Augmenting and Aligning Snippets for Few-Shot Video Domain Adaptation

Yuecong Xu[1*] Jianfei Yang[2*] Yunjiao Zhou[2] Zhenghua Chen[1†] Min Wu[1] Xiaoli Li[1]
[1]Institute for Infocomm Research, A*STAR, Singapore [2]Nanyang Technological University
{xuyu0014, yang0478, yunjiao001, chen0832}e.ntu.edu.sg {wumin, xlli}@i2r.a-star.edu.sg

Abstract

For video models to be transferred and applied seamlessly across video tasks in varied environments, Video Unsupervised Domain Adaptation (VUDA) has been introduced to improve the robustness and transferability of video models. However, current VUDA methods rely on a vast amount of high-quality unlabeled target data, which may not be available in real-world cases. We thus consider a more realistic Few-Shot Video-based Domain Adaptation (FSVDA) scenario where we adapt video models with only a few target video samples. While a few methods have touched upon Few-Shot Domain Adaptation (FSDA) in images and in FSVDA, they rely primarily on spatial augmentation for target domain expansion with alignment performed statistically at the instance level. However, videos contain more knowledge in terms of rich temporal and semantic information, which should be fully considered while augmenting target domains and performing alignment in FSVDA. We propose a novel SSA^{2}lign to address FSVDA at the snippet level, where the target domain is expanded through a simple snippet-level augmentation followed by the attentive alignment of snippets both semantically and statistically, where semantic alignment of snippets is conducted through multiple perspectives. Empirical results demonstrate state-of-the-art performance of SSA^{2}lign across multiple cross-domain action recognition benchmarks. Code will be provided at: https://github.com/xuyu0010/SSA2lign.

1. Introduction

Video Unsupervised Domain Adaptation (VUDA) [4, 7, 51, 46, 53] aims to improve the generalizability and robustness of video models by transferring knowledge to new domains, and is widely applied in scenarios where massive labeled videos are unavailable. Current VUDA methods assume that sufficient target data are accessible which enables domain alignment by minimizing cross-domain distribution discrepancies and obtaining domain invariant representations [4, 7, 54]. However, this assumption may not be feasible in real-world applications such as in smart hospitals and security surveillance where video models are leveraged for anomaly behavior recognition [35, 32], and are expected to be functional at all times even across different environments. It is more practical to obtain a few labeled videos during the early stage of model deployment to improve the transferred models' performances in the new (target) environment. A *Few-Shot Video Domain Adaptation* (*FSVDA*) task is hence formulated to enable knowledge learned from labeled source video to be transferred to the target video domain given only very limited labeled target videos.

With only several target domain samples, FSVDA is more challenging than VUDA, since aligning distributions with limited samples is harder. A few research have touched on the image-based Few-Shot Domain Adaptation (FSDA) [27, 45, 49, 11] by domain alignment, e.g., moment matching or adversarial training, between a spatial-augmented target domain and a filtered target-similar source domain. More recently, there have been a few early research on FSVDA [12, 13] which extends the above strategies to videos by viewing each video sample as a whole and obtaining frame-based video features.

However, there are two major shortcomings when the image-based FSDA is applied to video domains. Firstly, applying frame-level spatial augmentation towards individual video frames ignores and undermines temporal correlation across sequential frames, and we find that such augmentation would result in only minor or even negative effects on FSVDA performance. Secondly, the effectiveness of domain alignment methods is built upon sufficient source domain and target domain data that depicts the distribution discrepancy, which is not available in FSVDA. Even worse, statistical estimation of video data distribution is less accurate due to the complicated temporal structure of video data. In this paper, we aim to overcome these two challenges by designing more effective target domain augmentation and semantic alignment in the spatial-temporal domain.

To this end, we propose to address the FSVDA task by a **S**nippet-attentive **S**emantic-statistical **Align**ment with

[*]Equal contribution.
[†]Corresponding author.

Stochastic **S**ampling **A**ugmentation (**SSA²lign**). Instead of aligning features of whole video samples at the video level or frame level [12, 13], we align source and target video features at the snippet level. Snippets are formed from a limited series of adjacent sequential frames, thus they contain both spatial and short-term temporal information. Leveraging snippet features for FSVDA brings two unique advantages: i) a larger amount of target domain samples could be obtained via spatial-temporal augmentations on snippets, obtaining more diverse features across the temporal dimension; ii) additional alignment of the diverse but highly correlated snippet features of each video could further improve the discriminability of the corresponding videos, which has been proven to benefit the effectiveness of video domain adaptation [6, 59, 21, 53]. SSA²lign is therefore proposed. It firstly augments the source and target domain data by a simple yet effective stochastic sampling process that makes full use of the abundance of snippet information and then performs semantic alignment from three perspectives: alignment based on semantic information within each snippet, cross-snippets of each video, and across snippet-level data distribution. Our method is demonstrated to be very effective for the FSVDA problem, surpassing state-of-the-art methods by large margins on two VUDA benchmarks.

In summary, our contributions are threefold. (i) We propose a novel SSA²lign to address FSVDA at the snippet level by both statistical and semantic alignments that are achieved from three perspectives. (ii) We propose to augment target domain data and the snippet-level alignments by a simple yet effective stochastic sampling of snippets for more robust video domain alignment. (iii) Extensive experiments show the efficacy of SSA²lign, achieving a remarkable average improvement of 13.1% and 4.2% over current state-of-the-art FSDA/FSVDA methods on two large-scale cross-domain action recognition benchmarks.

2. Related Work

(Video) Unsupervised Domain Adaptation ((V)UDA). Current UDA and VUDA methods aim to transfer knowledge from the source to the target domain given that both domains contain sufficient data, improving the transferability and robustness of models [50, 57]. They could be generally divided into four categories: a) reconstruction-based methods [14, 56], where domain-invariant features are obtained by encoders trained with data-reconstruction objectives; b) adversarial-based methods [4, 51, 7], where feature generators obtain domain-invariant features leveraging domain discriminators, trained jointly in an adversarial manner [17, 10]; c) semantic-based methods [58, 53], which exploit the shared semantics across domains such that domain-invariant features are obtained; and d) discrepancy-based methods [33, 62], which mitigate domain shifts by applying metric learning, minimizing metrics such as MMD [25] and CORAL [36]. With the introduction of cross-domain video datasets such as Daily-DA [54] and Sports-DA [54], there has been a significant increase in research interest for VUDA [29, 5]. Despite the gain in video model robustness thanks to VUDA methods, they all assume that sufficient target data are accessible, which may not be feasible in real-world cases where a large amount of superior unlabeled target data are not available. A more related VUDA method concerns SAVA [7] which also utilizes the clips to design a self-supervised learning task (clip order prediction), but the adaptation is still performed with the video-level feature. We differ from SAVA [7] where our alignment is performed at the snippet level considering three different perspectives: within each snippet, cross-snippets of each video, and across snippet-level data distribution, therefore leading to better performances in FSVDA.

Few-Shot (Video) Domain Adaptation (FS(V)DA). It is more practical to obtain a few labeled target data to aid video models to adapt. There have been a few research that explores image-based FSDA. Among them, FADA [27] is adversarial-based and augments the domain discriminator to classify 4 types of source-target pairs. d-SNE [49] learns a latent space through SNE [16] with large-margin nearest neighborhood [9], and utilizes spatial augmentations to create sibling target samples. AcroFOD [11] explores FSDA for object detection by applying multi-level spatial augmentation and filtering target-irrelevant source data. There are also works as in [63, 38, 37, 60] that combine domain adaptation (DA) with few-shot learning (FSL), yet we differ them in the assumption of similar target and source classes and only limited target data accessible, which is more realistic. More recently, there have been a few early research on FSVDA, including PASTN [12] that constructs pairwise adversarial networks performed across source-target video pairs, while PTC [13] further leverages optical flow features. Both PASTN and PTC obtain video features from a frame-based video model. Despite some advances made in FS(V)DA, the above methods have not tackled FSVDA effectively by leveraging the rich temporal information as well as semantic information embedded within videos. We propose to engage in FSVDA by augmenting and attentively aligning snippet-level features which contain temporal information via both semantic and statistical alignments.

3. Proposed Method

For *Few-Shot Video Domain Adaptation*, we are given a labeled source domain $\mathcal{D}_S = \{(V_{S,i}, y_{S,i})\}_{i=1}^{N_S}$ with sufficient N_S i.i.d. source videos across $\mathcal{C}$ classes, characterized by a probability distribution of p_S. We are also given a labeled target domain $\mathcal{D}_T = \{(V_{T,j}, y_{T,j})\}_{j=1}^{N_T}$ with a limited number of $N_T \ll N_S$ i.i.d. target videos across the same $\mathcal{C}$ classes, where each video class only contains k target video samples (corresponding to the k-shot Video Domain Adaptation task), thus $N_T = k \times \mathcal{C}$. $\mathcal{D}_T$ is characterized by a probability distribution of p_T.

Owing to the absence of sufficient target data and the lack of target information, FSVDA is more challenging than VUDA. Current VUDA methods [4, 51] that are primarily moment matching-based are ineffective without target information for domain alignment. FSVDA should be tackled by leveraging the temporal information of videos fully for more temporally diverse features while aligning with the embedded semantic information to improve video discriminability for effective video domain adaptation. We propose SSA^{2}lign, a novel method to transfer knowledge from the source domain to the target domain with only limited labeled target data by obtaining, augmenting, and aligning snippet features attentively. We start by introducing how snippet features are obtained and augmented through the Stochastic Sampling Augmentation (SSA), followed by a detailed illustration of the proposed SSA^{2}lign.

3.1. Snippet Features with the Stochastic Sampling Augmentation

The key to effective target domain expansion and domain alignment in FSVDA is to obtain and augment features with temporal information such that the augmented features are diverse temporally. While various spatial augmentation methods (e.g., color jittering, flipping, cropping) have been adopted in supervised action recognition thanks to their capability in improving the robustness of video models, and in prior FSDA for expanding the target domain $\mathcal{D}_T$, they are performed at the frame-level across randomly selected individual frames. Meanwhile, the temporal information corresponds to the correlation of sequential frames and would be undermined by spatial augmentation since sequential frames may not be equally augmented. Augmentations for FSVDA must be performed above the frame level.

Snippets are formed from a limited series of adjacent sequential frames and have been utilized in multiple supervised action recognition methods (e.g., TSN [44] and STPN [48]) thanks to their ability in including both spatial and short-term temporal information. Therefore, we align source and target video features at the snippet level. Mathematically, given a target video $V = [f^1, f^2, ..., f^n]$ that contains n frames, we denote the i-th frame as f^i. We denote the length of a snippet s to be m, then video V would contain $n - m + 1$ snippets in total. We define a snippet $s^j = [f^j, f^{j+1}, ..., f^{j+m-1}]$ as the snippet starting from the j-th frame. While given only $N_T = k \times \mathcal{C}$ target videos, there are $N_T \times (n-m+1)$ target snippets, which can greatly expands the target domain for domain alignment while preserving essential temporal information.

While the target domain is largely expanded, utilizing all snippets for alignment is computationally inefficient (a 10-second 30-fps video contains more than 290 8-frame snippets). Moreover, snippets that are obtained adjacently would differ over only ONE frame, resulting in high redundancy in temporal information. To ensure that diverse temporal information is utilized, we adopt a simple Stochastic Sampling Augmentation (SSA) over the snippets. Formally, during training we sample $r > 1$ snippets s_b^a stochastically per target video per mini-batch, where $a \in [1, n - m + 1]$ denotes the starting frame of the snippet and $b \in [1, r]$ denotes the b-th snippet sampled. SSA further ensures that the sampled snippets are diverse from two perspectives. Firstly, SSA samples snippets with a minimum of $\hat{m}$ difference between the starting frame of any two snippets from the same target video, that is $\forall b_x \in [1, r], b_y \in [1, r]$ with $b_x \neq b_y$, we set $|a_x - a_y| \geqslant \hat{m}$. Secondly, since there are more source videos than target videos during training, it is likely that the same target video would be encountered across different mini-batches. SSA ensures that different snippets are sampled each time the same target video is included in a mini-batch across the same training epoch.

The SSA is also applied to the source videos to obtain source snippets. However, since there are sufficient source videos, it is more reasonable and efficient to exploit source knowledge with different source videos rather than the different snippets of a source video that would contain redundant source knowledge. Therefore, we only sample $r = 1$ snippet stochastically per source video via SSA.

Another crucial step towards transferring source knowledge to the target domain is to obtain rigorous snippet features that include both spatial and temporal information. We resort to the Transformer-based TimeSFormer [2] which extracts spatial and temporal features with separate space-time attention blocks based on self-attention [41]. While various Transformer-based video models achieve competitive performances on action recognition, TimeSFormer possesses the least amount of parameters, requiring only 60% parameters of Swin [24] and only 40% parameters of ViViT [1]. The feature of snippet s_b^a is $\mathbf{f}_b = Time(s_b^a)$ where $Time$ denotes the TimeSFormer.

3.2. Snippet-attentive Semantic-statistical Alignment with SSA

With the absence of sufficient target data, conventional VUDA methods that are primarily moment matching-based would not be fully effective since target data distribution is unknown. Alternatively, we tackle FSVDA at the snippet level by aligning the embedded semantic information from three perspectives: aligning based on the semantic information within each snippet, cross-snippets of each video, and across snippet-level data distribution. Statistical alignment is also adopted for more stable domain alignment, while both alignments attend to the more impactful snippets.

Following the above strategy, we propose the **Snippet-attentive Semantic-statistical Alignment (SSAlign)**, with

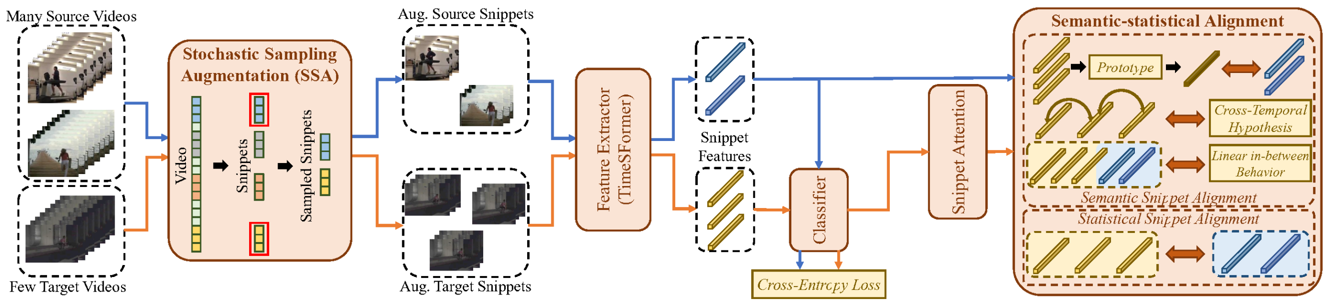


Figure 1. Pipeline of SSA^{2}lign. Source and target snippets are first obtained through the Stochastic Sampling Augmentation, whose features are obtained through the shared feature extractor. SSA^{2}lign then aligns the source and target domains at the snippet level with the Semantic-statistical Alignment, while weighing the impact of different target snippets through snippet attention, whose weight is built based on the output prediction of target snippets, obtained from a shared classifier with source snippets. The blue and orange lines imply the data flow for source and target videos respectively.

the input obtained through SSA introduced in Sec. 3.1, forming the **SSA^{2}lign**. The overall pipeline of SSA^{2}lign is presented in Fig. 1. We obtain the augmented source and target snippets through SSA whose features are extracted by applying TimeSFormer. We denote a source snippet from the i-th source video as $s_{S,i}$ and its feature as $\mathbf{f}_{S,i}$, while the l-th target snippet ($l \in [1, r]$) from the j-th target video as $s_{T,jl}$ and its feature as $\mathbf{f}_{T,jl}$. The superscript of the snippet expression is omitted for clarity. Domain alignment is achieved by performing both the Semantic Snippet Alignment and the Statistical Snippet Alignment. The snippet attention is applied to the augmented target snippets to weigh the snippets dynamically. The TimeSFormer feature extractor $Time$ is shared across source and target domains while a shared classifier H outputs a prediction o for the source and target snippets, optimized through a cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{pred} = &\frac{1}{N_S}\sum\nolimits_{i=1}^{N_S} l_{ce}(o_{S,i}, y_{S,i}) \\ &+ \frac{1}{N_T \times r}\sum\nolimits_{j=1}^{N_T}\sum\nolimits_{l=1}^{r} l_{ce}(o_{T,jl}, y_{T,j}),\end{aligned} \tag{1}$$

where $o_{S,i} = \sigma(H(\mathbf{f}_{S,i}))$ and $o_{T,jl} = \sigma(H(\mathbf{f}_{T,jl}))$ are the output predictions of snippet features $\mathbf{f}_{S,i}$ and $\mathbf{f}_{T,jl}$, while σ denotes the SoftMax function.

Semantic Snippet Alignment. The purpose of applying semantic alignment at the snippet level is to match the embedded semantic information (e.g., each individual feature or characteristic over a set of features) across source and target domains. Since both domains share the same TimeSFormer feature extractor, this implies that for each individual snippet feature, those of the same class should be close together across both domains. However, it is computationally expensive to compute the distances between each source and target snippet features given their large quantity. Inspired by the Prototypical Network [34, 23] designed for few-shot learning [65, 43], we resort to a more efficient solution where semantic alignment across each snippet is performed by minimizing the distance between source snippet features and target prototypes. The target prototypes are obtained for each individual class C_x as the mean feature of all target snippet features classified as C_x, formulated as:

$$Pr_x = \frac{1}{n_{T,x}}\sum\nolimits_{\forall s_{T,jl}\in C_x} \mathbf{f}_{T,jl}, \tag{2}$$

where $n_{T,x}$ is the number of target snippets classified as class C_x. For stable and effective alignment, the snippet features for computing the target prototypes are obtained after e training epochs. Target prototypes are subsequently updated per epoch by their exponential moving average as:

$$Pr_x \leftarrow \lambda_P Pr_x + (1-\lambda_P)Pr'_x, \tag{3}$$

where Pr_x and Pr'_x denote the target prototype of class C_x computed at the current and previous epochs. Aligning source snippet features towards target prototypes is thus achieved by minimizing the Euclidean distances between them and denoted as the prototype alignment loss as:

$$\mathcal{L}_{proto} = \frac{1}{N_S}\sum\nolimits_{x=1}^{\mathcal{C}}\sum\nolimits_{i=1}^{n_{S,x}}\sqrt{(\mathbf{f}_{S,i}-Pr_x)^2}. \tag{4}$$

$n_{S,x}$ is the number of source snippets classified as class C_x.

Besides the capability of obtaining temporally diverse features via SSA, leveraging snippet features for FSVDA is also more advantageous due to the inclusion of additional semantic information that exists across the diverse but highly correlated snippet features obtained from the same video, which should also be aligned. However, since we aim to exploit more source information with different source videos, the source cross-snippet semantic information cannot be directly obtained. Alternatively, the *cross-temporal hypothesis* introduced in [53] provides a thorough description of the cross-snippet semantic information for the source videos. Therefore, the equivalence of aligning the cross-snippet semantic information across source and target domains is to align the cross-snippet semantic information of the target domain to the *cross-temporal hypothesis*, that is the snippet features over the snippets obtained from the same target video through SSA must be consistent. Meanwhile, aligning the *cross-temporal hypothesis* would also drive target videos to be discriminative, while previous studies [6, 59, 21, 53] have proven that improving discriminability can benefit the effectiveness of domain adaptation.

Formally, the cross-snippet consistency is achieved by minimizing the Kullback–Leibler (KL) divergence of the predictions of target snippets corresponding to the same target video. It is computed between each snippet against the key snippet of each target video, which is identified such that it is classified correctly and is certain in its prediction (i.e., low prediction entropy). In cases where no snippets are classified correctly, the snippet with the lowest prediction entropy is identified as the key snippet. The cross-snippet consistency loss is computed as:

$$\mathcal{L}_{cross} = \frac{1}{N_T(r-1)} \sum_{j=1}^{N_T} \sum_{l=1, l\neq k}^{r} KL(\log(o_{T,jy}) \| \log(o_{T,jl})), \quad (5)$$

where $KL(p\|q)$ denotes the KL-divergence while y denotes y-th snippet corresponding to the target video $V_{T,j}$ identified as the key snippet.

Aligning semantically via matching the characteristics over differed snippet features could be further performed across the snippet-level data distribution. Since source snippets for training are obtained stochastically at each training epoch, semantic information embedded across the source snippet-level data distribution changes continuously, and would therefore be ineffective for the target snippet-level data distribution to be directly aligned. Alternatively, snippet features that are highly discriminative would imply effective domain adaptation since it has been proven that improving discriminability benefits domain adaptation [6, 59, 21, 53]. We thus aim to drive the feature extractor towards obtaining snippet features that are distributed more discriminatively. Specifically, results in model robustness [61] suggest that the discriminability of features can be improved if the feature extractor behaves linearly in-between training samples. The linear in-between behavior can be complied by employing the interpolation consistency training (ICT) technique [42] across both source and target snippets, which encourages the linearly interpolated features to produce a linearly interpolated prediction. Formally, given a pair of snippet features $\mathbf{f}_*$, $\mathbf{f}_{*'}$, and their corresponding output predictions o_*, $o_{*'}$, the ICT is conducted with the following process and optimization loss:

$$\begin{aligned} \tilde{\mathbf{f}} &= \lambda_v \mathbf{f}_* + (1-\lambda_v)\mathbf{f}_{*'}. \\ \tilde{\mathbf{o}} &= \lambda_v \mathbf{o}_* + (1-\lambda_v)\mathbf{o}_{*'}. \\ \mathcal{L}_{ICT}(*,*') &= l_{ce}(\sigma(H(\tilde{\mathbf{f}})), \tilde{\mathbf{o}}), \end{aligned} \quad (6)$$

Algorithm 1 Training with SSA[2]lign for FSVDA

Input: $\mathcal{D}_S = \{(V_{S,i}, y_{S,i})\}_{i=1}^{N_S}$, $\mathcal{D}_T = \{(V_{T,j}, y_{T,j})\}_{i=1}^{N_T}$, $N_T \ll N_S$.

while Training **do**
  Obtain r target snippets $s_{T,jl}$ from $V_{T,j}$ and one source snippet $s_{S,i}$ from $V_{S,i}$ via SSA.
  Obtain features $\mathbf{f}_{S,i}$, $\mathbf{f}_{T,jl}$, predictions $o_{S,i}$, $o_{T,jl}$.
  Compute prediction loss as Eq. 1.
  Obtain snippet attention as Eq. 9 and normalize. Update $\mathbf{f}_{T,jl}$ to $\mathbf{f'}_{T,jl}$.
  if epoch $> e$ **then**
    Obtain target prototypes Pr_x as Eq. 2.-3.
    Compute prototype alignment loss as Eq. 4.
  end if
  Compute cross-snippet consistency loss as Eq. 5.
  Compute snippet distribution loss as Eq. 6-7.
  Compute and optimize overall loss as Eq. 8.
end while

Output: Trained feature extractor $Time$ and classifier H.

where $\lambda_v \in Beta(\alpha_v, \alpha_v)$ is the weight assigned to $\mathbf{f}_{T,j_1l_1}$ sampled from a Beta distribution with α_v as the parameter. We refer to previous works [22, 55] and set $\alpha_v = 0.3$. $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{o}}$ are the linearly interpolated features and the interpolated output predictions. In practice, we drive snippets to comply with the linear in-between behaviour by forming a single stochastic snippet pair for every snippet, forming $(N_T \times r + N_S)$ snippet pairs. Aligning the snippet-level data distribution with the linear in-between behavior is achieved by optimizing the snippet distribution loss as:

$$\mathcal{L}_{sn-dist} = \frac{1}{N_T \times r + N_S} \sum_{*,*' \in \{i \cup jl\}} \mathcal{L}_{ICT}(*,*'). \quad (7)$$

It is possible that a snippet pair will include two snippets from the same target video. In such case, the corresponding $\mathcal{L}_{ICT}$ across the snippet pair can be viewed as a low-ordered cross-snippet consistency loss. This implies that optimizing $\mathcal{L}_{cross}$ and $\mathcal{L}_{sn-dist}$ share the common goal of improving feature discriminability for more effective video domain adaptation.

Statistical Snippet Alignment. To improve the stability of snippet-level alignment, we adopt a statistical alignment strategy apart from the aforementioned semantic alignment strategies. The statistical alignment is performed by minimizing the snippet-level distribution discrepancies $\mathcal{L}_{sn-stat}$ formulated as metrics such as MMD [25], CORAL [36], and MDD [62]. Compared to the adversarial-based adaptation strategy more commonly used in prior VUDA tasks [4, 51, 7], minimizing discrepancies does not require additional network structures (e.g., domain classifiers), thus is more stable. The MDD [62] metric is empirically selected. The overall optimization loss function for FSVDA is therefore:

$$\mathcal{L} = \mathcal{L}_{pred} + \lambda_{sem}(\mathcal{L}_{proto} + \mathcal{L}_{cross} + \mathcal{L}_{sn-dist}) + \lambda_{stat}\mathcal{L}_{sn-stat}, \quad (8)$$

where λ_{sem} and λ_{stat} are the tradeoff hyper-parameters for the semantic and statistical snippet alignment losses.

Snippet Attention. With multiple snippets leveraged per target video for both semantic and statistical snippet alignments, it is unreasonable to leverage each snippet equally

Methods	Publication	Daily-DA													Sports-DA						
		H→A	M→A	KD→A	A→H	M→H	KD→H	H→M	A→M	KD→M	M→KD	H→KD	A→KD	Avg.	KS→U	S→U	U→S	KS→S	U→KS	S→KS	Avg.
TSF	-	37.86	32.58	31.11	44.58	57.08	45.83	36.50	30.00	34.50	61.66	58.90	75.72	45.53	91.66	91.07	76.37	77.74	87.77	85.12	84.95
TSF w/ T	-	39.57	39.49	39.41	61.67	62.92	62.50	41.50	38.75	36.50	77.79	80.28	83.59	55.33	92.48	93.42	78.95	79.05	88.02	87.00	86.49
TRX	CVPR-21	31.42	31.42	31.03	42.08	49.17	44.00	31.25	30.00	26.75	69.10	73.10	65.52	43.74	87.07	86.49	76.95	73.47	83.13	83.76	81.81
STRM	CVPR-22	33.83	32.35	32.89	43.33	50.83	44.417	30.75	29.50	28.25	72.14	74.62	68.97	45.16	91.54	90.01	78.58	75.16	86.90	84.63	84.47
HyRSM	CVPR-22	38.09	35.38	33.75	45.83	54.58	48.17	33.75	31.50	29.50	75.17	76.14	70.34	47.68	92.71	90.72	79.53	76.68	87.05	84.88	85.26
DANN	ICML-15	37.47	39.72	38.56	65.42	61.67	55.83	43.75	41.25	42.00	73.66	79.17	83.17	55.14	93.06	92.13	79.21	81.32	85.53	88.63	86.65
MK-MMD	ICML-15	35.30	42.75	35.61	64.17	63.33	56.67	44.00	41.75	36.50	76.69	81.93	79.86	54.88	92.60	93.42	80.74	77.84	84.76	88.12	86.25
MDD	ICML-19	42.51	42.28	42.90	64.58	64.17	57.92	45.00	39.50	37.75	75.17	81.52	84.28	56.47	93.18	93.07	78.47	79.105	86.70	87.72	86.37
SAVA	ECCV-20	39.18	41.66	41.74	63.33	63.33	60.00	42.75	41.50	39.25	77.52	81.24	80.69	56.02	93.30	91.54	79.26	80.47	87.61	87.51	86.62
ACAN	TNNLS-22	43.83	43.76	43.68	65.42	66.67	66.25	45.75	43.00	40.75	82.48	84.97	84.41	59.25	95.77	96.71	80.16	80.26	88.33	88.58	88.30
FADA	NeurIPS-17	39.10	42.13	32.35	46.25	58.75	47.50	37.25	30.75	35.25	77.24	81.10	77.52	50.43	93.66	93.66	76.95	78.32	88.74	86.09	86.23
d-SNE	CVPR-19	41.58	44.07	38.01	67.08	65.42	61.67	44.50	43.25	41.00	78.76	82.76	83.45	57.63	95.42	94.83	81.11	82.32	89.76	83.51	87.82
SSA²lign	-	**52.13**	**52.21**	**51.75**	**78.33**	**75.42**	**74.58**	**47.75**	**46.75**	**48.25**	**84.69**	**86.48**	**89.66**	**65.67**	**98.59**	**98.24**	**87.26**	**88.11**	**92.97**	**93.02**	**93.03**

Table 1. Results for 10-shot ($k = 10$) FSVDA on Daily-DA and Sports-DA. Note that TSF is trained with *only* source data, while TSF w/T is trained with both source data and k target videos. Same for subsequent tables.

Methods	Publication	Daily-DA													Sports-DA						
		H→A	M→A	KD→A	A→H	M→H	KD→H	H→M	A→M	KD→M	M→KD	H→KD	A→KD	Avg.	KS→U	S→U	U→S	KS→S	U→KS	S→KS	Avg.
TSF	-	37.86	32.58	31.11	44.58	57.08	45.83	36.50	30.00	34.50	61.66	58.90	75.72	45.53	91.66	91.07	76.37	77.74	87.77	85.12	84.95
TSF w/ T	-	40.19	40.03	37.08	60.04	60.04	52.96	34.75	36.00	33.25	79.45	66.21	69.10	50.76	91.89	93.30	78.74	78.32	87.97	86.80	86.17
TRX	CVPR-21	32.79	30.26	28.99	39.43	47.45	40.35	29.00	27.75	24.75	69.55	63.03	55.88	40.77	86.53	86.96	76.34	70.95	81.83	80.41	80.50
STRM	CVPR-22	35.32	32.51	30.98	40.15	47.49	39.30	27.25	26.25	26.50	72.49	62.58	57.78	41.55	91.00	90.02	77.84	73.49	86.38	82.14	83.48
HyRSM	CVPR-22	39.07	35.09	31.06	42.74	52.10	43.74	31.00	29.25	28.00	75.65	63.89	58.60	44.18	92.17	90.77	79.12	75.27	86.25	81.71	84.22
DANN	ICML-15	40.50	38.79	36.39	60.83	58.75	52.92	41.75	38.50	39.75	74.35	66.76	69.66	51.58	92.25	93.07	78.42	75.63	85.12	82.16	84.44
MK-MMD	ICML-15	38.87	43.91	34.60	58.33	57.08	54.58	42.25	35.00	35.50	75.31	68.14	69.38	51.08	92.01	93.18	78.74	75.00	84.51	83.49	84.49
MDD	ICML-19	41.89	42.67	38.40	61.25	62.50	55.42	43.25	40.00	38.50	75.72	68.55	70.90	53.25	92.72	92.60	79.11	79.79	86.54	83.59	85.72
SAVA	ECCV-20	40.96	37.24	38.25	60.00	62.92	55.83	40.75	38.75	35.25	77.79	67.45	68.97	52.01	92.48	92.83	78.05	76.21	83.13	81.70	84.07
ACAN	TNNLS-22	44.45	44.30	41.35	63.33	63.33	56.25	39.00	40.25	37.50	80.55	70.90	73.79	54.58	95.18	96.59	79.95	79.53	88.18	88.38	87.97
FADA	NeurIPS-17	40.75	41.58	30.07	43.03	55.53	41.06	33.00	27.00	32.50	77.79	67.19	64.74	46.19	93.01	93.86	76.04	76.17	87.50	83.38	84.99
d-SNE	CVPR-19	41.99	44.16	35.74	64.93	63.37	56.94	41.25	41.75	39.75	79.45	72.53	73.30	54.60	94.99	94.73	80.98	81.77	89.37	81.75	87.27
SSA²lign	-	**52.37**	**51.98**	**47.40**	**76.67**	**72.92**	**70.42**	**47.00**	**46.25**	**47.50**	**86.76**	**79.31**	**81.79**	**63.36**	**97.06**	**97.89**	**84.05**	**86.21**	**91.18**	**90.21**	**91.10**

Table 2. Results for 5-shot ($k = 5$) FSVDA on Daily-DA and Sports-DA.

Methods	Publication	Daily-DA													Sports-DA						
		H→A	M→A	KD→A	A→H	M→H	KD→H	H→M	A→M	KD→M	M→KD	H→KD	A→KD	Avg.	KS→U	S→U	U→S	KS→S	U→KS	S→KS	Avg.
TSF	-	37.86	32.58	31.11	44.58	57.08	45.83	36.50	30.00	34.50	61.66	58.90	75.72	45.53	91.66	91.07	76.37	77.74	87.77	85.12	84.95
TSF w/ T	-	37.94	34.14	33.05	51.25	58.75	46.67	37.75	34.50	35.25	74.07	60.41	63.72	47.29	91.17	92.83	75.37	76.58	86.60	85.37	84.65
TRX	CVPR-21	24.68	23.33	25.06	32.05	43.34	30.23	28.00	27.50	23.25	66.19	52.09	49.38	35.42	84.90	86.30	71.85	69.97	80.82	79.78	78.94
STRM	CVPR-22	28.04	24.18	26.92	33.85	45.02	30.04	25.75	26.00	25.00	69.14	52.13	50.49	36.38	89.64	89.18	73.72	72.50	85.68	81.63	82.06
HyRSM	CVPR-22	30.94	27.14	26.97	35.67	48.42	34.36	30.00	28.75	26.50	71.71	52.93	50.58	38.66	90.68	90.24	74.73	73.27	84.71	81.05	82.45
DANN	ICML-15	30.57	28.63	34.06	53.75	51.67	42.08	39.75	37.25	33.00	73.66	52.97	64.28	45.14	91.64	91.42	72.90	76.90	84.71	82.55	83.35
MK-MMD	ICML-15	29.40	32.51	31.96	54.58	55.42	44.17	38.50	37.00	33.75	72.00	56.00	63.04	45.69	90.72	91.07	74.68	74.21	85.83	84.05	83.43
MDD	ICML-19	31.65	33.59	34.52	54.17	56.67	47.08	42.25	38.50	34.75	70.62	56.14	59.45	46.62	91.42	92.72	74.37	74.90	82.82	82.93	83.19
SAVA	ECCV-20	31.03	33.44	32.97	50.83	58.33	42.92	40.50	39.75	37.75	72.14	55.72	62.07	46.45	89.31	92.83	73.21	73.84	81.75	80.38	81.89
ACAN	TNNLS-22	38.64	35.61	35.30	55.00	61.67	46.67	38.25	38.75	35.75	76.28	59.31	62.62	48.65	93.75	96.24	75.37	77.05	85.66	87.11	85.86
FADA	NeurIPS-17	33.88	34.14	25.57	35.52	53.23	31.26	32.50	27.00	32.75	73.86	57.15	57.91	41.23	91.35	93.13	71.83	75.04	86.75	82.51	83.44
d-SNE	CVPR-19	36.26	37.86	32.13	59.20	60.91	50.01	40.50	41.00	38.50	76.29	64.30	62.93	49.99	93.90	94.49	76.44	77.11	87.86	81.05	85.14
SSA²lign	-	**44.83**	**46.78**	**45.31**	**68.75**	**70.83**	**62.08**	**46.75**	**46.50**	**45.00**	**79.72**	**65.79**	**71.59**	**57.83**	**96.59**	**97.42**	**80.05**	**80.95**	**88.94**	**89.76**	**88.95**

Table 3. Results for 3-shot ($k = 3$) FSVDA on Daily-DA and Sports-DA.

since it is intuitive that the importance of each target snippet is uneven. We thus propose a snippet attention to weigh the impact of different target snippets on the domain alignment dynamically. Intuitively, a snippet whose output prediction is the most accurate, i.e., whose classification is closest to its given ground truth, should be focused during alignment. A simple yet effective expression of how accurate the snippet's output prediction is would be the inverse of the cross-entropy loss. The snippet attention weights are therefore built upon the inverse of the cross-entropy loss of the snippet, along with a residual connection for more stable optimization, expressed as:

$$w_{jl} = 1 + \frac{1}{l_{ce}(o_{T,jl}, y_{j,T})}. \tag{9}$$

The snippet attention weights are subsequently normalized across the r snippets corresponding to the same target video, expressed as $\overline{w}_{jl} = w_{jl}/\frac{1}{r}\sum_{l'=1}^{r} w_{jl'}$. The normalized snippet attention weight $\overline{w}_{jl}$ is then applied to the target snippet features, forming the weighted target snippet features by $\mathbf{f'}_{T,jl} = \overline{w}_{jl}\mathbf{f}_{T,jl}$, which are then aligned with the source domain through the semantic and statistical snippet alignments by replacing the features $\mathbf{f}_{T,jl}$ with $\mathbf{f'}_{T,jl}$.

SSA²lign. Finally, we sum up our proposed SSA²lign in Algorithm 1. The snippet features, SSA, and snippet attention are leveraged only during training. During testing, target video representations are obtained by uniform sampling across the target testing videos, while the video features and their output predictions are obtained by directly applying the trained feature extractor and classifier to the uniformly sampled target video representations.

4. Experiments

In this section, we evaluate our proposed SSA²lign across two challenging cross-domain action recognition benchmarks: Daily-DA and Sports-DA [54], which cover a wide range of cross-domain scenarios. We present superior results on both benchmarks. Further, ablation studies and analysis of SSA^{2}lign are also presented to justify the design of SSA^{2}lign. *Code is provided in the appendix.*

4.1. Experimental Settings

Daily-DA is a challenging dataset that has been leveraged in prior VUDA works [54, 53, 55]. It covers both normal and low-illumination videos and is constructed from four datasets: ARID (A) [52], HMDB51 (H) [20], Moments-in-Time (M) [26], and Kinetics-600 (KD) [3]. HMDB51, Moments-in-Time, and Kinetics-600 are widely used for action recognition benchmarking, while ARID is a recent dark dataset, with videos shot under adverse illumination. Daily-DA contains 18,949 videos from 8 classes, with 12 cross-domain action recognition tasks. **Sports-DA** is a large-scale cross-domain video dataset, built from UCF101 (U), Sports-1M (S) [18], and Kinetics-600 (KS), with 40,718 videos from 23 action classes, and includes 6 cross-domain action recognition tasks. Refer to prior FSDA/FSVDA works [12, 13, 11], we evaluate SSA^{2}lign on both benchmarks with $k = (3, 5, 10)$ target videos per action class (i.e., 3-shot, 5-shot and 10-shot VDA tasks).

For a fair comparison, all methods examined and experiments conducted in this section adopt the TimeSFormer [64] as the feature extractor, pre-trained on Kinetics-400 [19]. All experiments are implemented with the PyTorch [30] library. We set the length of snippets and the number of snippets per target video via SSA empirically as $m = 8, r = 3$. Hyper-parameters λ_{sem}, λ_{stat} and λ_P are empirically set to 1.0, 1.0, and 0.6 and are fixed. *More specifications on benchmark details and network implementation are provided in the appendix.*

4.2. Overall Results and Comparisons

We compare SSA^{2}lign with state-of-the-art FSDA approaches, and prevailing UDA/VUDA and few-shot action recognition (FSAR) approaches. These methods include: FADA [27], d-SNE [49] designed for image-based FSDA; DANN [10], MK-MMD [25], MDD [62], SAVA [7] and ACAN [51], designed for UDA/VUDA; and TRX [31], STRM [39], and HyRSM [47] proposed for FSAR. To adapt the FSAR approaches for FSVDA, the source domain is used for meta-training and the target domain is used for the meta-testing, while target labels are available for optimizing the cross-entropy loss to adapt UDA/VUDA approaches for FSVDA. We also report the results of the source-only model (denoted as TSF) by applying the model trained with only source data directly to the target data; and the source with few-shot target model (denoted as TSF w/ T) by optimizing only the prediction loss $\mathcal{L}_{pred}$ for training. We report the top-1 accuracy on the target domains, averaged on 5 different settings of available target data randomly selected and each with 5 runs (25 runs in total). Tables 1-3 show comparison of SSA^{2}lign against the above methods.

Results in Tables 1-3 show that the novel SSA^{2}lign achieves the state-of-the-art results on all 18 cross-domain action recognition tasks across both cross-domain benchmarks, outperforming prior UDA/VUDA, FSDA or FSAR approaches by noticeable margins. Notably, SSA^{2}lign outperforms all prior FSDA approaches originally designed for image-based FSDA (i.e., FADA and d-SNE) consistently on all tasks, by a relative average of 13% over the second-best performances on Daily-DA (across 3 k-shot settings and 12 tasks), and a relative average of 4.2% on Sports-DA (across 3 k-shot settings and 6 tasks). The consistent improvements justify empirically the effectiveness of augmenting and aligning both semantic information and statistical distribution at the snippet level for FSVDA.

It is also observed that prior FSDA and UDA/VUDA methods could not perform well on FSVDA tasks. Notably, even when $k = 10$ target videos are available per class, all but one of the evaluated FSDA and UDA/VUDA approaches result in performances inferior to that trained with only $\mathcal{L}_{pred}$ without any adaptation (i.e., TSF w/ T). Prior FSDA approaches do not incorporate temporal features and their related semantic information, which are crucial for tackling FSVDA, while UDA/VUDA methods are not effective when target information is not fully available. Negative improvements are more severe when k decreases. It is also noted that at small k values (e.g., $k = 3$), the performance of TSF w/ T could be inferior to that trained without target data (i.e., TSF). This suggests that the few target data could be outliers of the target domain, whose distribution differs greatly from the other target data, resulting in a severe negative impact. Prior FSAR approaches could not tackle FSVDA as well, producing even poorer results than all UDA/VUDA approaches examined. This can be caused by domain shift that exists between data for the meta-training and meta-testing. Feature extractors trained via meta-training on the source domain could not be simply applied to the meta-testing phase on the target domain.

4.3. Ablation Studies, Analysis, and Discussion

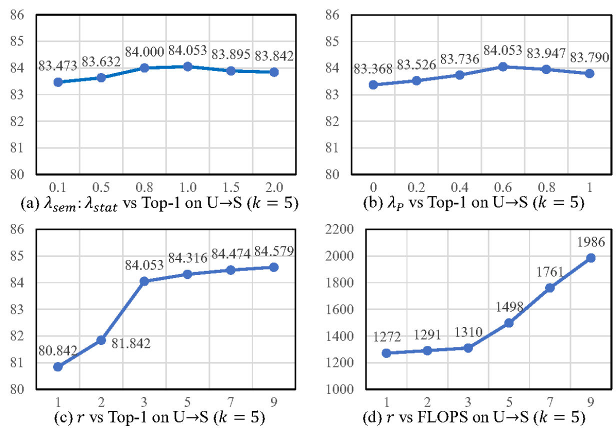
To gain a comprehensive understanding of SSA^{2}lign and justify its design, we perform extensive ablation studies as in Tables 4-5. The ablation studies explore the effects brought by its components, namely the semantic and statistical alignments, the SSA, and the snippet attention. It further validates the alignment details by assessing against 5 variants: SSA^{2}lign-CORAL and SSA^{2}lign-MMD formulate $\mathcal{L}_{sn-stat}$ as CORAL [36] and MDD [62]; SSA^{2}lign-FC computes $\mathcal{L}_{cross}$ over all $r \times (r-1)$ snippet pairs for the same target video; SSA^{2}lign-SP minimizes the distance between target snippet features and source class prototypes for $\mathcal{L}_{proto}$; SSAlign (w/ spatial aug.) augments target domain

Methods	Components						Daily-DA									Sports-DA						Avg.
							$k=10$			$k=5$			$k=3$			$k=10$		$k=5$		$k=3$		
	SSA	Sn-Attn	$\mathcal{L}_{proto}$	$\mathcal{L}_{cross}$	$\mathcal{L}_{sn-dist}$	$\mathcal{L}_{sn-stat}$	H→A	M→A	KD→A	H→A	M→A	KD→A	H→A	M→A	KD→A	U→S	KS→S	U→S	KS→S	U→S	KS→S	
TSF w/ T							39.57	39.49	39.41	40.19	40.03	37.08	37.94	34.14	33.05	78.95	79.05	78.74	78.32	75.37	76.58	53.86
	✓						41.66	41.97	42.05	42.82	42.28	38.25	38.79	36.39	35.22	81.05	81.21	80.79	80.42	75.89	77.21	55.73
SSA²lign		✓	✓	✓	✓	✓	45.62	46.32	45.70	46.47	45.69	41.74	39.17	40.96	39.49	84.32	85.47	81.05	83.58	77.26	77.37	58.68
	✓		✓	✓	✓		51.05	51.13	50.43	50.97	50.73	46.39	43.59	45.85	44.22	86.58	87.37	83.21	85.63	79.32	80.37	62.46
	✓	✓	✓				49.88	49.81	49.73	50.35	49.26	45.07	42.19	44.53	42.90	85.58	86.74	82.74	84.74	78.69	79.47	61.45
	✓	✓		✓			50.19	50.50	49.65	50.27	49.88	45.77	42.66	45.15	43.06	85.84	87.05	82.58	85.05	78.95	79.58	61.75
	✓	✓			✓		48.18	48.57	46.86	47.56	48.09	43.29	40.64	42.51	41.97	84.68	85.95	81.95	83.74	77.74	78.63	60.03
	✓	✓	✓	✓	✓		51.36	51.44	50.89	51.82	51.43	46.55	43.90	46.00	44.92	86.68	87.63	83.63	85.69	79.69	80.42	62.80
	✓	✓	✓	✓	✓	✓	**52.13**	**52.21**	**51.75**	**52.37**	**51.98**	**47.40**	**44.83**	**46.78**	**45.31**	**87.26**	**88.11**	**84.05**	**86.21**	**80.05**	**80.95**	**63.43**

Table 4. Ablation studies of the components of SSA²lign on 5 cross-domain tasks over Daily-DA and Sports-DA.

Methods	Daily-DA									Sports-DA						Avg.	Δ Avg.	GFLOPS	Δ GFLOPS
	$k=10$			$k=5$			$k=3$			$k=10$		$k=5$		$k=3$					
	H→A	M→A	KD→A	H→A	M→A	KD→A	H→A	M→A	KD→A	U→S	KS→S	U→S	KS→S	U→S	KS→S				
SSA²lign-CORAL	51.90	51.98	51.51	51.98	51.58	47.09	44.52	46.39	45.07	87.05	87.84	83.90	86.00	79.90	80.68	63.16	-0.27	1302	-8
SSA²lign-MMD	51.67	51.90	51.28	51.98	51.51	47.01	44.52	46.32	44.84	87.05	87.84	83.84	85.90	79.79	80.63	63.07	-0.36	1312	+2
SSA²lign-FC	52.83	52.91	52.37	52.99	52.36	47.56	45.30	47.25	45.31	87.16	88.26	84.53	86.69	79.79	81.26	63.78	+0.35	1472	+162
SSA²lign-SP	50.97	51.20	50.97	51.28	51.27	46.47	43.82	45.62	44.69	86.90	87.42	83.37	85.74	79.26	80.47	62.63	-0.80	1390	+80
SSAlign (w/ spatial aug.)	45.38	46.70	45.23	45.77	45.84	41.12	39.63	40.26	38.64	83.53	85.63	80.58	83.37	76.95	77.74	58.43	-5.00	1325	+15
SSA²lign	52.13	52.21	51.75	52.37	51.98	47.40	44.83	46.78	45.31	87.26	88.11	84.05	86.21	80.05	80.95	63.43	-	1310	-

Table 5. Ablation studies of the alignment details of SSA²lign on 5 cross-domain tasks over Daily-DA and Sports-DA.



(a) $\lambda_{sem}:\lambda_{stat}$ vs Top-1 on U→S ($k=5$) (b) λ_P vs Top-1 on U→S ($k=5$)

(c) r vs Top-1 on U→S ($k=5$) (d) r vs FLOPS on U→S ($k=5$)

Figure 2. Sensitivity of hyper-parameters on U→S task.

through random spatial augmentation across the frames of r snippets. The ablation studies are conducted on 5 tasks over Daily-DA and Sports-DA. If SSA is not applied, we sample r snippets sequentially from the 1^{st} frame of each target video and remain unchanged during training.

Semantic Alignment. As shown in Table 4, with only snippet-level semantic alignment (whether in full or any one of the three perspectives), the performance still surpasses all previous FSDA and UDA/VUDA methods compared. This conforms to our motivation that applying semantic alignment could tackle FSVDA more effectively. Moreover, statistical alignment and snippet attention further improve SSA²lign, but only by a marginal degree.

Superiority of SSA. Notably, a significant performance drop is observed when SSA is not applied, which proves the importance of expanding target domain data through SSA for subsequent alignment. The importance of SSA is further verified when we apply SSA for training with augmented snippets but without adaptation which shows a noticeable gain compared to the original TSF w/ T. Further, the significantly inferior performance of SSAlign (w/ spatial aug.) as shown in Table 5 conforms with the motivation of SSA, which aims for more effective target video domain augmentation while spatial augmentation may undermine temporal correlation across sequential frames.

Methods	Sports-DA						Avg.
	$k=10$		$k=5$		$k=3$		
	U→S	KS→S	U→S	KS→S	U→S	KS→S	
ACAN+MixUp	80.63	80.89	80.42	79.79	75.95	77.37	79.18
ACAN+RandAugment	78.68	79.84	78.26	80.05	75.74	76.58	78.19
ACAN+TrivialAugment	81.53	81.05	80.53	80.68	76.84	77.63	79.71
ACAN+MixUp+TrivialAugment	82.16	81.95	81.21	81.58	78.05	78.32	80.55
SSA²lign	**87.26**	**88.11**	**84.05**	**86.21**	**80.05**	**80.95**	**84.44**

Table 6. Compare with ACAN with up-to-date augmentations.

While the success of SSA²lign is built upon augmenting target videos with SSA, there have been more complex augmentation practices introduced for images, such as MixUp [61], RandAugment [8], and TrivialAugment [28]. To further prove the efficacy of SSA for FSVDA, we compare SSA2lign against a competitive VUDA method ACAN [51] with the aforementioned up-to-date augmentation practices applied to the target domain on Sports-DA, as shown in Table 6. Note that TimeSFormer [64] is leveraged in ACAN as the feature extractor for fair comparison. Results in Table 6 show that while there are improvements by applying augmentations to the target domain, SSA²lign with simple augmentation still outperforms multiple augmentations due to its snippet-level alignment, which further validates the superiority of SSA.

Alignment Methods. Table 5 shows that while formulating $\mathcal{L}_{sn-stat}$ as MDD [62] brings the best performance, selecting other metrics brings negligible impact. Further, computing $\mathcal{L}_{cross}$ with all target snippet pairs only brings trivial performance gain at a cost of significant computation over-

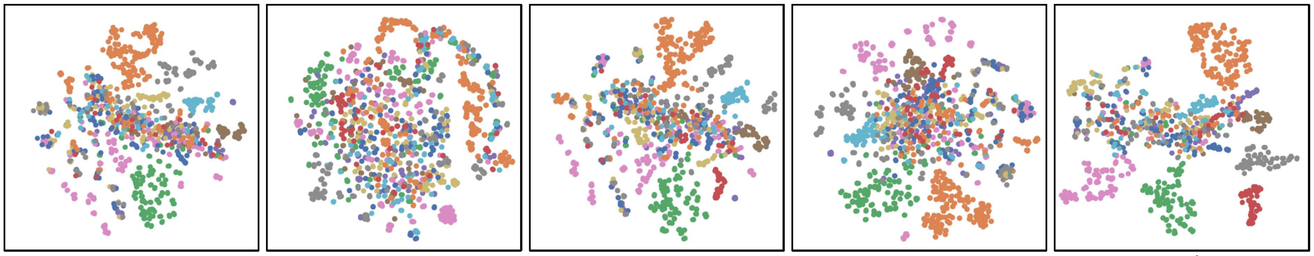


Figure 3. t-SNE visualizations of target features from (a) TSF w/T, (b) HyRSM, (c) MDD, (d) SSA^{2}lign. Colors denote classes.

head (12% computation increase for 0.54% gain). Further, matching target snippet features to source class prototypes for $\mathcal{L}_{proto}$ results in a performance drop with more computation. The inferior performance could be due to outliers in the source domain which could affect source class prototypes, bringing in source noise that should not be aligned.

Hyper-parameter Sensitivity. We focus on studying the sensitivity of λ_{sem} and λ_{stat} which control the strength of the semantic and statistical snippet alignment losses, λ_P which relates to the update of target prototypes and r the number of snippets per target video. Without loss of generality, we fix $\lambda_{stat} = 1.0$ and study the ratio $\lambda_{sem} : \lambda_{stat}$ in the range of 0.1 to 1.5. λ_P is in the range of 0 to 1 which corresponds to using only the initial prototypes or the updated prototypes, and r is in the range of 1 to 9. As shown in Fig. 2, SSA^{2}lign is robust to ratio $\lambda_{sem} : \lambda_{stat}$ and λ_P, falling within a margin of 0.683%, with the best results obtained at the current default where $\lambda_{sem} : \lambda_{stat} = 1.0$ and $\lambda_P = 0.6$. SSA^{2}lign is also robust to r when $r \geqslant 3$, i.e., when there are multiple snippets obtained via SSA per target video. $r = 3$ is selected as significant computation overhead would occur for $r > 3$ with marginal gain. Notably, SSA^{2}lign cannot perform when $r < 3$, especially when $r = 1$ where the $\mathcal{L}_{cross}$ does not work and the target domain is not expanded.

Feature Visualization. We further understand the characteristics of SSA^{2}lign by plotting the t-SNE embeddings [40] of target features with class information from the model trained without adaptation (TSF w/T), HyRSM, MDD, ACAN and SSA^{2}lign for U$\to$S with $k = 10$ in Fig. 3. It is observed that target features from SSA^{2}lign are more clustered and discriminable, corresponding to better performance. Such observation intuitively proves that video domain adaptation can be improved when feature extractors possess stronger discriminability. However, SSA^{2}lign is not designed to deal explicitly with classes that could be similar spatially or temporally, thus certain features observe lower discriminability, which denotes future work.

Limitations in Choice of Datasets. The datasets Sports-DA and Daily-DA [54] are leveraged as our benchmarks as they have been commonly used in the VUDA community in prior VUDA works [53, 54, 55]. However, it is noted that benchmarks that have been used for action recognition with stronger temporal reasoning assessment and fine-grained action classes (such as Something-Something V2 [15]) are not included in any current cross-domain video datasets as there is few datasets that offer overlapped fine-grained action classes. The current results show that the FSVDA is still a challenging task in our proposed benchmarks, we believe that exploring how to adapt models from coarse-category datasets to fine-grained datasets (SSv2) denotes future exploration.

5. Conclusion

In this work, we propose a novel SSA^{2}lign to tackle the challenging yet realistic Few-Shot Video Domain Adaptation (FSVDA), where only limited labeled target data are available. Without sufficient target data, SSA^{2}lign tackles FSVDA at the snippet level via a simple SSA augmentation and performing the semantic and statistical alignments attentively, where the semantic alignment is further achieved from three perspectives based on semantic information within and across snippets. Extensive experiments and detailed ablation studies across cross-domain action recognition benchmarks validate the superiority of SSA^{2}lign in addressing FSVDA.

Acknowledgements

This research is jointly supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-RP-2021-027), and A*STAR Singapore under its Career Development Award (Grant No. C210112046), and Nanyang Technological University, Singapore, under its NTU Presidential Postdoctoral Fellowship, “Adaptive Multimodal Learning for Robust Sensing and Recognition in Smart Cities” project fund.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video

vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[3] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 7

[4] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6321–6330, 2019. 1, 2, 3, 5

[5] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020. 2

[6] Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *International conference on machine learning*, pages 1081–1090. PMLR, 2019. 2, 5

[7] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 678–695. Springer, 2020. 1, 2, 5, 7

[8] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 8

[9] Carlotta Domeniconi, Dimitrios Gunopulos, and Jing Peng. Large margin nearest neighbor classifiers. *IEEE transactions on neural networks*, 16(4):899–909, 2005. 2

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 7

[11] Yipeng Gao, Lingxiao Yang, Yunmu Huang, Song Xie, Shiyong Li, and Wei-Shi Zheng. Acrofod: An adaptive method for cross-domain few-shot object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 673–690. Springer, 2022. 1, 2, 7

[12] Zan Gao, Leming Guo, Weili Guan, An-An Liu, Tongwei Ren, and Shengyong Chen. A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-r2. *IEEE Transactions on Image Processing*, 30:767–782, 2020. 1, 2, 7

[13] Zan Gao, Leming Guo, Tongwei Ren, An-An Liu, Zhi-Yong Cheng, and Shengyong Chen. Pairwise two-stream convnets for cross-domain action recognition with small data. *IEEE Transactions on Neural Networks and Learning Systems*, 33(3):1147–1161, 2020. 1, 2, 7

[14] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *European conference on computer vision*, pages 597–613. Springer, 2016. 2

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 9

[16] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002. 2

[17] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and J Doug Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 43–58, 2011. 2

[18] 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. 7

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 7

[20] 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. 7

[21] Jogendra Nath Kundu, Akshay R Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Anand Kulkarni, Varun Jampani, and Venkatesh Babu Radhakrishnan. Balancing discriminability and transferability for source-free domain adaptation. In *International Conference on Machine Learning*, pages 11710–11728. PMLR, 2022. 2, 5

[22] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 5

[23] Jinlu Liu, Liang Song, and Yongqiang Qin. Prototype rectification for few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 741–756. Springer, 2020. 4

[24] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[25] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2, 5, 7

[26] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments

in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 7

[27] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. *Advances in neural information processing systems*, 30, 2017. 1, 2, 7

[28] Samuel G Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 774–782, 2021. 8

[29] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *AAAI*, pages 11815–11822, 2020. 2

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 7

[31] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 475–484, 2021. 7

[32] Abdel Mlak Said, Aymen Yahyaoui, and Takoua Abdellatif. Efficient anomaly detection for smart hospital iot systems. *Sensors*, 21(4):1026, 2021. 1

[33] 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. 2

[34] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 4

[35] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6479–6488, 2018. 1

[36] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 2, 5, 7

[37] Guangyu Sun, Zhang Liu, Lianggong Wen, Jing Shi, and Chenliang Xu. Anomaly crossing: New horizons for video anomaly detection as cross-domain few-shot learning. *arXiv e-prints*, pages arXiv–2112, 2021. 2

[38] Takeshi Teshima, Issei Sato, and Masashi Sugiyama. Few-shot domain adaptation by causal mechanism transfer. In *International Conference on Machine Learning*, pages 9458–9469. PMLR, 2020. 2

[39] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19958–19967, 2022. 7

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

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[42] Vikas Verma, Alex Lamb, Juho Kannala, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. In *International Joint Conference on Artificial Intelligence*, pages 3635–3641, 2019. 5

[43] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019. 4

[44] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 3

[45] Tao Wang, Xiaopeng Zhang, Li Yuan, and Jiashi Feng. Few-shot adaptive faster r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7173–7182, 2019. 1

[46] Xiyu Wang, Yuecong Xu, Jianfei Yang, and Kezhi Mao. Calibrating class weights with multi-modal information for partial video domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3945–3954, 2022. 1

[47] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19948–19957, 2022. 7

[48] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1529–1538, 2017. 3

[49] Xiang Xu, Xiong Zhou, Ragav Venkatesan, Gurumurthy Swaminathan, and Orchid Majumder. d-sne: Domain adaptation using stochastic neighborhood embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2506, 2019. 1, 2, 7

[50] Yuecong Xu, Haozhi Cao, Zhenghua Chen, Xiaoli Li, Lihua Xie, and Jianfei Yan. Video unsupervised domain adaptation with deep learning: A comprehensive survey. *arXiv preprint arXiv:2211.10412*, 2022. 2

[51] Yuecong Xu, Haozhi Cao, Kezhi Mao, Zhenghua Chen, Lihua Xie, and Jianfei Yang. Aligning correlation information for domain adaptation in action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 1, 2, 3, 5, 7, 8

[52] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Arid: A new dataset for recognizing action in the dark. In *International Workshop on*

Deep Learning for Human Activity Recognition, pages 70–84. Springer, 2021. 7

[53] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, and Zhenghua Chen. Source-free video domain adaptation by learning temporal consistency for action recognition. In *European Conference on Computer Vision*, pages 147–164. Springer, 2022. 1, 2, 4, 5, 7, 9

[54] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, Zhengguo Li, and Zhenghua Chen. Multi-source video domain adaptation with temporal attentive moment alignment network. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1, 2, 6, 7, 9

[55] Yuecong Xu, Jianfei Yang, Min Wu, Xiaoli Li, Lihua Xie, and Zhenghua Chen. Extern: Leveraging endo-temporal regularization for black-box video domain adaptation. *arXiv preprint arXiv:2208.05187*, 2022. 5, 7, 9

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

[57] Jianfei Yang, Yuecong Xu, Haozhi Cao, Han Zou, and Lihua Xie. Deep learning and transfer learning for device-free human activity recognition: A survey. *Journal of Automation and Intelligence*, 1(1):100007, 2022. 2

[58] Jianfei Yang, Jiangang Yang, Shizheng Wang, Shuxin Cao, Han Zou, and Lihua Xie. Advancing imbalanced domain adaptation: Cluster-level discrepancy minimization with a comprehensive benchmark. *IEEE Transactions on Cybernetics*, 0:1 – 12, 2021. 2

[59] Jianfei Yang, Han Zou, Yuxun Zhou, Zhaoyang Zeng, and Lihua Xie. Mind the discriminability: Asymmetric adversarial domain adaptation. In *European Conference on Computer Vision*, pages 589–606. Springer, 2020. 2, 5

[60] Xiangyu Yue, Zangwei Zheng, Hari Prasanna Das, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Multi-source few-shot domain adaptation. *arXiv preprint arXiv:2109.12391*, 2021. 2

[61] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations*, 2018. 5, 8

[62] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. 2, 5, 7, 8

[63] An Zhao, Mingyu Ding, Zhiwu Lu, Tao Xiang, Yulei Niu, Jiechao Guan, and Ji-Rong Wen. Domain-adaptive few-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1390–1399, 2021. 2

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

[65] Xiatian Zhu, Antoine Toisoul, Juan-Manuel Perez-Rua, Li Zhang, Brais Martinez, and Tao Xiang. Few-shot action recognition with prototype-centered attentive learning. *arXiv preprint arXiv:2101.08085*, 2021. 4