag{7}$$

where λ is a hyperparameter to control the strength of intra-domain alignment. We empirically set the λ to 15 for the convergence of $\mathcal{L}_{t-intra}$ due to the predictor. Our framework can be easily extended to two-stream network by adding Eqn. 8 for optical flow modality as $\mathcal{L}_{all}^{flow}$. Also, the optical flow correspondence of target videos can be integrated as intra-domain positives in $\mathcal{L}_{t-intra}$ (Eqn.15). Our algorithm with pseudo code will be found in appendix Alg.1.

4 Experiment

4.1 Datasets

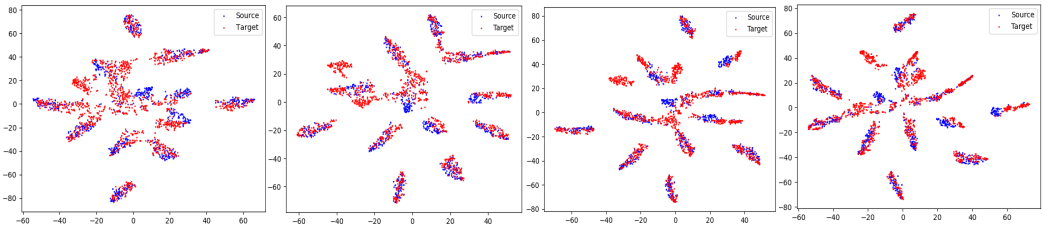
We evaluate the performance of our approach on two publicly benchmark datasets for video domain adaptation, namely UCF↔ HMDB Chen et al. (2019a) and Epic-Kitchens Munro & Damen (2020). We use the standard training and testing splits provided by the authors in Chen et al. (2019a); Munro & Damen (2020) to conduct our experiments on each dataset.

UCF↔ HMDB is first assembled by Chen et al. Chen et al. (2019a)) for studying video domain adaptation problem. It is a subset of the original UCF Soomro et al. (2012) and HMDB datasets Kuehne et al. (2011), containing $3,209$ videos across 12 overlapping classes for human activity recognition. We show the performance of our method on two setting: UCF→ HMDB and UCF← HMDB. **Epic-Kitchens** is a challenging egocentric fined-grained action recognition 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 Damen et al. (2018). It is released by the authors in Munro & Damen (2020) for exploring video domain adaptation on the fined-grained setting.

4.2 Comparison with State-of-the-art Methods

Baselines. We compare our approach with the following baselines. (1) source only and supervised target only baselines that trains the network using labeled source data and labeled target data respectively, (2) existing UDA methods based on adversarial learning, including DANN Ganin et al. (2016)and ADDA Tzeng et al. (2017), (3) existing UDA based on pseudo-labeling, including CRST Zou et al. (2019) and MSTN Xie et al. (2018). (4) existing video domain adaptation methods, including TA^{3}N Chen et al. (2019a), ABG Luo et al. (2020), TCoN Pan et al. (2020a), MM-SADA Munro & Damen (2020), SAVA Choi et al. (2020), STCDA Song et al. (2021) and LCMC Kim et al. (2021).

Results on UCF-HMDB. Table 1 shows results of our method with other competing approaches on UCF-HMDB dataset. Our framework achieves the best performance than previous state-of-the-art on different backbones. Compared to the recent methods incorporating self-supervision such as clip-order prediction (SAVA Choi et al. (2020)), appearance-motion correspondence (MM-SADA Munro & Damen (2020)) and contrastive learning (STCDA Song et al. (2021) and LCMC Kim et al. (2021)), our method makes a significant improvement using I3D backbone with $\mathbf{1.8\%}$ on UCF→HMDB and $\mathbf{3.5\%}$ on HMDB→UCF task respectively. Though not specifically designed for motion space, our method can extend to two-stream backbone with additional gains to $\mathbf{86.1\%}$ on UCF→HMDB and $\mathbf{95.4\%}$ on HMDB→UCF, achieving the state-of-the-art performance.



(a) Source only (b) Intra-domain (c) Cross-domain (d) Ours (all)

Figure 3: Feature Visualizations using t-SNE on HMDB→UCF.

Results on Epic-Kitchens. Table 2 shows the results on Epic-Kitchens, which is another challenging dataset consisting of total 6 transfer tasks with a data imbalance across different action classes. Our method achieves the best performance on all 6 tasks with the average performance of **52.3%**. Compared to adversarial alignment (AdaBN Li et al. (2018) and MCD Saito et al. (2018)), self-correspondence learning (MM-SADA Munro & Damen (2020)) and contrastive learning methods (STCDA Song et al. (2021) and LCMC Kim et al. (2021)), similar observations can be found that our method outperforms all the competing methods in the challenging Epic-Kitchens dataset by **1.1%**.

Table 1: **Results on UCF-HMDB dataset.** Two-stream denotes the methods using I3D two-stream network for both RGB and flow stream respectively. Supervised Target denotes the baseline training with labeled target data only. * stands for the self-implementation.

Method	Backbone	UCF → HMDB	HMDB → UCF
Source-only Chen et al. (2019a)	ResNet-101	71.6	73.9
DANN Ganin et al. (2016)	ResNet-101	75.3	76.4
JAN Long et al. (2017)	ResNet-101	74.7	79.3
AdaBN Li et al. (2018)	ResNet-101	75.5	77.4
MCD Saito et al. (2018)	ResNet-101	74.4	79.3
TA^{3}N Chen et al. (2019a)	ResNet-101	78.3	81.8
ABG Luo et al. (2020)	ResNet-101	79.1	85.1
TCoN Pan et al. (2020a)	ResNet-101	87.2	89.1
Supervised Target Chen et al. (2019a)	ResNet-101	82.8	94.9
Source-only Choi et al. (2020)	I3D	80.3	88.8
DANN Ganin et al. (2016)	I3D	80.7	88.0
ADDA Tzeng et al. (2017)	I3D	79.1	88.4
MSTN* Ganin et al. (2016)	I3D	80.2	89.9
CRST* Long et al. (2017)	I3D	80.9	90.2
TA^{3}N Chen et al. (2019a)	I3D	81.4	90.5
SAVA Choi et al. (2020)	I3D	82.2	91.2
STCDA Song et al. (2021)	I3D	81.9	91.9
Ours	I3D	**84.0**	**94.7**
Supervised Target Choi et al. (2020)	I3D	95.0	96.8
Source-only Song et al. (2021)	Two-stream	82.8	90.7
MM-SADA Munro & Damen (2020)	Two-stream	84.2	91.1
STCDA Song et al. (2021)	Two-stream	83.1	92.1
LCMC Kim et al. (2021)	Two-stream	84.7	92.8
Ours	Two-stream	**86.1**	**95.4**
Supervised Target Song et al. (2021)	Two-stream	95.8	97.7

4.3 Ablation Study and Analysis

Effectiveness of Proposed Positives. As presented from Table 3, intra-domain positives bring a significant improvement on UCF→HMDB by **2.6%** and HMDB→UCF by **5.0%**. Cross-domain positives also improve the performance by a large margin. Surprisingly, each proposed component individually outperforms the previous state-of-the-art methods (e.g., SAVA Choi et al. (2020) and STCDA Song et al. (2021)) on the I3D backbone. Incorporating these two components into the same framework brings further gains on both UCF→HMDB by **3.7%** and HMDB→UCF by **5.9%**. These results suggest that intra/cross-domain positives could benefit each other. Precisely, the better target domain discrimination brought by *intra-domain positives* helps the better estimation of target distribution in Eqn.5. In return, *cross-domain positives* help cross-domain alignment and result in more accurate target NNs for *intra-domain positives* in Eqn. 4. Last, the performance could be further boosted with our optimization without negatives.

Effect of Optimization w/o Negatives. As shown in Table 3, optimization without negatives boosts the performance of both intra-domain positive, cross-domain positives and jointly training by a significant margin. Explicitly, when jointly training, it improves **1.5%** on UCF→HMDB and **1.6%** on HMDB→UCF compared to the baseline of conventional contrastive optimization with negatives.

Effect of Augmentation Strategies. When $k = 1$, our *Intra-Pos* could generalize to self-supervised learning where the target anchor videos are pushed to be close to their augmentations. Explicitly, we consider spatial augmentations (e.g., random cropping, color distortion, and temporal augmentations including video playrate change and clips shuffling. Table 4 shows that all types of augmentations benefit the performance, and integrating all of them gives the best performance. Interestingly, UCF→HMDB favors the temporal augmentations more than the spatial augmentations while an opposite observation is found in HMDB→UCF.

Table 3: **Ablation Study on UCF-HMDB for each component. Intra-Pos**: Intra-domain Positives in Eqn.4, **Cross-Pos**: Cross-domain Positives in Eqn.6, **w/o Negatives**: Optimization w/o negatives in Sec.3.1

Intra-Pos	Cross-Pos	w/o Neg	UCF → HMDB	HMDB → UCF
✗	✗	✗	80.3	88.8
✓	✗	✗	81.8	92.6
✓	✗	✓	82.9	93.8
✗	✓	✗	80.9	90.9
✗	✓	✓	82.6	92.3
✓	✓	✗	82.5	93.1
✓	✓	✓	**84.0**	**94.7**

Table 4: **Ablation Study on augmentation strategies and cross-domain positive.** The gray rows are the **Intra-Pos** with all augmentations and the **Cross-Pos** with covariance estimiation and w/o predictor.

Method	UCF → HMDB	HMDB → UCF
Intra-Pos (default, k=1)	81.8	92.3
w/ Spatial Aug.	80.8	92.1
w/ Speed Aug.	81.1	91.2
w/ Shuffle Aug.	81.3	91.4
Cross-Pos	82.6	92.3
w/o Σ_c^t	82.2	91.8
w/ predictor	80.2	89.1

Effect of Class-conditional Covariance. Our *cross-domain* alignment pulls a source video to the synthesized target video representations from the estimated target distribution. We claim that the estimated class mean feature vector provides the shared action semantics from the target domain. Its corresponding covariance brings target spatial-temporal variations to the source samples. Table 4 shows a performance drop on *Cross-domain* without the covariance Σ_c^t on both tasks of UCF-HMDB, demonstrating the importance of integrating target variations.

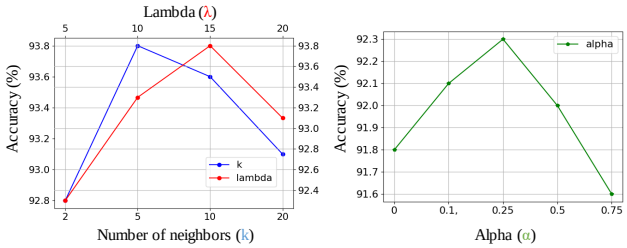


Figure 4: Hyperparameters sensitivity on the number of neighbors k (bottom x-axis, left y-axis), intra-target coefficient λ (top x-axis, right y-axis) and class covariance coefficient α_0 on HMDB→UCF.

Effect of Predictor in Cross-domain postives. Unlike previous non-contrastive learning method Chen & He (2021), we find that applying the predictor results in poor performance in Tab. 4. We further tried increasing the optimization strength with larger loss ratio but still failed. *We hypothesize that, due to the domain shift, pulling the source features closer to the target requires non-trivial optimization effort and thus the model is less prone to the "collapsing solutions". Therefore, adding a predictor would make the optimization even harder.*

Hyperparameters Sensitivity. Figure 4 presents the effect of choosing different neighbors k with $\lambda = 15$. We observe that adding neighbors in Eqn.4 generally boosts the performance and $k = 5$ gives the best performance. For λ, we set it λ from $\{5, 10, 15, 20\}$ with $k = 5$. Empirically, we observe that relatively large λ helps the convergence on the predictor in Figure 4, and $\lambda = 15$ achieves the best performance. For α_0, it can be implicitly considered as the radius of target class distribution. It is shown that Σ_c^t becomes most effective when $\alpha_0 = 0.25$ in Figure 4. In this case, the sampled target variations from Σ_c^t are more likely to be clean. In summary, our method is robust to different hyperparameters.

Feature Visualizations. We visualize the feature embedding by t-SNE Maaten & Hinton (2008) among each component of proposed methods, including the source only model, intra-domain, cross-domain and ours. As seen from Figure 3, intra-domain shows better class discrimination while cross-domain aligns two domains more tightly. Ours incorporates the merits of both two methods.

5 Conclusion and Future Work

In this work, we introduce the bottlenecks of in existing contrastive-based video DA methods and propose a unified solution to address them without relying on negatives by mining informative and robust intra-domain positives and cross-domain positives. Extensive experiments on several benchmark datasets demonstrate the superiority of our approach over state-of-the-art methods. Regarding the future work, multimodal learning would be an interesting direction to be integrated with our algorithm by mining cross-modal nearest neighbors and estimating target distribution from other modalities.

ETHICS STATEMENT

In our paper, we strictly follow the ICLR ethical research standards and laws. All datasets we employed are publicly available, and all related publications and source codes are cited appropriately.

REPRODUCIBILITY STATEMENT

We highlight several components within this submission that effectively support the reproducibility of this work. **1)** Implementation details are provided in Appendix C, where the training dataset, backbone architecture, training schemes and hyperparameters are carefully stated. **2)** A more detailed pseudo code for our algorithm can be found in Alg. 1.

REFERENCES

Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can temporal information help with contrastive self-supervised learning? 2021.

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*, pp. 9922–9931, 2020.

Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017a.

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*, pp. 6299–6308, 2017b.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

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*, pp. 6321–6330, 2019a.

Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *IEEE International Conference on Computer Vision*, pp. 2090–2099, 2019b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

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*, pp. 678–695, 2020.

Inseop Chung, Daesik Kim, and Nojun Kwak. Maximizing cosine similarity between spatial features for unsupervised domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1351–1360, 2022.

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)*, pp. 720–736, 2018.

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*, pp. 6202–6211, 2019.

Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3636–3645, 2017.

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.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020.

Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *NeurIPS*, 2020.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning Spatio-Temporal Features with 3D Residual Networks for Action recognition. In *IEEE International Conference on Computer Vision Workshops*, pp. 3154–3160, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014a.

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*, pp. 1725–1732, 2014b.

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

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *Proceedings of the International conference on Computer Vision*, 2011.

Shuang Li, Mixue Xie, Kaixiong Gong, Chi Harold Liu, Yulin Wang, and Wei Li. Transferable semantic augmentation for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11516–11525, 2021.

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive Batch Normalization for Practical Domain Adaptation. *Pattern Recognition*, 80:109–117, 2018.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pp. 97–105. PMLR, 2015.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International Conference on Machine Learning*, pp. 2208–2217. JMLR. org, 2017.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pp. 1640–1650, 2018.

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*, pp. 19–27, 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2018.

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*, pp. 122–132, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, pp. 69–84, 2016.

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*, pp. 11815–11822, 2020a.

Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3764–3773, 2020b.

Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. *arXiv preprint arXiv:2106.15326*, 2021.

Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *Advances in Neural Information Processing Systems*, 34, 2021.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3723–3732, 2018.

Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *International Conference on Machine Learning*, pp. 1275–1282, 2012.

Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems*, pp. 568–576, 2014.

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

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

Hui Tang, Ke Chen, and Kui Jia. Unsupervised domain adaptation via structurally regularized deep clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8725–8735, 2020.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Xiaofu Wu, Suofei Zhang, Quan Zhou, Zhen Yang, Chunming Zhao, and Longin Jan Latecki. Entropy minimization versus diversity maximization for domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Shaoan Xie, Zibin Zheng, Liang Chen, and Chuan Chen. Learning semantic representations for unsupervised domain adaptation. In *Proceedings of the International conference on machine learning*, pp. 5423–5432. PMLR, 2018.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training. *arXiv preprint arXiv:1810.07911*, 2018.

Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *IEEE International Conference on Computer Vision*, pp. 5982–5991, 2019.

A APPENDIX

In this material, we show the additional optimization detail, comparison to Imagge-based Domain adaptation methods and algorithm pseudo code to our main submission.

B OPTIMIZATION DETAIL

In summary, our framework can be simply implemented and optimized in an end-to-end deep learning framework through the overall objective function:

$$\mathcal{L}_{all}^{rgb} = \mathcal{L}_{s-ce}^{rgb} + \mathcal{L}_{s-cross}^{rgb}{}^{\infty} + \lambda \mathcal{L}_{t-intra}^{rgb}, \tag{8}$$

B.1 CLOSE-FORM UPPER-BOUND FOR CROSS-POSITIVES OPTIMIZATION

The naive implementation is not computational efficient when M is large. To tackle this issue, we consider the case that the number of M goes to infinity, and then derive a close-form upper-bound for the expectation of Eqn. 6 under all possible target features as:

$$\begin{aligned}
\mathcal{L}_{s-cross}^{\infty} &= \mathbb{E}_{\tilde{z_i}^t}[\log e^{\mathcal{D}(z^s, \tilde{z_i}^t)}], && (9)\\
&\leq -\log \mathbb{E}_{\tilde{z_i}^t}[e^{z^s \tilde{z_i}^t}], && (10)\\
&= -\log e^{z^s \mu_{y_s}^t + \frac{\alpha}{2}(z^s)^T \Sigma_{y_s}^t z^s}, && (11)\\
&= \mathcal{D}(z^s, \mu_{y_s}^t) - \frac{\alpha}{2}(z^s)^T \Sigma_{y_s}^t z^s. && (12)
\end{aligned}$$

For simplicity, all the z^s, $\tilde{z_i}^t$, $\mu_{y_s}^t$ and $\Sigma_{y_s}^t$ above are l_2-normalized if not coupled with $\mathcal{D}(\cdot,\cdot)$. The Eqn.10 emerges from the Jensen inequality $\mathbb{E}\log(X) \leq log\mathbb{E}[X]$ on the concave function $\log(\cdot)$. Based on the moment generation function $\mathbb{E}[e^{aX}] = e^{a\mu + \frac{1}{2}a^T \Sigma a}$ where $X \sim \mathcal{N}(\mu, \Sigma)$, Eqn.11 can be obtained from Eqn.10 under Gaussian assumption. The overall cross-domain alignment loss above is averaged for all source videos N_s. Notably, the computational cost of Eqn. 12 is different from backpropagating losses over $M \times N_s$ source-target pairs. As the $\mathcal{N}(\mu_{y_s}^t, \alpha\Sigma_{y_s}^t)$ is estimated offline based on the memory bank, Eqn. 12 can be computed efficiently without explicitly sampling target features.

B.2 EXTENSION TO TWO-STREAM NETWORK

For fair comparison to the methods that utilizing optical flow such as MM-SADA Munro & Damen (2020), STCDA Song et al. (2021) and LCMC Kim et al. (2021), our method could be extended to two-stream network using both RGB and optical flow as inputs as well. Following the setting above, we have:

$$\mathcal{L}_{all}^{flow} = \mathcal{L}_{s-ce}^{flow} + \mathcal{L}_{s-cross}^{flow}{}^{\infty} + \lambda \mathcal{L}_{t-intra}^{flow}. \tag{13}$$

Besides $\mathcal{L}_{all}^{rgb}$ and $\mathcal{L}_{all}^{flow}$, we also incorporate the cross-modality self-supervision in target domain as:

$$\mathcal{L}_{t-intra}^{cross}(x_t^{rgb}, x_t^{flow}) = \mathcal{D}(z_t^{rgb}, z_t^{flow}) + \mathcal{D}(z_t^{flow}, z_t^{rgb}), \tag{14}$$

where $z_t^{rgb} = F^r(x_t^{rgb}), z_t^{flow} = F^f(x_t^{flow})$ are the features of RGB and flow images. Similar to the cross-domain positives case, we do not add predictor after features in optimization as cross-modality alignment in features space does not lead to the collapsed outputs.

To sum up, the overall objective function for two-stream network is shown as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{all}^{rgb} + \mathcal{L}_{all}^{flow} + \lambda \mathcal{L}_{t-intra}^{cross}, \tag{15}$$

Algorithm 1 Robust Cross-domain Positives for video DA

Input: Labeled source domain $\mathbb{D}_s = \{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)|_{i=1}^{N_s}\}$, unlabeled target domain $\mathbb{D}_t = \{(\boldsymbol{x}_i^t)|_{i=1}^{N_t}\}$; warm-start iteration T_0, maximum iteration T and batch size B; hyper-parameters: λ and α_0.
Output: Parameters of model $\boldsymbol{\Theta}_F$ and predictor $\boldsymbol{\Theta}_H$.

for $t = 1$ to T_0 **do**
  Update $\boldsymbol{\Theta}$ with labeled source videos $\mathcal{L}_{ce}^s$.
end for
for $t = T_0$ to T **do**
  $\alpha_= (t/T) \times \alpha_0$
  Sample $\{(\boldsymbol{x}_i^s, \boldsymbol{y}_i^s)\}_{i=1}^B$ and $\{(\boldsymbol{x}_i^t)\}_{j=1}^B$ from $\mathbb{D}_s$ and $\mathbb{D}_t$, respectively.
  Obtain deep features $\{\mathbf{z}_i^s\}_{i=1}^B, \{\mathbf{z}_j^t\}_{j=1}^B$ and softmax outputs $\{\hat{\boldsymbol{p}}_i^s\}_{i=1}^B, \{\hat{\boldsymbol{p}}_j^t\}_{j=1}^B$ for source and target samples, respectively.
  Generate target pseudo labels based on $\{\hat{\boldsymbol{p}}_j^t\}_{j=1}^B$.
  Update memory bank $\mathbb{M}_t = \{(z_i^t, \hat{y_i}^t)|_{i=1}^{N_t}\}$ with target features and pseudo labels in current batch t.
  Randomly apply strong spatial or temporal augmentation on the target samples as the second view $\{(\hat{\boldsymbol{x}_i^t})\}_{j=1}^B$.
  Obtain their deep features and transform them as $\{\mathbf{h}_j^t\}_{j=1}^B$ by the predictor $\boldsymbol{\Theta}_H$.
  Given the current target features $\{\mathbf{z}_j^t\}_{j=1}^B$, we mine the top-K nearest neighbors from $\mathbb{M}_t$ as $\{\hat{z_j}^t|_{j=1}^k\}$ and compute the intra-domain positives loss $\mathcal{L}_{t-intra}$.
  for each class c **do**
    Estimate target features means $\boldsymbol{\mu}_t^c$ according to memory module $\mathbb{M}_t$.
    Estimate the target intra-class covariance matrix $\boldsymbol{\Sigma}_t^c$ according to memory module $\mathbb{M}_t$.
  end for
  Based on $\mathcal{N}(\mu_c^t, \alpha\Sigma_c^t)$, compute the cross-domain alignment loss $\mathcal{L}_{s-cross}^{\infty}$.
  Update $\boldsymbol{\Theta}_F$, $\boldsymbol{\Theta}_H$ by minimizing the loss $\mathcal{L}_{all}^{rgb}$ in Eqn. 8 with stochastic gradient descent (SGD).
end for

C Implementation Details

Following Choi et al. (2020); Song et al. (2021); Kim et al. (2021), we use I3D Carreira & Zisserman (2017b) as the backbone feature encoder network, initialized with Kinetics pre-trained weights. The dimension of the features extracted from the I3D encoder is 1024. We randomly sample 16 consecutive frames out of a video clip with a size of 224×224. For inference, we use 16 uniformly sampled frames per video to recognize the action. We follow the standard 'pre-train then adapt' procedure used in prior works Tzeng et al. (2017); Choi et al. (2020) to train the model with only source data for 3 epochs as a warm start before the proposed approach is employed. Then we train models in Eqn. 8 for 40 epochs in total. For the hyperparameters, we select λ from $\{5, 10, 15, 20\}$, α_0 from $\{0.1, 0.25, 0.5, 0.75\}$ and the number of neighbors k from $\{1, 2, 5, 10, 20\}$. We conduct hyperparameter sensitivity experiments in sec. 4.3. We set $k = 5$, $\lambda = 15$ and $\alpha_0 = 0.25$ for all datasets. We 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 0.01 for the I3D with a cosine learning rate scheduler for our experiments. We use a batch size of 32 equally split over the two domains. We report the average action recognition accuracy over 3 random trials. We use four 12G NVIDIA GPUs for training.

D Comparison to Image-based Domain adaptation Baseline

We compare our approach with the following Image-based DA baselines. (1) existing UDA based on advanced pseudo-labeling and entropy minimization methods which can be considered as cross(inter)-domain alignment approaches, including CRST Zou et al. (2019) and MSTN Xie et al. (2018) and MEDM Wu et al. (2021). (2) existing DA method based on intra-domain discrimination IDA Pan et al. (2020b).

Table 5: **Results on UCF-HMDB dataset for Image-based DA baselines.** Supervised Target denotes the baseline training with labeled target data only. * stands for the self-implementation.

Method	**Backbone**	**UCF $\rightarrow$ HMDB**	**HMDB $\rightarrow$ UCF**
Source-only Choi et al. (2020)	I3D	80.3	88.8
MSTN* Ganin et al. (2016)	I3D	80.2	89.9
CRST* Long et al. (2017)	I3D	80.9	90.2
MEDM* Wu et al. (2021)	I3D	80.6	89.8
IDA* Pan et al. (2020b)	I3D	81.6	91.4
Ours	I3D	**84.0**	**94.7**
Supervised Target Choi et al. (2020)	I3D	95.0	96.8

We shows results of our method with other competing approaches on UCF-HMDB dataset in Table 5. Our framework outperforms both inter-domain and intra-domain methods by a significant margin.