Overcoming Label Noise for Source-free Unsupervised Video Domain Adaptation

Avijit Dasgupta
CVIT, IIIT Hyderabad
India

C. V. Jawahar
CVIT, IIIT Hyderabad
India

Karteek Alahari
Univ. Grenoble Alpes, Inria
CNRS, Grenoble INP, LJK, France

ABSTRACT

Despite the progress seen in classification methods, current approaches for handling videos with distribution shifts in source and target domains remain source-dependent as they require access to the source data during the adaptation stage. In this paper, we present a self-training based *source-free* video domain adaptation approach (without bells and whistles) to address this challenge by bridging the gap between the source and the target domains. We use the source pre-trained model to generate pseudo-labels for the target domain samples, which are inevitably noisy. We treat the problem of source-free video domain adaptation as learning from noisy labels and argue that the samples with correct pseudo-labels can help in the adaptation stage. To this end, we leverage the cross-entropy loss as an indicator of the correctness of pseudo-labels, and use the resulting small-loss samples from the target domain for fine-tuning the model. Extensive experimental evaluations show that our method termed as *CleanAdapt* achieves ~ 7% gain over the source-only model and outperforms the state-of-the-art approaches on various open datasets.

CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; *Neural networks*; *Computer vision representations.*

KEYWORDS

action recognition, domain adaptation, transfer learning

ACM Reference Format:
Avijit Dasgupta, C. V. Jawahar, and Karteek Alahari. 2022. Overcoming Label Noise for Source-free Unsupervised Video Domain Adaptation . In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'22), December 8–10, 2022, Gandhinagar, India,* Soma Biswas, Shanmuganathan Raman, and Amit K Roy-Chowdhury (Eds.). ACM, New York, NY, USA, Article 21, 9 pages. https://doi.org/10.1145/3571600.3571621

1 INTRODUCTION

Action recognition models [2, 28, 34, 40] often encounter new domains with *distribution-shift* [37] when deployed in the real world. Such shifts can occur in videos for several reasons: relative differences in the speed and duration of the action, camera movement, viewpoints, etc. Thus, the resulting difference in data distributions of the training (source domain) and the test (target domain) data produces a degraded performance. Furthermore, the source domain data usually comes with fully labeled videos, whereas the target domain data is typically unlabeled to reduce the annotation cost. The primary goal for unsupervised domain adaptation (UDA) is to reduce this performance gap by adapting the model to the label-scarce target domain by transferring the knowledge learned from the label-rich source domain dat [11, 24, 33, 39, 45]. Source-free UDA [13, 19, 23, 46] takes this approach one step further by assuming the unavailability of the source domain data for adaptation. This is a more practical setup than traditional source-dependent UDA mainly due to privacy issues, computation cost, and storage complexity [13, 19, 23, 46].

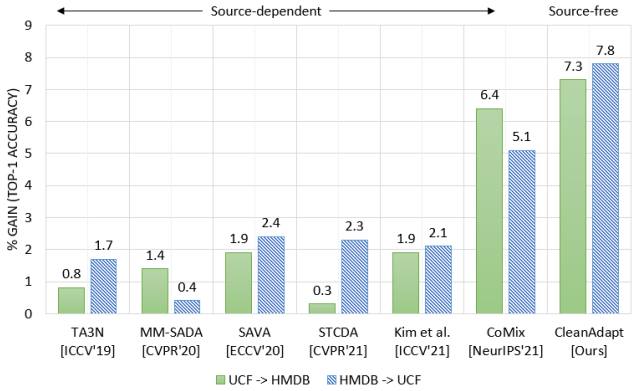


Figure 1: Existing approaches [3, 5, 16, 27, 32, 35] have a *source-dependent* adaptation stage achieving marginal performance gain over the source-pretrained models. On the other hand, our proposed method CleanAdapt achieves significant performance improvement over the source-only model while being *source-free* (*i.e.*, the adaptation stage does not require videos from the source domain). (Best viewed in color.)

There has been a surge of interest in domain adaptation for videos recently [3–6, 14, 16, 27, 29, 35]. These approaches either propose to directly extend the adversarial learning framework [14] from image-based methods [11] or couple it with some temporal attention weights [3, 5] and self-supervised pretext tasks [5, 27] to align the segment-level features between the domains. However, these strategies produce only a modest ~ 2% gain over the source-only model (see Figure 1). Recently, there has been a paradigm shift from adversarial to contrastive learning framework [16, 32, 35] for video domain adaptation which proved to be beneficial. As shown in Figure 1, the most recent method [32] achieves 6.4% and 5.1% gain over the source-only model on UCF → HMDB and HMDB →

ICVGIP'22, December 8–10, 2022, Gandhinagar, India

ACM ISBN 978-1-4503-9822-0/22/12.
https://doi.org/10.1145/3571600.3571621

UCF datasets, respectively. However, all of these existing methods are inherently complex and use source domain videos during the adaptation stage, which is untenable in several scenarios [13, 23, 46], as discussed earlier.

In this work, we present an effective approach that leverages the self-training framework [52] for source-free video UDA where we do not have access to the source-domain videos during the adaptation stage. We generate *pseudo-labels* for the unlabeled target domain videos using a source pre-trained model. These pseudo-labels are indeed noisy due to the existing domain gap. Finetuning the source pre-trained model with these noisy pseudo-labels is a sub-optimal solution as the presence of incorrect pseudo-labels hinders the adaptation stage as discussed in Sec. 4.4. However, we observe that these pseudo-labeled target domain videos are not completely unusable, and in fact, there is a substantial number of target domain videos with correct pseudo-labels. For example, in the case of HMDB $\rightarrow$ UCF, the HMDB pre-trained model produces pseudo-labels with $\sim$ 90% accuracy on the UCF dataset, and we experimentally show that this amount of data is sufficient for adaptation. Throughout this paper, we term these samples with correct pseudo-labels as *clean*, whereas the samples with incorrect pseudo-labels are termed as *noisy*. We observe that the network learns clean samples first before memorizing the noisy samples, and this acts as the core idea behind the adaptation stage in our proposed method (Figure 2). We discuss this further in Sec. 3.3.

To our knowledge, we are the first to address the video domain adaptation problem in a source-free setup. We treat this problem as learning from noisy labels and propose a self-training based approach that selects the clean samples from the noisy pseudo-labeled target domain samples to re-train the model for gradually adapting to the target domain in an iterative manner. Thus, we name our approach as CleanAdapt. In contrast to the previous methods [3–6, 14, 16, 27, 29, 35], CleanAdapt is inherently source-free as it only requires target domain videos and their corresponding pseudo-labels. Our proposed method surpasses all other source-dependent state-of-the-art methods by a large margin on UCF $\leftrightarrow$ HMDB and EPIC-Kitchens datasets, despite being source-free.

2 RELATED WORK

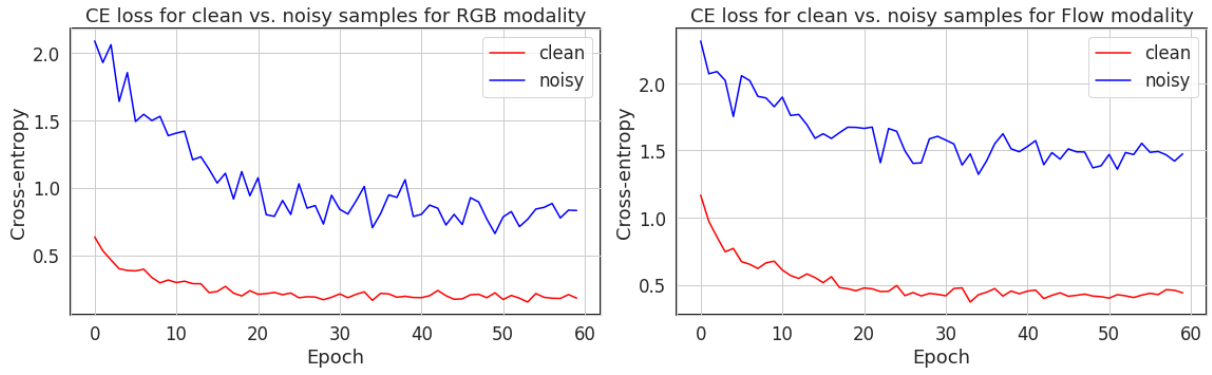
Supervised Action Recognition. Convolutional neural networks (CNNs) are now the de-facto solution for action recognition tasks. Various efforts have been made in this context to capture spatio-temporal information in videos, starting from two-stream networks with 2D [34, 40, 51] to 3D CNNs [2, 9, 38]. Recent advances in action recognition focus on capturing long-term context from videos [8, 41, 44]. Despite their success, these methods suffer from a common limitation: a subtle difference in testing data distribution from training data limits their ability to generalize in the new domain. Thus, these methods require a large number of labeled data in the new domain for fine-tuning, which is often time-consuming and expensive. In contrast, we focus on unsupervised video domain adaptation to eliminate the need for labeled data from the target domain.

Domain Adaptation for Action Recognition. Early works [3, 5, 14, 27, 29] on video UDA are inspired by image-based UDA's adversarial framework [11]. Jamal *et al.* [14] propose to align the source and the target domains using a subspace alignment technique and outperform all previous shallow methods. Chen *et al.* [3] show the efficacy of attending to the temporal dynamics of video for domain adaptation. TCoN [29] used cross-domain co-attention module for matching the source and the target domain features with appearance and motion streams. Munro *et al.* [27] were among of the first to show the effectiveness of learning multi-modal correspondence for video domain adaptation. SAVA [5] proposed an attention-augmented model with a clip order prediction task to re-validate the effectiveness of self-supervised learning for video domain adaptation, as shown in [27]. However, the adversarial methods are complex and sensitive to the choice of hyperparameters [32].

There has been a recent shift from adversarial to contrastive learning-based methods for the video UDA task. Song *et al.* [35] propose to bridge the domain gap using a self-supervised contrastive framework named cross-modal alignment. In a similar direction, Kim *et al.* [16] use a cross-modal feature alignment loss for learning domain adaptive feature representation. CoMix [32] represents videos as graphs and uses temporal-contrastive learning over graph representations for transferable feature learning. Additionally, these methods [16, 32, 35] generate pseudo-labels from the source pre-trained model for the target domain videos and use only the target domain samples with high-confident pseudo-labels in their contrastive loss in each iteration. However, the source-only model often makes wrong predictions with high confidence due to the distribution shift for target domain videos which can hinder the adaptation. To address this, we treat target pseudo-labels as noisy and formulate the domain adaptation problem as learning from noisy labels. Moreover, the adaptation stage in these methods [3–5, 14, 16, 27, 29, 32, 35] is *source-dependent*. This is an impractical assumption as the source data transfer during the deployment phase of the model is often infeasible. In contrast, we propose a *source-free* video domain adaptation approach that achieves state-of-the-art results with *only* target domain data.

Source-free Domain Adaptation for Images. There has been a significant effort for adaptation to the target domain without the source-domain data for images [13, 21, 46]. These approaches consider a closed-set setting where the label set does not change across domains. Recently, Kundu *et al.* [19] proposed a universal source-free setup where unknown classes can appear in the target domain. We follow [13, 21, 46] and assume a closed-set setup in our work for simplicity.

Learning from Noisy-labels. Self-training based methods with careful design choices may still produce over-confident incorrect predictions. To alleviate this issue, we resort to learning from label-noise literature. One of the popular approaches to reducing the effect of noisy-labels is to design noise-robust losses [10, 25, 42]. However, these methods fail to handle real-world noises [50]. According to [1], deep neural networks produce small loss values for the samples with correct pseudo-labels. Thus, a popular direction for handling label-noise is to use the cross-entropy loss as an indicator of label correctness [12, 48] and use these small-loss samples for re-training the networks. In this work, we demonstrate that the small-loss samples are the potential clean samples and are enough to help our model adapt to the target domain. Therefore,



(a) (b)

Figure 2: Average cross-entropy loss per epoch of training with pseudo-labeled target domain videos for clean vs. noisy samples with (a) RGB and (b) Flow modalities. We refer to the target domain samples with correct pseudo-labels as *clean* samples and those with incorrect pseudo-labels as *noisy* samples. Note that the groundtruth labels are only used to identify the clean vs. noisy samples for visualization purposes and not for training the model. The networks learn the clean samples first before memorizing the noisy samples according to the deep memorization effect presented in [1]. In our proposed approach CleanAdapt, we exploit this connection to select the clean samples for fine-tuning the model to adapt to the target domain.

our proposed approach is more straightforward and requires only pseudo-labeled target domain samples.

3 APPROACH

3.1 Problem Definition

In this paper, we consider the *source-free* UDA task for videos. Here, we are given a labeled source domain dataset of videos $D_s = \{(x_s, y_s) : x_s \sim p\}$, where p is the source data distribution and y_s is the corresponding label of x_s. We are also given an unlabeled target domain dataset $D_t = \{x_t : x_t \sim q\}$, where q is the target distribution that is different from the source distribution p. We assume that the source and the target domains share the same label-set C *i.e.* closed-set domain setup.

For a video clip x from any domain, we consider two modalities, $x = \{x_a, x_m\}$, where x_a is the appearance (RGB) stream and x_m is the motion (optical flow) stream. We use two 3D CNN backbones f_a and f_m, one for each modality that classify a video into one of the $|C|$ classes. We aim to adapt the 3D CNNs (f_a and f_m) to the target domain. We also note that, the source domain videos are only available during the pre-training stage. However, we do not use the source dataset $\mathcal{D}_s$ during the adaptation stage as we are interested in the more realistic source-free setup. We show an overview of the proposed method in Figure 3.

3.2 Self-training based Domain Adaptation

Contrary to the adversarial learning based approaches [3, 5, 27], we take the path of self-training [22, 26, 52] primarily due to its simplicity in the adaptation stage. First, we pre-train the 3D CNN models using the labeled source videos from D_s. Second, we generate labels for the unlabeled target dataset D_t using the source pre-trained model referred to as pseudo-labels. Third, we retrain the networks f_a and f_m using the pseudo-labeled target domain videos from D_t for adaptation. One of the possibilities is to use all the samples with their corresponding pseudo-labels to retrain the networks. However, the pseudo-labels contain noise due to the existing domain gap between the source domain D_s and the target domain D_t. Retraining the f_a and f_m with all these pseudo-labels lead to a sub-optimal result, as discussed in Section 4. We aim to answer the following question in this paper: what kind of D_t can help us in adaptation?

3.3 Clean Samples are All You Need

The pseudo-labels contain several samples with correct pseudo-labels (clean samples). For example, there are ~90% samples with correct pseudo-labels in UCF dataset when generated using the HMDB pre-trained networks. Thus, if we can filter out the noisy samples and keep only the clean samples, we can easily finetune our networks (f_a and f_m) using these clean samples and their corresponding correct pseudo-labels. Thus, we argue that these clean samples are the ones, which can help in domain adaptation. Now, the important question is how to sample the clean samples from the noisy ones?

To this end, we cast the problem of video domain adaptation as learning from noisy labels due to noisy pseudo-labels. In Figure 2, we observe that deep neural networks learn the clean samples easily and have difficulty learning from the noisy samples due to the memorization effect [1]. Thus, samples with *low loss* values are the potential clean samples. In this work, we design an approach without bells and whistles, *CleanAdapt*, aiming to select the clean samples based on the loss generated by the model against their corresponding pseudo-labels for adaptation. In each epoch of the adaptation stage, we select these clean samples from the target domain and use them to re-train the source-only models f_a and f_m.

There are three key advantages to this: (1) we do not need to modify the overall training regime (*e.g.* contrastive learning for domain alignment [16, 32, 35]) during adaptation, (2) we do not need to make any domain adaptation-specific design choices (*e.g.*,

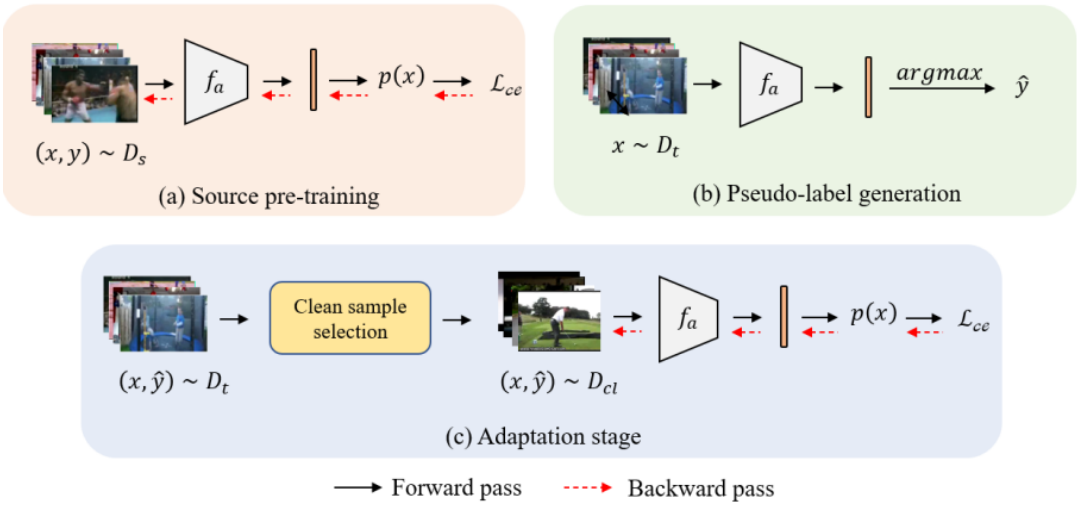


Figure 3: Overview of our CleanAdapt framework for source-free video domain adaptation, which has three stages. (a) The model (f_a) is first pre-trained on the *labeled* source domain videos from $\mathcal{D}_s$. For brevity, only the single-stream model is shown here. (b) This source pre-trained model is then used to generate pseudo-labels $\hat{y}$ for the *unlabeled* target domain videos from $\mathcal{D}_t$. Inevitably, these pseudo-labels are noisy due to the domain shift between the source and the target domains. (c) A *clean sample selection* module is used to select a set $\mathcal{D}_{cl}$ of small-loss samples as potential clean samples. The source pre-trained model is finetuned on these clean samples from $\mathcal{D}_{cl}$ using their corresponding pseudo-labels $\hat{y}$. We repeat this step multiple times. See text in Sec. 4 for implementation details. (Best viewed in color.)

background mixing [32]), and (3) we implicitly design an adaptation method that *does not* need any source dataset during the adaptation stage. The overall training pipeline of our CleanAdapt framework is shown in Figure 3.

3.4 Source Pre-training

In the source pre-training stage, we train the 3D CNNs f_a and f_m using the labeled source-domain dataset D_s and we term these as a *source only* model. For a sample $(x, y) \in D_s$, we average the logits obtained from $f_a(x)$ and $f_m(x)$ to compute the final score $p(x)$ as follows -

$$p(x) = \sigma(f_a(x) + f_m(x)). \tag{1}$$

We use the conventional cross-entropy loss between the predicted class probabilities $p(x)$ and the one-hot encoded ground-truth label y as the loss function for training -

$$\mathcal{L}_{ce}(x) = -\sum_{c=1}^{|C|} y^c \log(p^c(x)), \tag{2}$$

where y^c and p^c represent the c^{th} element of y and $p(x)$ respectively for class c. The main goal for this pre-training step is to equip our model with the initial knowledge of the classes present in the source dataset D_s. Figure 3(a) depicts this step.

3.5 Pseudo-label Generation

As illustrated in Figure 3(b) the next step is to generate the pseudo-labels for the unlabeled target domain samples. Once the model is pre-trained on the source domain videos, we use the learned notion of the class semantics of the model to generate labels for the target domain data. Note that these generated labels are not the actual labels for the target domain videos. Thus, we term these source-only model-generated labels as *pseudo-labels* $\hat{y}$. Formally,

$$\hat{y}(x) = \arg\max_c p^c(x), \tag{3}$$

where $x \in D_t$. Due to the domain shift between the source and the target, these pseudo-labels $\hat{y}$ are noisy.

3.6 Adaptation

Once the pseudo-labels are obtained from the source pre-trained model for the target domain videos, we use these pseudo-labeled target videos for adaptation, as shown in Figure 3(c). As discussed earlier, the pseudo-labels are noisy, and we would like to extract the samples with correct pseudo-labels (clean samples) for adaptation. Each epoch of the adaptation stage has two key steps in our CleanAdapt framework: (a) clean sample selection, and (b) fine-tune the models f_a and f_m using these clean samples.

Clean sample selection. To filter out the target domain videos with noisy pseudo-labels, we exploit the connection between the small-loss and the clean samples. The videos are first grouped into $|C|$ classes based on their pseudo-labels and sorted in ascending order of their cross-entropy loss values computed using the *prediction* made by the model and their corresponding *pseudo-labels*. If the pseudo-labels are correct, the model will likely produce a small loss.

Inspired by [12, 43], we define a hyper-parameter named *keep-rate* τ. For each groups, we select τ proportion of the total number

Figure 4: The *clean sample selection* module. The pseudo-labeled target domain videos from $\mathcal{D}_t$ are grouped according to their pseudo-labels $\hat{y}$ and sorted in ascending order of the loss generated by the model against their pseudo-labels. The *keep-rate* τ ($\tau = 0.6$ in this example) determines the number of samples to be selected for adaptation, with small-loss values for each class. We have used only four classes here for simplicity. We enclose the videos with the correct pseudo-labels in a green box, and the ones with incorrect pseudo-labels in a red box for visualization purposes. (Best viewed in color.)

of samples with small losses. We call this updated dataset of small-loss samples as $D_{cl} \subset D_t$ and discard the rest of the samples. This step is illustrated in Figure 4.

Fine-tuning. In this step, the networks f_a and f_m are re-trained using the samples x and their corresponding pseudo-labels $\hat{y}$ from D_{cl} using the cross-entropy loss as shown in Eq. 4.

$$\mathcal{L}_{ce}(x) = -\sum_{c=1}^{|C|} \hat{y}^c \log(p^c(x)), \tag{4}$$

where $(x, \hat{y}) \in D_{cl}$. We repeat these two steps in an iterative manner until the networks reach convergence.

4 EXPERIMENTS

4.1 Datasets and Metrics

We consider both first-person and third-person videos for benchmarking our proposed approach. Following [3, 5, 16, 27, 32, 35], we use publicly available UCF101 [36] and HMDB51 [18] for third-person videos and EPIC-Kitchens [7] for first-person videos. We show experimentally that our approach adapts well for first-person as well as for third-person videos.

UCF ↔ HMDB. We use the official split released by Chen *et al.* [3] for UCF ↔ HMDB to evaluate our CleanAdapt on video domain adaptation. In total, this dataset has 3209 third-person videos with 12 action classes. Specifically, all videos are a subset of the original UCF101 [36] and HMDB51 [18] datasets with 12 classes common between them. Following [3], we use two settings: UCF101 → HMDB51, and HMDB51 → UCF101.

Table 1: Performance comparisons with state-of-the-art video domain adaptation methods on UCF101↔ HMDB51. Result for MM-SADA [27] is taken from Kim *et al.* [16]. The results for our methods are highlighted in gray color.

Method	Two-stream?	Source-free?	Datasets	
			UCF → HMDB	HMDB → UCF
Source only [3]			80.6	88.8
TA3N [3]	✗	✗	81.4	90.5
Target supervised [3]			93.1	97.0
Source only [5]			80.3	88.8
SAVA [5]	✗	✗	82.2	91.2
Target supervised [5]			95.0	96.8
Source only [35]			82.8	89.8
STCDA [35]	✓	✗	83.1	92.1
Target supervised [35]			95.8	97.7
Source only [16]			82.8	90.7
MM-SADA [27]	✓	✗	84.2	91.1
Kim *et al.* [16]	✓	✗	84.7	92.8
Target supervised [16]			98.8	95.0
Source only [32]			80.3	88.8
CoMix [32]	✗	✗	86.7	93.9
Target supervised [32]			95.0	96.8
Costa *et al.* [6]	✗	✗	87.8	95.8
Source only			80.6	89.3
CleanAdapt	✗	✓	86.1 ▲+5.5	96.1 ▲+6.8
Target supervised			93.6	98.4
Source only			82.5	91.4
CleanAdapt	✓	✓	89.8 ▲+7.3	99.2 ▲+7.8
Target supervised			95.3	99.3

EPIC-Kitchens. This is the largest video domain adaptation dataset which contains egocentric videos of fine-grained actions recorded in different kitchens. We follow the official split provided by [27]. This dataset contains videos from the three largest kitchens *i.e.* D1, D2, and D3, with 8 action categories. EPIC-Kitchens has more class-imbalance than UCF ↔ HMDB making it more challenging [27].

Metrics. We follow the standard protocol defined by [3, 27] to compare our approach with state-of-the-art unsupervised domain adaptation methods [3, 5, 16, 21, 24, 27, 32, 33, 35] in terms of top-1 accuracy. We perform cross-domain retrieval experiments to evaluate the feature space learned by our model before and after adaptation. We report retrieval performance in terms of Recall at k (R@k), implying, if k closest nearest neighbours contain one video of the same class semantics, a correct retrieval is counted.

4.2 Implementation Details

We use the Inception I3D [2] network as the backbone for both RGB and Flow modalities. Following the prior video domain adaptation works [5, 16, 27, 35], we use the Kinetics [15] pre-trained weights to initialize the I3D network. We randomly sample 16 consecutive frames and perform the same data augmentation used in [5, 16, 27] for all our steps. We set the batch size to 48 for both UCF ↔ HMDB and EPIC-Kitchens datasets. We pre-compute optical flow using the TV-L1 algorithm [49].

Source pretraining stage. We train the model on the source dataset for 40 and 100 epochs with learning rates $1e-2$ and $2e-2$ for UCF ↔ HMDB and EPIC-Kitchens dataset, respectively. We reduce the learning rate by a factor of 10 after 10, 20 epochs for UCF ↔ HMDB. For EPIC-Kitchens, we decrease the learning rate by 10 after 50 epochs. We follow [5] for other hyperparameters.

Table 2: Performance comparisons with state-of-the-art video domain adaptation methods on EPIC-Kitchens dataset. All models reported are two-stream networks. The results for our methods are highlighted in gray color.

Method	Source-free?	D2→D1	D3→D1	D1→D2	D3→D2	D1→D3	D2→D3	Mean
Source only		42.5	44.3	42.0	56.3	41.2	46.5	45.5
AdaBN [21]	✗	44.6	47.8	47.0	54.7	40.3	48.8	47.2
MMD [24]	✗	43.1	48.3	46.6	55.2	39.2	48.5	46.8
MCD [33]	✗	42.1	47.9	46.5	52.7	43.5	51.0	47.3
MMSADA [27]	✗	48.2	50.9	49.5	56.1	44.1	52.7	50.3
STCDA [35]	✗	49.0	52.6	52.0	55.6	45.5	52.5	51.2
Kim *et al.* [16]	✗	49.5	51.5	50.3	56.3	46.3	52.0	51.0
Target Supervised		62.8	62.8	71.7	71.7	74.0	74.0	69.5
Source only		41.8	40.0	46.0	45.6	38.9	44.4	42.8
CleanAdapt	✓	46.2 ▲+4.4	47.8 ▲+7.8	52.7 ▲+6.7	54.4 ▲+8.8	47.0 ▲+8.1	52.7 ▲+8.3	50.3 ▲+7.5
Target Supervised		62.1	62.1	72.8	72.8	72.3	72.3	69.1

Adaptation stage. We use the source pre-trained weights during the adaptation stage to initialize I3D [2]. This network is trained for 60 epochs with learning rates $1e-2$ and $2e-3$ for UCF ↔ HMDB and EPIC-Kitchens respectively. The learning rate is reduced by 10 after 20, 40 epochs for UCF ↔ HMDB. In the case of EPIC-Kitchens, we reduce the learning rate by 10 after 10, 20 epochs.

Our entire framework is implemented in PyTorch [30] and uses 4 NVIDIA 2080Ti GPUs. On average, training takes around 1 hour for UCF ↔ HMDB and about 5 hours for EPIC-Kitchens datasets.

4.3 Comparisons to the State-of-the-art Methods

UCF ↔ HMDB. We present the quantitative results of our approach for UCF ↔ HMDB dataset in Table 1 and compare our results with the state-of-the-art unsupervised video domain adaptation approaches. For each paper in Table 1, we also report *source only* and *target supervised* results for fair comparisons. The source-only method refers to the f_a and f_m models trained only on the `train` split of the source dataset as described in Section 3.4 and tested directly on the `validation` split of the target dataset, which serves as a lower bound of the adaptation performance. The target supervised model is trained and tested on the `train` and `validation` split of the target dataset, respectively that serves as an upper bound to the adaptation performance.

TA3N [3], SAVA [5], CoMix [32], and Costa *et al.* [6] use only appearance stream in their methods, whereas approaches like STCDA [35], MM-SADA [27], and Kim *et al.* [16] leverage both appearance and motion streams. We show the results for both single-stream and two-stream versions of our model.

Our single-stream model (*i.e.*, RGB only) achieves **86.1%** and **96.1%** top-1 accuracy with a gain of **5.5%** and **6.8%** over the source-only model for UCF → HMDB and HMDB → UCF datasets, respectively. In comparison, the best performing recent existing model CoMix [32] gives 6.4% and 5.1% gains for these two datasets, respectively. Note that all of these methods use the source data along with the target data during adaptation, whereas we use only target data in our approach and attain similar gains.

Similarly, our two-stream model achieves state-of-the-art performance on both UCF → HMDB and HMDB → UCF datasets in terms of top-1 accuracy with the values of **89.8%** and **99.2%**, respectively. This is a significant gain of **7.3%** for UCF → HMDB and **7.8%** for HMDB → UCF over the source-only model without using any source-domain data, which is much higher than the other source-dependent adaptation models [3, 5, 16, 27, 32, 35]. This validates the effectiveness of using the small-loss target domain samples in the adaptation stage for source-free unsupervised video domain adaptation.

What happens if we use only high-loss samples for adaptation? We trained our two-stream network with the high-loss samples instead of the proposed low-loss samples. For UCF → HMDB, we obtained 84.7% of accuracy after adaptation with the high-loss samples which is 5.1% less when adapted with the low-loss samples. We observe a similar drop for HMDB → UCF. This difference is even more significant when the noisy pseudo-labels are dominant (e.g., more than 12% on Epic-Kitchens).

Does the overconfident pseudo-labels trigger error accumulation? Although error accumulation is possible, we have found error accumulation to be negligible in practice. For example, the UCF pre-trained model selects low loss samples with ∼98% accuracy in each epoch of the adaptation stage from HMDB.

Comparisons with self-training based methods. In Table 1, we compare our approach with the other self-training approaches [16, 32, 35]. Our method re-purposed the LNL based pseudo-label selection method performs better than all these.

Comparisons with image-based source-free methods. In Table 3, we compare our approach with state-of-the-art image-based source-free methods [17, 20, 23, 31, 46, 47]. For [17, 20, 31, 47], we report the values with TRN [51] as their backbone network. Our model CleanAdapt achieves higher gain over their corresponding source-only model than all these image-based source-free methods. We have also adopted the frameworks proposed by [23, 46] with our 3D backbone network. Liang *et al.* [23] perform marginally better than the source-only model. Yang *et al.* [46] performance is comparable to ours on UCF → HMDB, but significantly worse on HMDB→UCF.

EPIC-Kitchens. In Table 2, we compare the results of our approach with state-of-the art image-based [21, 24, 33] as well as video-based domain adaptation [16, 27, 35] methods. All these methods reported in Table 2 use the two-stream I3D network, including our methods for fair comparisons. We quote the numbers in Table 2 for the previous works from Kim *et al.* [16] and Song *et al.* [35]. We implement our model from scratch to replicate the source only and target supervised performance, as reported in [16]. Note that there is a minor difference (∼2.7%) in the performance of the source-only model reported in MM-SADA [27] and ours. A similar difference

Table 3: Comparison with state-of-the-art image-based source-free domain adaptation techniques.

Method	Backbone	UCF → HMDB	HMDB → UCF
Source only	TRN	72.7	72.2
Kim *et al.* [17]	TRN	69.9	74.9
Li *et al.* [20]	TRN	74.4	67.3
Yang *et al.* [47]	TRN	75.3	76.3
Qiu *et al.* [31]	TRN	75.8	68.2
Source only	I3D	80.6	89.3
Yang *et al.* [46]	I3D	**86.6**	91.4
Liang *et al.* [23]	I3D	82.5	91.9
CleanAdapt	I3D	86.1	**96.1**

Figure 5: Class activation maps (CAM) on the target videos of the UCF ↔ HMDB dataset. The actions in green are correct predictions, whereas the ones in red are incorrect. Note that the adapted model (bottom row) focuses more on the action part instead of the scene context part as CleanAdapt learns domain-invariant action-relevant features as opposed to the source-only model. (Best viewed in color.)

with [27] can be seen in [32] due to the reimplementation. However, such a minor difference in source-only accuracy is not a concern for evaluating domain adaptation performance. The most important metric here is the gain achieved after adaptation over the source-only model.

MM-SADA [27] is the first to report domain adaptation results on the EPIC-Kitchens dataset achieving an average of 4.8% gain on top of their source-only model followed by Song *et al.* [35] reporting an average gain of 5.7%. Kim *et al.* [16] show an improvement of 5.5% averaged over 6 datasets. However, all of these methods use the source dataset for adaptation. In contrast to these prior approaches, our simple yet powerful source-free approach CleanAdapt, achieves an average of **7.5%** gain over the source-only model. The performance comparisons with the state-of-the-art video domain adaptation approaches for the single-stream model are in supplementary.

Visualization. In Figure 5, we show the Class Activation MAP (CAM) visualizations of our adapted model and compare them with the source-only model. The visualization shows that the source-only model attends to the irrelevant part of the scene and makes incorrect predictions, while the adapted model focuses on the important part of the scene to make correct predictions.

4.4 Hyperparameter Search

The only hyperparameter our model introduces is the keep-rate τ. It controls the number of target domain samples chosen from each class with low loss values in the adaptation stage. Figure 6 shows the ablation results of varying τ in terms of validation accuracy for the target domain.

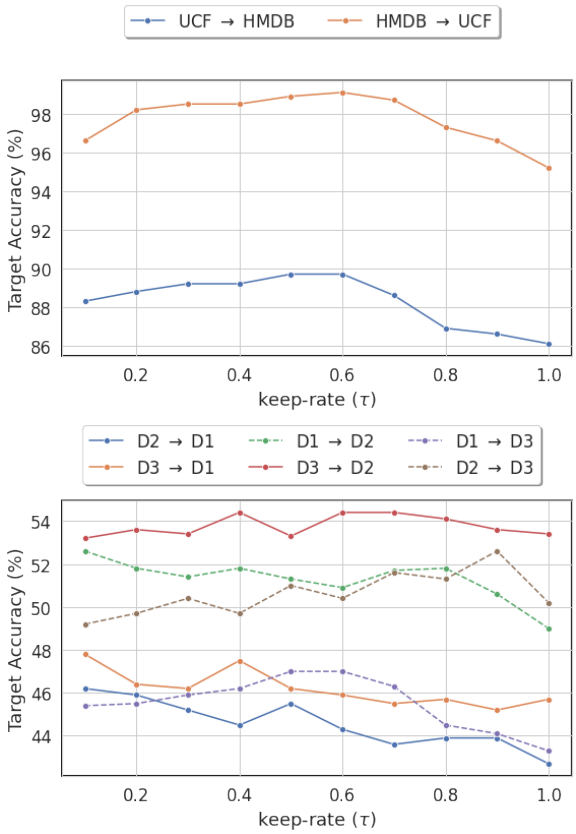


Figure 6: Hyperparameter search for keep-rate τ on UCF101↔ HMDB51 and EPIC-Kitchens datasets. The keep-rate τ controls the number of samples to be selected as clean having low-loss values computed against the pseudo-labels generated by the source-only model. All results reported here are for a two-stream network. (Best viewed in color.)

Empirically, we verify that the choice of proper keep-rate τ is particularly important. As mentioned earlier, the samples from the target domain `train` set pseudo-labeled by the source-only model have inherently noisy labels. The choice of keep-rate $\tau = 1$ is equivalent to choosing all the samples for retraining the model on the target domain. However, the noisy pseudo-labels lead to a sub-optimal adaptation performance for all the datasets. For example, the adapted model gives top-1 accuracy of 86.1% on UCF → HMDB and 95.2% on HMDB → UCF respectively for the value of $\tau = 1$. However, the value of keep-rate $\tau = 0.6$ gives top-1 accuracy of 89.8% and 99.2% on UCF → HMDB and HMDB → UCF, respectively.

4.5 Cross-domain Video Retrievals

We explore the feature space learnt by our adapted CleanAdapt model to better understand of the predictions made by the model. We evaluate the cross-domain video retrieval performance of the adapted model to better understand the feature space learnt by it. Given a query video of a particular class from the target domain, our goal here is to retrieve videos from the source domain with the same semantic category. We show the results for the two-stream networks here. We first compute the similarity scores for

Figure 7: Nearest neighbour retrieval results for the UCF → HMDB and the HMDB → UCF datasets. The left column shows the query videos from the target domain. The middle column shows the retrieved source videos using the source-only model, and the right column shows the source videos retrieved using our proposed model. (Best viewed in color.)

the individual modalities and average them for final retrieval. We evaluate both the source only and the proposed method CleanAdapt quantitatively as well as qualitatively.

Table 4: Cross-domain video retrieval results on UCF ↔ HMDB dataset. Given queries from the target domain, we evaluate retrieved videos from the source domain in terms of R@k, where $k \in \{1, 5, 10\}$. Note that, all models reported here are two-stream networks and we average the similarity score from each modality to retrieve the source videos.

Method	UCF → HMDB			HMDB → UCF		
	R@1	R@5	R@10	R@1	R@5	R@10
Source Only	0.82	0.87	0.90	0.88	0.94	0.95
CleanAdapt	**0.92**	**0.97**	**0.99**	**0.91**	**0.97**	**0.98**

In Table 4, we show the quantitative results for the cross-domain video retrieval task for the UCF ↔ HMDB dataset. Our model retrieves better source videos from the target queries with R@1 of 0.92 and 0.91 compared to the source-only model, which achieves only 0.82 and 0.88 on UCF → HMDB and HMDB → UCF datasets, respectively. In Figure 7, we show some qualitative retrieval results for the UCF → HMDB. Our model can correctly retrieve the source videos of the same semantic categories as the target query videos. See supplementary for more qualitative and quantitative results.

5 CONCLUSION

In this work, we address the unexplored problem of source-free video domain adaptation and propose a simple yet effective approach CleanAdapt. Our framework is based on self-training in which we generate noisy pseudo-labels for the target domain data using the source pre-trained model. We argue that the presence of noise in the pseudo-labels hinders the adaptation performance and exploit the deep memorization effect to select the clean samples in order to increase the quality of the pseudo-labels. Our method CleanAdapt consistently outperforms the state-of-the-art image-based and video-based UDA methods without any source domain videos. We hope this perspective for video domain adaptation will help approach other domain adaptation settings for videos.

ACKNOWLEDGMENTS

We thank Vinod K Kurmi, Aditya Arun, Aniket Singh, Minesh Mathew, Siddhant Bansal, Trupthi Ann John, Deepayan Das, and Subham Dokania for their feedback. Avijit Dasgupta is supported by a Google Ph.D. India Fellowship. Karteek Alahari is supported in part by the ANR grant AVENUE (ANR-18-CE23-0011).

REFERENCES

[1] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *ICML*.

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

[3] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. 2019. Temporal attentive alignment for large-scale video domain adaptation. In *ICCV*.

[4] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. 2020. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *WACV*.

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

[6] Victor G Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. 2022. Dual-Head Contrastive Domain Adaptation for Video Action Recognition. In *WACV*.

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *ECCV*.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *ICCV*.

[9] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional two-stream network fusion for video action recognition. In *CVPR*.

[10] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. 2021. Can cross entropy loss be robust to label noise?. In *IJCAI*.

[11] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*.

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *NeurIPS* (2018).

[13] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. 2021. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *NeurIPS* (2021).

[14] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. 2018. Deep Domain Adaptation in Action Space.. In *BMVC*.

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[16] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. 2021. Learning Cross-modal Contrastive Features for Video Domain Adaptation. *ICCV* (2021).

[17] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. 2021. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence* (2021).

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

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

[20] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. 2020. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*.

[21] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. 2018. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition* (2018).

[22] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*.

[23] J. Liang et al. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*.

[24] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *ICML*.

[25] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. In *ICML*.

[26] Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. 2020. Generative pseudo-label refinement for unsupervised domain adaptation. In *WACV*.

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

[28] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. 2021. Video transformer network. In *ICCV*.

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

[30] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS* (2019).

[31] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. 2021. Source-free domain adaptation via avatar prototype generation and adaptation. *IJCAI* (2021).

[32] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. 2021. Contrast and Mix: Temporal Contrastive Video Domain Adaptation with Background Mixing. *NeurIPS* (2021).

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

[34] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *NeurIPS* (2014).

[35] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. 2021. Spatio-temporal Contrastive Domain Adaptation for Action Recognition. In *CVPR*.

[36] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[37] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR*.

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

[39] Haoran Wang, Tong Shen, Wei Zhang, Ling-Yu Duan, and Tao Mei. 2020. Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In *ECCV*.

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

[41] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *CVPR*.

[42] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*.

[43] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*.

[44] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*.

[45] Jihan Yang, Shaoshuai Shi, Zhe Wang, Hongsheng Li, and Xiaojuan Qi. 2021. ST3D: Self-training for unsupervised domain adaptation on 3d object detection. In *CVPR*.

[46] S. Yang et al. 2021. Exploiting the Intrinsic Neighborhood Structure for Source-free Domain Adaptation. *NeurIPS* (2021).

[47] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. 2020. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427* (2020).

[48] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption?. In *ICML*.

[49] Christopher Zach, Thomas Pock, and Horst Bischof. 2007. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*.

[50] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*.

[51] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal Relational Reasoning in Videos. *ECCV* (2018).

[52] Xiaojin Jerry Zhu. 2005. Semi-supervised learning literature survey. *Technical Report, University of Wisconsin-Madison, Department of Computer Sciences* (2005).