Source-Free Video Domain Adaptation with Spatial-Temporal-Historical Consistency Learning

Kai Li, Deep Patel, Erik Kruus, Martin Renqiang Min
NEC Labs, America
{kaili,dpatel,kruus,renqiang}@nec-labs.com

Abstract

Source-free domain adaptation (SFDA) is an emerging research topic that studies how to adapt a pretrained source model using unlabeled target data. It is derived from unsupervised domain adaptation but has the advantage of not requiring labeled source data to learn adaptive models. This makes it particularly useful in real-world applications where access to source data is restricted. While there has been some SFDA work for images, little attention has been paid to videos. Naively extending image-based methods to videos without considering the unique properties of videos often leads to unsatisfactory results. In this paper, we propose a simple and highly flexible method for Source-Free Video Domain Adaptation (SFVDA), which extensively exploits consistency learning for videos from spatial, temporal, and historical perspectives. Our method is based on the assumption that videos of the same action category are drawn from the same low-dimensional space, regardless of the spatio-temporal variations in the high-dimensional space that cause domain shifts. To overcome domain shifts, we simulate spatio-temporal variations by applying spatial and temporal augmentations on target videos and encourage the model to make consistent predictions from a video and its augmented versions. Due to the simple design, our method can be applied to various SFVDA settings, and experiments show that our method achieves state-of-the-art performance for all the settings.

1. Introduction

Action recognition is a crucial task in video understanding and has been receiving tremendous attention from the vision community. In recent years, it has made significant progress, primarily due to the development of deep learning techniques [11, 43, 45] and the establishment of large-scale annotated datasets [2, 13, 42]. However, it is acknowledged that an action recognition model trained with annotated data drawn from one distribution typically experiences a performance drop when tested on out-of-distribution data [4]. This is the so-called domain shift problem.

To tackle this problem, Unsupervised Video Domain Adaptation (UVDA) has been proposed. The goal is to learn an adaptive model using labeled video data from one domain (source) and unlabeled video data from another domain (target). Typical UVDA methods use videos from both domains as input and train a model by minimizing the classification risk on labeled source videos and explicitly aligning videos from both domains in a class-agnostic fashion. Although most image-based domain alignment techniques can be applied to video domain alignment, such as adversarial learning [22, 37, 44], methods that align domains by considering the richer temporal information in videos have shown superior performance [6, 33, 36].

While UVDA methods help alleviate the domain shift problem, their assumption that labeled source videos are available for domain alignment can be problematic in real-world applications where access to source videos is restricted due to privacy or commercial reasons [24, 50]. This motivates a new research topic, Source-Free Video Domain Adaptation (SFVDA) [50], which aims to learn an adaptive action recognition model using unlabeled target videos and a source model pre-trained with labeled source videos. SFVDA is similar to UVDA in learning an adaptive model using labeled source and unlabeled target videos but differs in that labeled source videos are only used for learning the source model. Adaptation only involves target videos, which avoids leaking annotated source videos. However, the absence of labeled source videos makes SFVDA a more challenging problem than UVDA since there is no reliable supervision signal, and no data drawn from the distribution to be aligned, which makes it even more challenging.

Very recently, Xu et al. [50] proposed a pioneering approach to SFVDA based on temporal consistency. They adapt the source model by encouraging it to keep the capability of understanding motion dynamics despite domain shifts. They train the model to produce features/predictions for a video clip consistent with those of other clips within the same video or that of the entire video. Despite im-

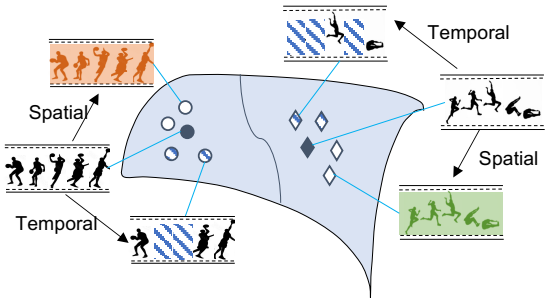

Figure 1. Conceptual illustration of applying spatial and temporal augmentations to simulate domain shifts and encouraging prediction consistency for SFVDA.

proved performance over baseline methods, this method only considers adapting the source model from a temporal perspective and ignores spatial factors (the appearance of frames) that also account for domain shifts. Adapting the model without encouraging it to surpass the visual appearance variations could still produce sub-optimal adaptation results. Besides, clips from the same video often share high similarity, and the model can produce consistent features/predictions even though it has not been well adapted.

In this paper, we propose a novel SFVDA method that overcomes the limitation of the existing methods by exploiting Spatial-Temporal-Historical Consistency (STHC). Our underlying assumption is that *videos of the same action category are drawn from the same low-dimensional space, regardless of spatio-temporal variations in the high-dimensional space that cause domain shifts*. To achieve this, we simulate spatio-temporal variations with target videos and adapt the source model by encouraging it to surpass the variations and produce consistent predictions. Specifically, we apply spatial and temporal augmentations to each unlabeled target video in a *stochastic* manner to simulate spatio-temporal variations. By encouraging consistent classification predictions for the video and its augmented versions, we ensure that they are drawn from the same low-dimensional space. After adapting the model in this way, it is expected to generalize well on test videos that fall into the same low-dimensional space as the training videos. Figure 1 provides an illustration of this concept.

More concretely, we randomly select a clip from the video and apply stochastic frame-wise spatial augmentation, resulting in a perturbed version of the clip. In addition, we also apply stochastic temporal augmentation by randomly masking some frames to generate a temporally-perturbed clip. To ensure prediction consistency, we enforce the spatial consistency (SC) of the clip with its perturbed version and the temporal consistency (TC) of the clip with its temporally-perturbed version. Besides these two techniques, we propose a third technique that enforces consistent predictions for the clip and other clips from the same video. This technique is similar to that in [50], but we implement this in a nearly no-cost way: We store historical predictions of all the clips (with randomly sampled frames) from each video in a memory bank and retrieve predictions from the bank to enforce prediction consistency for the current clip. This technique reinforces temporal consistency and we call it *historical consistency (HC)*. Notably, TC and SC produce "hard" versions of a clip and encourage the model to overcome the hard factors and make consistent predictions. Therefore, the model must have a strong understanding of the target domain to fulfill these tasks, facilitating model adaptation.

Thanks to simplicity in design, our STHC method can be easily extended to other SFVDA settings, including the open-set setting where the target domain contains classes that are absent in the source domain, the partial setting where the source domain contains classes that are absent in the target domain, and the black-box setting where only outputs of the source model are available and the model weights are not accessible. Experiments show that STHC outperforms existing methods for all the SFVDA settings. Our contributions can be summarized as follows:

- We comprehensively exploit consistency learning for videos and propose STHC model for SFVDA. STHC performs stochastic spatio-temporal augmentations on each video and enforces prediction consistency from spatial, temporal, and historical perspectives.
- We extend STHC to address various domain adaptation problems under the SFVDA setting. To our best knowledge, most of these problems have not been studied before and we establish the evaluation benchmarks that will help future development.
- STHC achieves state-of-the-art performance for SFVDA in various problem settings.

2. Related Work

2.1. Video Domain Adaptation

While image-based domain adaptation has been extensively investigated [20, 21, 28, 37, 44], video domain adaptation (VDA) has only recently been explored. A straightforward approach to VDA is to extend image-based methods to videos by applying adaptation techniques to the feature representation level. However, this naive extension often yields unsatisfactory results due to the failure of modeling the temporal information [4–6, 8, 29, 33, 36]. Various techniques have been proposed to address this issue. Some methods use attention mechanisms to perform alignment in the temporal direction, either by modeling the temporal relationship [4] or highlighting common key frameworks [33]. Some methods employ self-supervised learning techniques by learning pretext tasks with videos from both domains, e.g., clip order prediction [6], contrastive learn-

ing [5, 36], etc. Other methods extend adversarial learning to the temporal direction [5] or leverage frame graph to bridge the domain gap between video datasets [29]. While most existing VDA methods study the unsupervised setting where adaptation is from one labeled source dataset to another unlabeled target dataset, some works also investigate other adaptation scenarios, including the multi-modality setting [14,32,41,51] where both RGB and motion information are available for domain alignment, the partial domain adaptation setting [48] where action categories in the target domain is a subset of those of the source domain, and the source-free setting [50] which is studied in this paper. Due to the absence of source data for adaptation, most existing video domain alignment techniques are not applicable to SFVDA. [50] addresses this problem with temporal consistency learning. We differ from it in that we also consider spatial consistency. Additionally, [50] enforces consistency between video clips and other clips or the whole video, we use spatio-temporal augmentation to encourage consistency among different augmented versions.

2.2. Source-Free Domain Adaptation

The absence of source data for adaptation renders most mature domain alignment techniques impractical for Source-Free Domain Adaptation (SFDA). To work around this issue, several SFDA methods propose to generate "proxy" source samples, such as generating images with a learned image generation model [19], producing class prototypes by a conditional feature generator [34], or generating features with a parametric distribution estimation model [35]. Other methods adapt the model without relying on existing unsupervised domain alignment techniques. Liang et al. suggest adapting the feature extractor with target data and encouraging it to produce feature representations that result in certain predictions by the classifier [24]. Xia et al. propose a dual-classifier model to achieve contrastive category-wise matching and adversarial domain-level alignment [46]. Yang et al. address this problem by ensuring similar samples are assigned the same labels, either for semantically similar samples [53] or for spatially similar samples in the feature space [52]. Some works also study variant problems of SFDA, for example, the multi-source adaptation variant [1] and universal adaptation variant [18]. Some works also study the SFDA problem for other tasks beyond image classification, e.g., semantic segmentation [17, 27] and object detection [23]. Our method studies the SFDA problem for videos and explores both spatial and temporal information for model adaptation, and is thus essentially different from these image-based methods.

3. Algorithm

Source-Free Video Domain Adaptation (SFVDA) is derived from unsupervised video domain adaptation (UVDA): It shares the same goal as UVDA of learning an adaptive video classification model H using a source dataset $\mathcal{S} = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \dots, (\mathbf{X}_M, y_M)\}$ of M labeled videos and a target dataset $\mathcal{T} = \{\mathbf{U}_1, \mathbf{U}_2, \dots, \mathbf{U}_N\}$ of N unlabeled videos. The difference lies in how they access these datasets. UVDA learns model H *simultaneously* with access to both $\mathcal{S}$ and $\mathcal{T}$, whereas SFVDA learns model H in two steps, first on $\mathcal{S}$ in a standard supervised manner, and then on $\mathcal{T}$ by adapting the learned model in an unsupervised manner. While this process is more complicated, separating $\mathcal{S}$ from adaptation prevents the labeled data from being disclosed, enabling a privacy-safe adaptation solution.

Source model generation. To generate the source model, we follow the same learning protocols as the existing SFVDA method [50]. We decouple model H as $H = F \circ G \circ C$ where F, G, and C are the frame feature extractor, the temporal information encoder, and the classifier, respectively. For each video $(\mathbf{X}, y) \in \mathcal{S}$, we perform segment-wise frame sampling to obtain clips. Specifically, let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ be the frames. We divide the frames evenly into K segments and sample one frame from each segment. This produces a K-frame video snippet of class y, i.e., $(\bar{\mathbf{X}}, y)$ where $|\bar{\mathbf{X}}| = K$. We input $\bar{\mathbf{X}}$ to F to produce a sequence of frame features, which are then sent to G to encode the temporal information. After that, we perform average pooling of all frame features (outputted by G) and get a vector representation for $\bar{\mathbf{X}}$, which we send to C for calculating the cross-entropy loss with y. H is updated accordingly with the standard gradient back-propagation.

3.1. Spatial-Temporal-Historical Consistency

Given model H with weights learned from $\mathcal{S}$, we propose the STHC model to adapt it to the target domain using unlabeled data $\mathcal{T}$. We achieve this by encouraging H to surpass spatio-temporal variations that account for domain shifts. To simulate spatio-temporal variations, we apply stochastic augmentations and generate spatially-augmented and temporally-augmented views for each target video. We then train H to encourage it to make consistent predictions for the video and its augmented versions from spatial, temporal, and historical aspects. Figure 2 shows our framework. In the following sections, we will delve into the specifics of the three consistency learning techniques.

3.1.1 Spatial Consistency

For an unlabeled target video $\mathbf{U} \in \mathcal{T}$ with n frames, we perform segment-wise random sampling which evenly divides the n frames into K parts, and one frame is randomly selected from each part, producing a K-frame sequence $\bar{\mathbf{U}} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_K\}$. We apply per-frame data augmentation and get $\bar{\mathbf{U}}_s$ as

$$\bar{\mathbf{U}}_s = \{\psi(\mathbf{u}_1), \psi(\mathbf{u}_2), \dots, \psi(\mathbf{u}_K)\}, \tag{1}$$

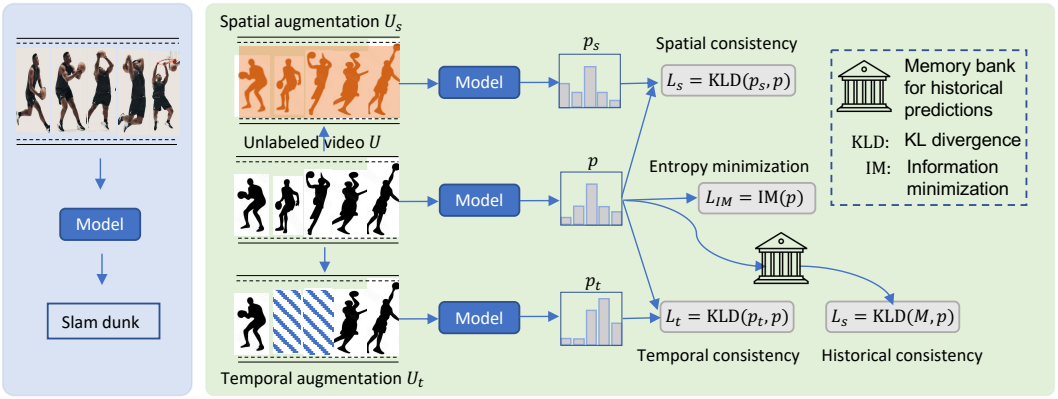

Figure 2. Framework of the proposed STHC method. It includes a pretraining stage that learns a source model using labeled source videos (Left) and an adaptation stage that adapts the source model using unlabeled target videos (Right). During the adaptation stage, for each unlabeled target video, we apply spatial and temporal augmentations and encourage consistent predictions with three types of consistency learning techniques, e.g., spatial consistency, temporal consistency, and historical consistency. The entropy minimization loss is used to encourage the model to make individually uncertain and globally diverse predictions.

where ψ is a stochastic function, which implies that different augmented versions can be obtained when ψ is applied at different time. This is important because a potentially infinite number of augmented videos can be generated, and all these videos should lie around $\bar{\mathbf{U}}$ in the low-dimensional space. Inspired by recent semi-supervised learning methods [40, 47], we adopt RandomAugment [7] as the stochastic function ψ which randomly selects image transformations out of a pool (including color inversion, translation, contrast adjustment, etc.) and apply them on each frame. After that, CutOut [9] is applied which sets a random square patch of pixels to gray.

As mentioned earlier, $\bar{\mathbf{U}}_s$ is supposed to be spatially adjacent with $\bar{\mathbf{U}}_s$ in the low-dimensional manifold space, as it is generated from $\bar{\mathbf{U}}_s$ with frame content perturbed; the semantic category information should be preserved. We enforce this spatial proximity constraint by minimizing the discrepancy of the predictions of $\bar{\mathbf{U}}$ and $\bar{\mathbf{U}}_s$ as

$$\mathscr{L}_s = \mathrm{KLD}\Big(H(\bar{\mathbf{U}}_s), H(\bar{\mathbf{U}})\Big), \tag{2}$$

where KLD represents the KL-divergence. Since $\bar{\mathbf{U}}_s$ can be viewed as a harder version of $\bar{\mathbf{U}}$ with the content being perturbed, enforcing prediction consistency between them encourages the model to surpass spatial factors that account for domain shifts and thus facilitates model adaptation.

3.1.2 Temporal Consistency

Similarly, we generate temporally augmented videos and enforce prediction consistency to learn the manifold structure in the temporal direction. For the sampled frame sequence $\bar{\mathbf{U}} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_K\}$, we apply stochastic temporal augmentation and get $\bar{\mathbf{U}}_t$ as

$$\bar{\mathbf{U}}_t = \{\phi(\mathbf{u}_1), \phi(\mathbf{u}_2), \ldots, \phi(\mathbf{u}_K)\}, \tag{3}$$

where ϕ denotes a stochastic function which drops $\mathbf{u}_k$ out of the sequence at a rate 0.5. This makes $\bar{\mathbf{U}}_t$ a sparser version of $\bar{\mathbf{U}}$. Despite with fewer frames, we expect $\bar{\mathbf{U}}_t$ still preserves the motion dynamics as in $\bar{\mathbf{U}}$ and enforce prediction consistency between them as,

$$\mathscr{L}_t = \mathrm{KLD}\Big((H(\bar{\mathbf{U}}_t), H(\bar{\mathbf{U}})\Big). \tag{4}$$

Since every frame from $\bar{\mathbf{U}}$ has a random possibility of being dropped, we can thus produce numerous augmented versions of $\bar{\mathbf{U}}_t$. All the augmented versions should be adjacent with $\bar{\mathbf{U}}$ in the low-dimensional manifold space, which helps learn the neighboring manifold structure of $\bar{\mathbf{U}}$. As $\bar{\mathbf{U}}_t$ can be viewed as the harder version of $\bar{\mathbf{U}}$ with different motion dynamics, enforcing prediction consistency between them thus encourages the model to surpass the motion dynamics variations between domains, and facilitate model adaptation.

3.1.3 Historical Consistency

The historical consistency is proposed to reinforce the temporal consistency. Recall that $\bar{\mathbf{U}} = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k\}$ is randomly sampled from K evenly divided video segments. We generate $\bar{\mathbf{U}}_t$ to encourage temporal consistency by masking some frames from $\bar{\mathbf{U}}$ to make a hard augmented version. However, this technique only considers the inter-segment consistency but ignores the intra-segment consistency. When a video is long (e.g., 1000 frames) and the segment number k is small (e.g., $k = 5$), the variations within each segment (200 frames) could be large. The above temporal consistency does not consider this aspect.

The straightforward solution is to sample another sequence, in the same way as constructing $\bar{\mathbf{U}}$, and enforce its prediction consistently with that of $\bar{\mathbf{U}}$. However, this

will introduce additional memory and computation burden. Here we introduce a nearly cost-free solution by leveraging historical predictions. As in each training round, we use the same way to randomly sample a sequence from a video, the past sequences thus can be viewed as other versions of the current sequence; we can enforce consistency between past predictions to the current one to achieve temporal consistency. Specifically, we construct a memory bank $\mathcal{M}$ which stores past Q predictions for each video. We employ $\mathcal{M}$ to calculate the historical consistency loss as,

$$\mathscr{L}_h = \mathop{\mathbb{E}}_{\mathbf{p}_u \sim \mathcal{M}} \Big[\mathrm{KLD}\Big(\mathbf{p}_u, H(\bar{\mathbf{U}})\Big)\Big], \tag{5}$$

where $\mathbf{p}_u$ is a historical prediction for a video clip sampled from $\mathbf{U}$. We should not set Q to be very large, as otherwise predictions produced by obsolete models are stored in the memory bank; enforcing consistency with this obsolete knowledge prevents the model from learning new knowledge. Empirically, we find setting the Q as 2 reaches well generalizable results.

3.1.4 Overall Learning Objective

We train our STHC model with the following learning objective,

$$\mathscr{L} = \mathscr{L}_{im} + \alpha\big(\mathop{\mathbb{E}}_{\bar{\mathbf{U}} \sim \mathcal{T}} (\mathscr{L}_s + \mathscr{L}_t + \mathscr{L}_h)\big) \tag{6}$$

where α is a hyper-parameter. $\mathscr{L}_{im}$ is the information maximization (IM) loss [12, 15] which was used by previous source-free domain adaptation methods [24, 50] as,

$$\mathscr{L}_{im} = -\mathop{\mathbb{E}}_{\bar{\mathbf{U}} \sim \mathcal{T}} \sum_{r=1}^{R} H(\bar{\mathbf{U}}) \log H(\bar{\mathbf{U}}) + \sum_{r=1}^{R} p_r \log p_r, \tag{7}$$

where $\mathbf{p} = -\mathbb{E}_{\bar{\mathbf{U}} \sim \mathcal{T}} H(\bar{\mathbf{U}})$ is the mean predictions over all target videos; p_r is r-th dimension of $\mathbf{p}$. The first term of Eq. (7) minimizes the entropy of the probability, encouraging the model to make confident predictions. The second term maximizes the entropy of $\mathbf{p}$, encouraging the samples are evenly distributed over all classes. **Algorithm 1** outlines the main steps of the proposed method.

Algorithm 1 Proposed STHC model

Input: Model H and unlabeled target data $\mathcal{T}$.
Output: Adapted model H.

```
Initialize an empty memory bank M for storing predictions for all training videos.
while not done do
    Randomly sample U ~ T.
    Get a sequence of frames Ū from U with segment-wise random sampling.
    Get Ū_s from Ū by spatial augment. with Eq. (1).
    Get Ū_t from Ū by temporal augment. with Eq. (3).
    Calculate predictions H(Ū), H(Ū_s), and H(Ū_t).
    Calculate spatial consistency loss L_s with H(Ū), H(Ū_s) according to Eq. (2).
    Calculate temporal consistency loss L_t with H(Ū), H(Ū_t) according to Eq. (4).
    if M is not empty for U then
        Calculate historical consistency loss L_h with H(Ū) and M according to Eq. (5).
    end if
    Calculate IM loss L_im according to Eq. (7).
    Update M with H(Ū) with the first-in-first-out principle.
    Update H with the overall loss according to Eq. (6).
end while
```

3.2. Extending to Other DA Settings

Thanks to the neat design, our STHC model can be easily extended to address other video domain adaptation problems under the source-free constraint.

Partial Domain Adaptation (PDA). PDA studies the domain adaptation scenario where classes in the target domain are a subset of classes in the source domain. Under the source-free constraint, this implies that only samples from a part of all the classes the source model is trained are used for adapting the model. Our STHC model can be directly applied to this setting. The three consistency term in Eq. (6) has no assumption on the class distribution in the target domain. The only necessary modification comes from the class-balancing term (the second term in Eq. (7)) of the IM loss. This term encourages samples to evenly distribute across all classes and thus contradicts the unbalancing reality of the target domain. Thus, we drop this term for the PDA setting.

Open-Set Domain Adaptation (OSDA). OSDA studies class-asymmetric domain adaptation too, but for the scenario where source classes are a subset of target classes. This brings challenge for model adaptation as samples from unknown class may cast negative impact on the source model. We adopt the same strategy proposed in [24] to avoid this problem. We utilize the entropy of predictions as a measurement of uncertainty and divide target samples into two groups using K-means clustering with the uncertainty scores. The group with mean uncertainty higher than the global mean of uncertainty for the whole dataset is regarded as samples from unknown classes and will be rejected for adaptation. Note that the uncertainty scores are updated during the adaptation process; samples initially regarded as unknown could be included for adaptation later, and vice versa.

Black-Box Model Adaptation (BBMA). BBMA is a variant of source-free domain adaptation which assumes that

even the source model is not available for adaption; it serves as a black-box which produces outputs for given inputs [25]. We extend our STHC model to this setting with a simple two-stage based solution. In the first stage, we treat the black-box model as the teacher model and train a student model of randomly initialized weights via knowledge distillation on target data. In the second stage, we adapt the student model in the same ways as adapting an accessible source model in the SFVDA setting.

4. Experiments

4.1. Experimental Setup

Benchmarks. We conduct experiments on the following four common benchmarks in this field [36, 50]. **(1)** ***UCF-HMDB*** includes videos from the 12 overlapping classes from the UCF101 (U) dataset [42] and the HMDB51 (H) dataset [16]. We evaluate 2 tasks, adapting U↔H videos. **(2)** ***UCF-Kinetics*** is from the *SportsDA* benchmark [50] which originally included 3 datasets, *UCF101* (U) [42], *Kinetics-600* (K) [2], and *Sports-1M* [13] from 23 action classes. These 3 datasets form 6 cross-domain tasks. However, *Sports-1M* is crawled from Youtube and we found many of the videos no longer exist, our dataset was much smaller than that of the *SportsDA* benchmark. For ease of future experiments, we exclude *Sports-1M* and have 2 tasks adapting U↔K videos. **(3)** ***Jester*** is a large-scale hand gesture dataset [30]. A subset of this dataset is used to construct two domains, JS and JT, which containing $51,498$ and $51,415$ video clips, respectively from 7 classes. Following the previous method [36], we evaluate the adaptation from JS to JT. **(4)** ***DailyDA*** is another large-scale cross-domain action recognition benchmark. It includes 4 datasets, namely, ARID (A) [49], HMDB51 (H) [16], Moments-in-Time (M) [31], and Kinetics (K) [2]. Videos from 8 shared classes are used for cross-domain evaluation. The four datasets result in a total of 12 tasks.

Implementation details. We follow the existing SFVDA method, ATCoN [50], and adopt the Temporal Relation Network [54] as the model for experiments. Specifically, we use ResNet-50 [11] as our frame feature extractor F and a Multi-Layer Perceptron (MLP) as our temporal encoder G. The frame features are averaged as a single vector for action prediction by the classifier C, which is implemented as one fully-connected layer. When performing segment-wise random sampling to train the model. we evenly divide a video into 5 segments, i.e., $K = 5$, and sample one frame from each segment. The hyper-parameter α in Eq. (6) is set as $\alpha = 0.1$. For all experiments, we train the model with 15 epochs at an initial learning rate of 10^{-3} for the two small benchmarks, *UCF-HMDB* and *UCF-Kinetics*, and a smaller initial learning rate of 10^{-4} for the large-scale benchmarks, *Jester* and *DailyDA*.

Methods	*UCF-HMDB*			*UCF-Kinetics*		
	U→H	H→U	Avg.	K→U	U→K	Avg.
TRN (source)	82.2	88.1	85.2	92.7	82.5	87.6
SHOT	82.2	81.2	81.7	94.1	75.3	84.7
ATCoN	85.6	90.2	87.9	95.3	87.3	91.3
STHC (ours)	**90.9**	**92.1**	**91.5**	**96.1**	**89.8**	**93.0**

Table 1. Results on *UCF-HMDB* and *UCF-Kinetics*. The best results are in **bold**.

	C1	C2	C3	C4	C5	C6	C7	Avg.
TRN (source)	43.4	99.4	2.7	14.2	60.8	**97.9**	**94.2**	50.3
SHOT	19.9	**99.6**	**88.2**	21.0	98.7	91.4	78.8	63.6
ATCoN	0.4	99.4	22.8	46.1	**99.1**	91.4	88.4	54.0
STHC (ours)	**66.1**	99.4	76.5	**60.3**	98.3	96.6	83.5	**78.4**

Table 2. Results on *Jester*. C1∼C7 represent the 7 classes from the datasets.

Baseline methods. We mainly compare with SHOT [24] and ATCoN [50]. SHOT is a well established source-free domain adaptation method. It was for image classification but can be easily extended to videos by applying the techniques on video feature embeddings. ATCoN is an existing SFVDA method. Please note *while we use the same networks and learning recipe as ATCoN, and develop based on the released code*[1]*, we find that, by using the released code with the suggested instructions, we can get source models of much better results than the reported ones. This means we start from better source models than those reported in the ATCoN paper. To make fair comparison, we do not cite the results from the paper and use our reproduced results instead, which are much better than the reported ones.*

4.2. Comparative Studies

Small-scale benchmarks. Table 1 shows the results on the two small-scale benchmarks, i.e., *UCF-HMDB* and *UCF-Kinetics*. We can see that in both benchmarks, the proposed STHC method improves the source TRN model and reaches the best performance compared with the other two baseline methods. Surprisingly, the performance gets worse after using SHOT to adapt the source model. The reason might be that SHOT treats videos in the same way as images and does not consider the crucial temporal information in videos, which leads to negative adaptation effect. In contrast, both ATCoN and STHC include techniques for adaptation in the temporal aspect and hence achieve positive adaptation results. Compared with ATCoN, STHC further incorporates adaptations to spatial aspects, which contributes to the superior performance.

Large-scale benchmarks. Table 2 and Table 3 show the results on the two large-scale benchmarks, i.e., *Jester* and

[1]https://github.com/xuyu0010/ATCoN

	TRN (source)	SHOT	ATCoN	STHC (ours)
K→A	**24.4**	20.1	14.6	15.5
K→H	**50.0**	49.1	49.1	48.7
K→M	32.5	**36.8**	35.8	34.8
M→A	**31.2**	16.1	13.6	18.4
M→H	50.8	53.3	**58.3**	56.3
M→K	75.9	42.8	71.7	**76.6**
H→A	**17.4**	14.3	10.2	13.8
H→M	32.3	35.0	38.8	**39.8**
H→K	43.7	36.9	45.8	**50.1**
A→H	17.9	34.2	40.0	**44.6**
A→M	18.3	**27.3**	**27.3**	**27.3**
A→K	22.3	41.8	36.8	**44.7**
Avg.	34.7	34.0	36.8	**39.2**

Table 3. Results on *DailyDA*. "K", "A", "H" and "M" are short for the four datasets in the benchmark.

DailyDA, respectively. We can see from Table 2 that the proposed STHC model achieves impressive results. It lifts the source TRN model by 28.1 for the average accuracy, and meanwhile beats the two existing methods by large margins. It is observed that all the three baseline methods have sharp result variances for different classes, for example, ATCoN achieves nearly perfect performance for $C5$, but near zero performance for $C1$. This extremely imbalanced performance towards different classes indicates that the embedding spaces of these methods are dominated by several major classes, such that data are collapsed into the spaces of these several major classes. The accuracies could be very high for the major classes but very low for the minor classes. In contrast, our method exploits stochastic augmentations from both spatial and temporal aspects, and is essentially able to mitigate class imbalance and prevents overfitting. This explains the astonishing performance gains.

Table 3 shows quite different phenomenon. While STHC still owns advantages over the other methods for the average accuracy and gets the best results for 6 out of the 12 tasks, it performs worse even to the source TRN model in several tasks. We analyze the main reason could be that this benchmark is so hard; the source models in most cases can only get less than 30% in accuracy, which makes the following adaptation process extremely vulnerable. SHOT, ATCoN and our STHC model all encourage confident predictions; if the starting source model is not able to produce reasonably reliable prediction, the error will be accumulated and leads to negative adaptation.

4.3. Property Analysis

Ablation study. The main technical contribution of this paper is the three consistency learning techniques. Table 4 show the affect of removing the three techniques on one small-scale benchmark (*UCF→ HMDB*) and one large-scale benchmark (*JS→ JT*). We can see that in both cases, removing any of the three technique leads to performance drop, which verifies the effectiveness. Remarkably, while the spatial consistency is most important for the small-scale benchmark, removing the temporal consistency results in the most significant performance drop. In both cases, the historical consistency is least important, which is reasonable as it is used to reinforce the temporal consistency.

	UCF→HMDB	*JS→JT*
w/o *spatial consistency*	87.8	75.0
w/o *temporal consistency*	88.9	70.1
w/o *historical consistency*	89.8	76.6
w/o training the classifier	88.1	70.4
Full Model	**90.8**	**78.4**

Table 4. Ablation studywith *UCF→HMDB* and *JS→JT*, which represent small-scale and large-scale benchmarks, respectively.

	PDA	OSDA	
		OS	OS*
TRN (source)	73.81	61.0	61.7
SHOT	65.24	63.0	71.4
ATCoN	71.90	65.8	**74.5**
STHC (ours)	**75.00**	**69.5**	73.9

Table 5. Results for partial domain adaptation (PDA) and open-set domain adaptation (OSDA) with *UCF→HMDB*. "OS" denotes accuracy over all classes and "OS*" measures accuracy only for known classes.

Effect of updating the classifier. Starting from SHOT [24], many SFDA works adapt the source model by fixing the classifier and only updating the feature extractor [26, 50]. ATCoN [50] adopts this learning paradigm for SFVDA too. However, we empirically found that freezing the classifier leads to performance drop for our method. As shown in Table 4, the accuracy drops by 2.7 and 8 points for the small-scale and large-scale benchmarks, respectively. We speculate the reason could be that we perform model adaptation by forcing it to produce augmentation-variant predictions in which the classifier plays a vital role. The previous methods performed model adaptation by forcing the feature extractor to produce source-like feature representations for target data; so the classifier served as a guide and was better not updated.

Partial domain adaptation (PDA) We employ a benchmark established in [48] for PDA evaluation. The benchmark was originally for unsupervised video domain adaptation, and comprises of $2,780$ videos from 14 common classes of *UCF101* and *HMDB51*. We evaluate the adaptation task from *UCF101* to *HMDB51*. The first 7 categories in alphabetic order of *HMDB51* are chosen as the target categories. To utilize this benchmark for the source-free setting, we first train a source model with videos from all the 14 classes in *UCF101*, then we adapt the source model using unlabeled videos from the 7 classes in HMDB. Table 5 shows the results. We can see that while both SHOT and AT-

TRN (black-box)	82.2
TRN (source)†	81.1
SHOT	76.7
ATCoN	86.4
STHC (ours)	**87.8**

Table 6. Results for black-box model adaptation with *UCF→HMDB*. †The source model here is trained from scratch by taking the black-box source model as the teacher with knowledge distillation on target data.

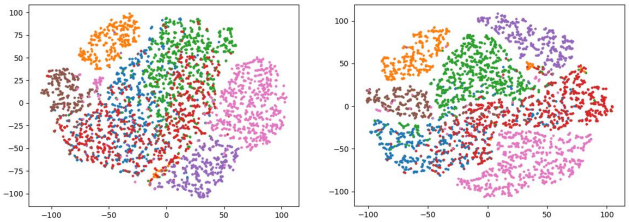

Figure 3. t-SNE visualization of video embeddings before (Left) and after (Right) adaptation on the target domain of the *Jester* benchmark. Samples from the same class are in the same color.

CoN suffer from negative transfer and obtain results worse than the source model, our STHC model manages to get some improvements. This shows the robustness of STHC on coping with different video domain adaptation problems.

Open-set domain adaptation (OSDA). We adapt the above PDA benchmark for OSDA evaluation. As in OSDA, the target domain contains unknown classes absent in the source domain, we select the first 7 categories in alphabetic order in *UCF101* as the source categories and all the 14 classes in *HMDB51* as the target categories. We train a source model using data from the 7 source categories and adapt the model with data from the 14 target categories. For evaluation, we follow previous methods [24, 38] and measure the accuracy over all classes $OS = \frac{1}{K+1}\sum_{k=1}^{K+1} Acc_k$, where K denotes the number of known classes, and $(K+1)$-th class represent the unknown class. We also calculate accuracy only for *known* classes as $OS^* = \frac{1}{K}\sum_{k=1}^{K}$. Table 5 shows the results. We can see that all the three adaptive methods achieve a positive adaptation and reach better performance than the baseline source method. While ATCoN performs slightly better than STHC for the OS^* measurement, it is much worse than STHC for the OS measurement, which indicates that ATCoN wrongly predicts much more known class samples as unknown than our method.

Black-box model adaptation (BBMA). We propose a two-step learning strategy to extend our method to the BBMA setting, first training a student model using the black-box source model as the teacher via knowledge distillation, and adapting the student model. We use the same strategy to extend SHOT and ATCoN to this setting too. Table 6 shows the results with *UCF→HMDB*. We can see that the knowledge distillation step looks very effective: while the teacher source model gets an accuracy of 82.2, the student model performs only slightly lower and obtains 81.1 in accuracy. With this fairly good student model, our STHC model achieves the best adaptation performance. This further verifies our advantage for the SFVDA problem.

α	0.001	0.01	0.1	1	10
Acc.	84.9	89.2	90.9	83.0	8.2

Table 7. Sensitive analysis for α with *UCF→HMDB*.

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	Avg.
0	0	0	0	0	100	0	0	0	0	0	0	8.2

Table 8. Per-class accuracy with *UCF→HMDB* when $\alpha = 10$.

Parameter analysis. Table 7 shows the parameter analysis results for α which balances the consistency loss terms and the Information Minimization (IM) loss term in Eq. (6). It is shown that the model degrades with a big α. We analyze the reason could be that the proposed consistency learning techniques can only help learn the space structure around each individual sample; the separation of the samples from different classes still relies on the IM term, which enforces the model to make certain predictions, and thus drives decision boundaries away from data-dense regions [3, 10, 39]. If the consistency terms dominate the loss, the model will degenerate to a point where samples from all classes are collapsed to the same local space. This is verified in Table 8 where we can see the major class gets perfect accuracy, while other classes get zero accuracy.

t-SNE visualization. Figure 3 shows the t-SNE visualization of the feature embeddings of test videos from *JT* in the *Jester* benchmark. We can see that the embeddings exhibit better clustering structure after adaptation. The number of clusters overlapping with each other is reduced from 3 to 2 out of the 7 clusters

5. Conclusions

We introduced in this paper the STHC model for SFVDA. STHC adapts a source model by encouraging it to surpass spatio-temporal variations in the video space that account for domain shifts and to make consistent predictions for video with different variations. To simulate spatio-temporal variations, we constantly apply stochastic spatial and temporal augmentations on each target video and enforce three types of consistency learning on the video and its augmented versions, including spatial consistency, temporal consistency and historical consistency. STHC is very flexible and can be easily extended to various other domain adaptation settings. Extensive experiments show that all the three consistency learning techniques help improve performance, and STHC outperforms existing methods for both small-scale and large-scale benchmarks for the standard SFVDA setting, as well as other adaptation settings.

References

[1] Sk Miraj Ahmed, Dripta S Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *CVPR*, 2021. 3

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1, 6

[3] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *AIStats*, 2005. 8

[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 *ICCV*, 2019. 1, 2

[5] Peipeng Chen, Yuan Gao, and Andy J Ma. Multi-level attentive adversarial learning with temporal dilation for unsupervised video domain adaptation. In *WACV*, 2022. 2, 3

[6] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *ECCV*, 2020. 1, 2

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2019. 4

[8] Victor G Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Dual-head contrastive domain adaptation for video action recognition. In *WACV*, 2022. 2

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 4

[10] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *NeurIPS*, 2004. 8

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 6

[12] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *ICML*, 2017. 5

[13] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 6

[14] 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 *ICCV*, 2021. 3

[15] Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. *NeurIPS*, 2010. 5

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 6

[17] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *ICCV*, 2021. 3

[18] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *CVPR*, 2020. 3

[19] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *WACV*, 2021. 3

[20] Kai Li, Chang Liu, Handong Zhao, Yulun Zhang, and Yun Fu. Ecacl: A holistic framework for semi-supervised domain adaptation. In *ICCV*, 2021. 2

[21] Kai Li, Curtis Wigington, Chris Tensmeyer, Handong Zhao, Nikolaos Barmpalios, Vlad I Morariu, Varun Manjunatha, Tong Sun, and Yun Fu. Cross-domain document object detection: Benchmark suite and method. In *CVPR*, 2020. 2

[22] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. Adversarial feature hallucination networks for few-shot learning. In *CVPR*, 2020. 1

[23] Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *AAAI*, 2021. 3

[24] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020. 1, 3, 5, 6, 7, 8

[25] Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In *CVPR*, 2022. 6

[26] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 7

[27] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, 2021. 3

[28] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *NeurIPS*, 2018. 2

[29] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. Adversarial bipartite graph learning for video domain adaptation. In *ACM Multimedia*, 2020. 2, 3

[30] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *ICCV Workshops*, 2019. 6

[31] 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. 6

[32] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020. 3

[33] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *AAAI*, 2020. 1, 2

[34] 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. In *IJCAI*, 2021. 3

[35] Mohammad Rostami and Aram Galstyan. Sequential unsupervised domain adaptation through prototypical distributions. *arXiv e-prints*, pages arXiv–2007, 2020. 3

[36] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *NeurIPS*, 2021. 1, 2, 3, 6

[37] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 1, 2

[38] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *ECCV*, 2018. 8

[39] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *ICLR*, 2018. 8

[40] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020. 4

[41] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *CVPR*, 2021. 3

[42] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 6

[43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1

[44] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, 2017. 1, 2

[45] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1

[46] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *ICCV*, 2021. 3

[47] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *NeurIPS*, 2020. 4

[48] Yuecong Xu, Jianfei Yang, Haozhi Cao, Zhenghua Chen, Qi Li, and Kezhi Mao. Partial video domain adaptation with partial adversarial temporal attentive network. In *ICCV*, 2021. 3, 7

[49] 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*, 2021. 6

[50] 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 *ECCV*, 2022. 1, 2, 3, 5, 6, 7

[51] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition. In *CVPR*, 2022. 3

[52] Shiqi Yang, Joost van de Weijer, Luis Herranz, Shangling Jui, et al. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. *NeurIPS*, 2021. 3

[53] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *ICCV*, 2021. 3

[54] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 6