Unsupervised Video Domain Adaptation with Masked Pre-Training and Collaborative Self-Training

Arun Reddy[1,2], William Paul[1], Corban Rivera[1], Ketul Shah[2], Celso M. de Melo[3], Rama Chellappa[2]
[1]Johns Hopkins University Applied Physics Laboratory
[2]Johns Hopkins University, Department of Electrical & Computer Engineering
[3]DEVCOM U.S. Army Research Laboratory

Abstract

In this work, we tackle the problem of unsupervised domain adaptation (UDA) for video action recognition. Our approach, which we call UNITE, uses an image teacher model to adapt a video student model to the target domain. UNITE first employs self-supervised pre-training to promote discriminative feature learning on target domain videos using a teacher-guided masked distillation objective. We then perform self-training on masked target data, using the video student model and image teacher model together to generate improved pseudolabels for unlabeled target videos. Our self-training process successfully leverages the strengths of both models to achieve strong transfer performance across domains. We evaluate our approach on multiple video domain adaptation benchmarks and observe significant improvements upon previously reported results.

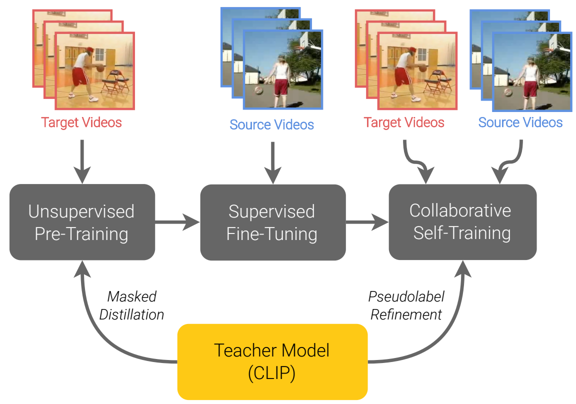

Figure 1. Overview of the UNITE pipeline for video UDA. A teacher model is used to guide the self-supervised learning process in the first stage, and is again used to improve pseudolabeling of target videos during self-training in the third stage.

1. Introduction

In recent years, the field of video action recognition has undergone significant advancement, largely driven by deep learning-based techniques. These include approaches based on convolutional neural networks (CNNs) [6, 14, 15, 47, 51, 53] and vision transformers (ViTs) [1, 5, 13, 33, 39, 42]. Further enriching this landscape, models integrating video and language data have emerged [36, 55, 57], which effectively leverage large collections of captioned videos to achieve impressive video understanding capabilities.

Despite increasing performance on benchmark datasets, deep networks for video action recognition suffer from the same fundamental challenges as other machine learning models in the presence of *distribution shift*. This phenomenon, where a model's performance degrades when applied to data distributions different from those it was trained on, remains a critical obstacle in deploying the models in varied real-world scenarios.

Domain adaptation (DA) seeks to mitigate the effects of distribution shift and has been a major research area within the computer vision community for years [41]. A subproblem within DA is unsupervised domain adaptation (UDA), which tries to make use of unlabeled data from the target domain to improve transfer performance. UDA approaches specific to video have also been developed, with many focusing on explicit alignment of the source and target domains using adversarial techniques or domain discrepancy measures. In contrast, we present here a method for video UDA that relies on self-supervised masked pre-training and self-training, both aided by a powerful image-based teacher model (CLIP [44]). The authors of DALL-V [67] demonstrate the effectiveness of CLIP in the source-free video domain adaptation setting by directly adapting it to action recognition tasks despite fundamentally being an image-based technique. Instead, we leverage CLIP as a spatial teacher for training a spatiotemporal student model that can better operate on target domain videos.

This paper presents, to the best of our knowledge, the first exploration of self-supervised masked distillation techniques in the context of unsupervised video domain adaptation. Our approach, Unsupervised Adaptation with Teacher-Enhanced Learning (UNITE), successfully harnesses the capabilities of an image-based teacher in two ways: (1) to form the target representations in a masked self-supervised pre-training stage on target domain videos, and (2) to improve the quality of pseudolabels used during self-training in collaboration with the student video network (see Figure 1).

We summarize the contributions of our work as follows:

- We introduce the UNITE pipeline for video domain adaptation, which consists of a novel combination of masked video modeling and self-training techniques.
- We conduct extensive experimentation by evaluating UNITE on three video domain adaptation benchmarks (*Daily-DA*, *Sports-DA* and *UCF-HMDB*) and observe significant performance gains compared to previously reported results.
- We present a series of ablation experiments that study the effectiveness of various aspects of the UNITE pipeline and assess alternative design choices. In particular, we show that masked distillation pre-training and masked self-training are best applied in concert to achieve the strongest domain transfer performance.

2. Related Work

Video Unsupervised Domain Adaptation (VUDA). Techniques for video unsupervised domain adaptation typically fall within a few common categories [59]. Adversarial methods attempt to align representations between the source and target domains by enforcing a domain confusion loss using a separate domain discriminator network, where the main network must learn to fool the discriminator. Some techniques which were developed for images, such as DANN [18] or MK-MMD [34], can be applied directly to video architectures. TA^{3}N [8] is a video-specific approach that uses separate adversaries for the spatial and temporal dimensions. These approaches can result in unstable training due to competing objectives.

Another class of methods use contrastive learning and exploit the intrinsic structure of video. CoMix [45] uses contrastive learning with pseudolabeling and video-based augmentations such as sampling at varying frame rates or mixing the backgrounds of source videos and target videos. CO^{2}A [52] optimizes 6 different losses simultaneously, including contrastive losses at the clip- and video-level, along with supervised contrastive learning across domains. UDAVT [10] adapts vision transformers using the information bottleneck principle by having separate projectors for source and target features and estimating a cross correlation matrix between features whose labels and pseudolabels match.

Some video domain adaptation techniques have been developed to address the related problem of source-free video video domain adaptation (SFVUDA), which focuses on scenarios where source data is unavailable during the adaptation phase. ATCoN [62] tries to ensure the trained model is as consistent with its features as the original model. EXTERN [63] additionally uses masking augmentations as a factor for features to be consistent over. DALL-V [67] takes a different approach by adapting a CLIP [44] image encoder to source and target datasets using an adapter.

Masked Image & Video Modeling. Masked modeling first achieved success in natural language processing [11], and has since been applied to image and video data as a way to learn powerful feature encoding from unlabeled data. The basic idea behind masked modeling is to partially occlude patches of the input data and train a model to reconstruct the missing patches. This can be done at the pixel-level [16, 19, 31, 50, 54, 58], or the targets of the reconstruction can be formed by a neural network— either an exponential moving average (EMA) of the network being trained [2–4, 24, 32, 70], or even a separate teacher model. MVD [56] uses the latter concept to train a video transformer network using both image and video teachers, which can be viewed as a form of knowledge distillation [21].

Unmasked Teacher (UMT) [30] uses a different form of masked distillation to train powerful video encoders, where instead of reconstructing representations of masked regions, the student attempts to match the teacher representations of visible patches [66]. UMT is able to use a spatial encoder (CLIP [44]) to train a powerful spatiotemporal model. Because the video encoder only needs to process visible patches, and there is no need for an expensive decoder for reconstruction, UMT offers a highly efficient approach for self-supervision on videos. As such, we choose to leverage the UMT masked distillation objective to perform unsupervised video learning in UNITE.

Masked Modeling for Domain Adaptation. Masked modeling is fundamentally about ensuring a given feature is invariant to whether its corresponding input is masked or not. Inducing such invariances on novel domains can be a useful property to enforce for the purposes of domain adaptation. MAE-TTT [17] finetunes a masked autoencoder [19] on single images at test time, adapting the feature extractor to the statistics of the current image while leaving the prediction head unchanged. MIC [22] uses an EMA teacher model on unmasked images to produce pseudolabels for training on masked target images. PACMAC [43] selects target domain samples for pseudolabeling based on prediction consistency over masked views, and also applies the classification loss to masked target inputs. Since these approaches have been developed for image-based tasks, there is still an open question of how similar concepts could be applied to video recognition.

3. Preliminaries

3.1. Problem Formulation

We formalize the problem of unsupervised domain adaptation for action recognition as follows. Given a set of labeled videos $\mathcal{D}_S := \{(\mathbf{x}_i^S, y_i^S)\}_{i=1}^{N_S}$ drawn from a source data distribution $\mathcal{P}_S$, and a set of unlabeled videos $\mathcal{D}_T := \{\mathbf{x}_i^T\}_{i=1}^{N_T}$ drawn from a target distribution $\mathcal{P}_T$, our objective is to learn a function that can successfully classify videos drawn from $\mathcal{P}_T$. More precisely, we seek to learn the parameters θ of a function f_θ, in our case a transformer neural network, that minimize the empirical risk on a sampling of videos from $\mathcal{P}_T$. The challenge in domain adaptation lies in the fact that $\mathcal{P}_S$ and $\mathcal{P}_T$ differ from one another, leading to a gap in performance when directly applying a source trained model on target domain data. Thus, UDA techniques need to exploit information in unlabeled target domain samples to reduce this gap.

3.2. Self-Supervised Initialization

In contrast to some other video domain adaptation works [10, 65] that use network weights from Kinetics-400 supervised pre-training, we initialize our network from *self-supervised* pre-training. As contended in [43] for UDA with images, we argue that supervised pre-training can potentially complicate the study of VUDA techniques. When the pre-trained network has been trained to classify categories present in the DA dataset, some DA techniques could perform well simply by preserving the capabilities of the pre-trained model despite not generalizing well to DA tasks that do not have category overlap with the pre-training dataset. Moreover, the UDA problem formulation implies that labeled instances of relevant classes are only available from the source domain. This condition does not hold for the *Daily-DA* and *Sports-DA* benchmarks that we use for evaluation, as 6 out of their 8 classes also appear in the Kinetics-400 dataset. Further exacerbating the issue with supervised Kinetics pre-training is the fact that Kinetics serves as a target domain in both benchmarks. Thus, we initialize the network weights from UMT [30] single-modality (*i.e.*, video only) self-supervised pre-training on Kinetics-710 [29]. Please refer to Sec. 10 in the Appendix for select results using a UMT pre-trained network with additional supervised fine-tuning on Kinetics-400, which unsurprisingly achieves better baseline performance.

4. Method

4.1. UNITE

Our approach to VUDA consists of three stages: (1) unsupervised pre-training on target domain data, (2) supervised fine-tuning on source domain data and (3) collaborative self-training using videos from both domains. We now describe each stage in detail.

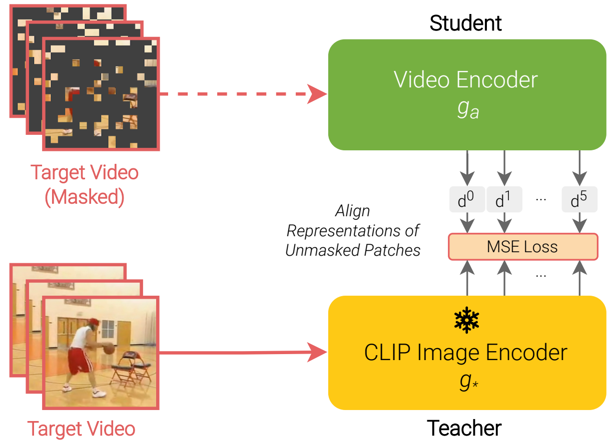

Figure 2. Unmasked Teacher (UMT) training used in Stage 1 of UNITE to perform unsupervised representation learning on target domain videos.

Stage 1: Unsupervised Target Domain Pre-Training. To enable the extraction of discriminative features from target domain videos, the UNITE pipeline begins by performing additional unsupervised pre-training on the target dataset $\mathcal{D}_T$. For this, we adopt the UMT [30] self-supervised training objective. UMT training uses a pre-trained and frozen spatial teacher model g_* to train a spatiotemporal student model g_a by enforcing alignment between feature representations of the two networks at multiple layers (depicted in Figure 2). A key aspect of the training process is that the student model takes as input a masked video view $m(\mathbf{x}^T)$, while the teacher model processes unmasked video frames from $\mathbf{x}^T$. Through UMT training, g_a must learn to match the semantically rich (albeit only spatially informed) features of g_* from heavily masked target domain videos. To ensure sampling of meaningful video patches, UMT performs an attention-guided masking operation $m(\cdot)$ drawn from a multinomial distribution defined by the attention map of the final layer CLS token in g_*. Attention-guided masking allows the use of a high masking ratio, which makes UMT training extremely efficient.
We formalize the Stage 1 objective as follows. Let $\mathbf{z}_a^l$ denote the L2 normalized l-th layer representation of g_a for a masked input $m(\mathbf{x})$, and let $\mathbf{z}_*^l$ denote the normalized l-th layer representation of g_* for the full input $\mathbf{x}$, but only at the locations corresponding to the visible patches in $m(\mathbf{x})$. The loss function in Stage 1 can be represented as:

$$\mathcal{L}_{\text{UMT}} = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_T}\left[\frac{1}{|\mathcal{A}|}\sum_{l\in\mathcal{A}} \text{MSE}\left(d^l(\mathbf{z}_a^l), \mathbf{z}_*^l\right)\right]$$

where $\mathcal{A}$ is a set of aligned layers, $d^l(\cdot)$ is a linear projection

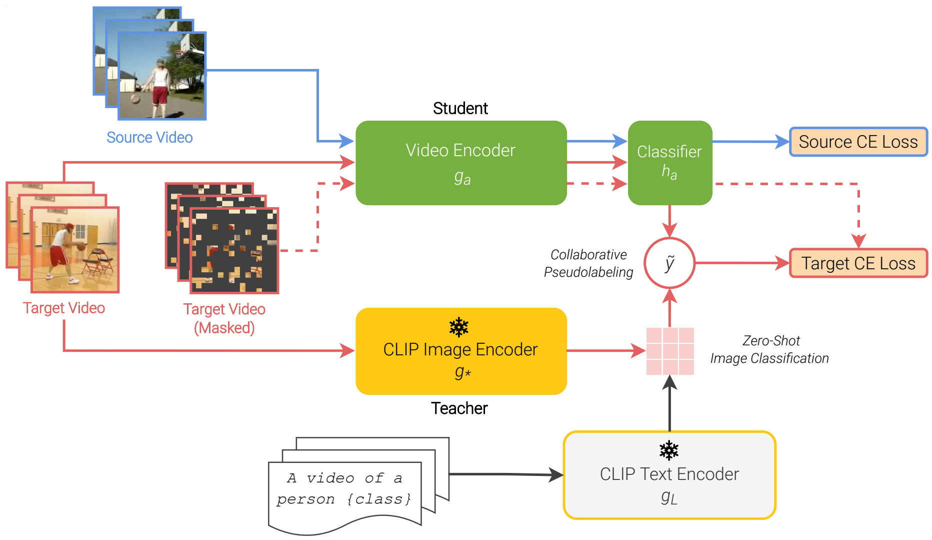

Figure 3. Overview of the collaborative self-training stage in UNITE. The student and teacher models work together to produce more accurate pseudolabels for target domain videos. The target domain classification loss is enforced on masked target videos to encourage stronger context learning. Source domain classification is included to stabilize training, which is especially important at the beginning of training when pseudolabels may have low accuracy.

for the l-th layer and MSE indicates mean squared error. Following the original UMT method, we leverage a CLIP [44] ViT-B image encoder as g_*. We keep the standard UMT masking ratio of $r = 0.8$ and perform alignment between the last 6 layers of the networks. g_a is initialized as described in Sec. 3.2, along with the projections $d^l(\cdot)$, which are kept frozen throughout Stage 1.

Stage 2: Source Domain Fine-Tuning. The second stage of UNITE fits a classifier on source domain data using g_a from Stage 1. We introduce label space information in a separate stage from target domain self-supervision because, as the authors of [17] observe for MAE training, we find that imposing a classification loss and UMT loss simultaneously can hinder convergence. To perform supervised fine-tuning, we introduce a linear classification head h_a that operates on the mean pooled outputs of the final transformer layer in g_a. The full student classifier network can now be represented as $f_a = h_a(g_a(\cdot))$. As is the case with other masked pre-training strategies, we find that representations learned via UMT pre-training are not linearly separable. Thus, we perform full fine-tuning of f_a by minimizing a cross entropy (CE) loss $\mathcal{L}_{CE}$ on unmasked source domain videos. The Stage 2 loss function is represented as:

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(\mathbf{x}^S, y^S) \sim \mathcal{P}_S} \left[\mathcal{L}_{CE}(f_a(\mathbf{x}^S), y^S) \right] \tag{1}$$

To help preserve target domain feature extraction learned in Stage 1, we apply a layer-wise learning rate decay when fine-tuning on source domain data.

Stage 3: Collaborative Self-Training (CST). After enhancing the ability of g_a to extract meaningful features in the target domain (Stage 1) and fitting f_a for the classification task using source videos (Stage 2), we further adapt f_a to target data using a self-training process (depicted in Figure 3). In a conventional self-training framework (denoted as $\mathcal{L}_{\text{ST}}$ in Eq. (2)), the overall loss is a sum of losses on labeled source domain samples and select target domain samples, chosen using mask $s(\cdot)$. The target domain loss leverages the model's own predictions $\hat{y}$ (where $\hat{y} = \text{argmax}_c\, \sigma(f(\mathbf{x}^T))[c]$ and $\sigma(\cdot)$ denotes the softmax function) in lieu of labels in the CE loss computation. Various strategies have been used to form the selection mask $s(\cdot)$, with the goal of choosing target samples whose predictions are more likely to be correct [43, 48, 68].

$$\begin{aligned} \mathcal{L}_{\text{ST}} =& \mathbb{E}_{(\mathbf{x}^S, y^S) \sim \mathcal{P}_S} \left[\mathcal{L}_{CE}(f_a(\mathbf{x}^S), y^S) \right] + \\ & \lambda \mathbb{E}_{\mathbf{x}^T \sim \mathcal{P}_T} \left[s(\mathbf{x}^T) \mathcal{L}_{CE}(f_a(\mathbf{x}^T), \hat{y}_a(\mathbf{x}^T)) \right] \end{aligned} \tag{2}$$

We introduce a modified form of self-training, which we refer to as collaborative self-training (CST), that makes use of the student-teacher setup from Stage 1 to improve the process of pseudolabel estimation. Specifically, we use the zero-shot classification capabilities of the CLIP image teacher model f_* (see Sec. 5.2 for details), in combination with predictions from the video student model f_a, to create refined target pseudolabels for self-training. We employ the MatchOrConf scheme proposed in [69] to combine the outputs of the two models as shown in Eq. (3), where $\text{Conf}(\hat{y})$ denotes the maximum softmax probability

[20], $\max_c \sigma(f(\mathbf{x}^T))[c]$, to create pseudolabel $\tilde{y}$:

$$\tilde{y} = \begin{cases} \hat{y}_a & \text{if } \hat{y}_a = \hat{y}_* \\ \hat{y}_a & \text{if } \hat{y}_a \neq \hat{y}_* \text{ and } \mathrm{Conf}(\hat{y}_a) > \gamma \text{ and } \mathrm{Conf}(\hat{y}_*) \leq \gamma, \\ \hat{y}_* & \text{if } \hat{y}_a \neq \hat{y}_* \text{ and } \mathrm{Conf}(\hat{y}_a) \leq \gamma \text{ and } \mathrm{Conf}(\hat{y}_*) > \gamma, \\ -1 & \text{otherwise.} \end{cases} \tag{3}$$

$\mathbf{x}^T$ is an argument of the above, but is omitted for clarity. The MatchOrConf scheme selects only those target samples where there is agreement between the predictions of the two models, or where one model's confidence exceeds threshold γ while the other's does not, resulting in the following selection mask:

$$s(\mathbf{x}^T) = \begin{cases} 1 & \text{if } \tilde{y}(\mathbf{x}^T) \neq -1, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Taking inspiration from [22] and [43], we compute the target domain loss on masked videos, where we employ the same teacher-guided attention masking strategy (with $r = 0.8$) used during UMT pre-training in Stage 1. We find that training on masked videos in this stage results in substantial performance boosts compared to training on unmasked videos (see Table 6). Like [22], we include a loss weighting q for each target video, based on the student network confidence:

$$q(\mathbf{x}^T) = \mathrm{Conf}(\hat{y}_a(\mathbf{x}^T)) \tag{5}$$

The complete loss function for the collaborative self-training stage can be expressed as:

$$\begin{aligned} \mathcal{L}_{\text{CST}} =& \mathbb{E}_{(\mathbf{x}^S, y^S) \sim \mathcal{P}_S} \left[\mathcal{L}_{CE}(f_a(\mathbf{x}^S), y^S) \right] + \\ & \lambda \mathbb{E}_{(\mathbf{x}^T) \sim \mathcal{P}_T} \left[s(\mathbf{x}^T) q(\mathbf{x}^T) \mathcal{L}_{CE}(f_a(m(\mathbf{x}^T)), \tilde{y}(\mathbf{x}^T) \right] \end{aligned} \tag{6}$$

5. Experiments

5.1. Datasets

We evaluate our approach on three video domain adaptation benchmarks: *Daily-DA*, *Sports-DA*, *UCF*↔*HMDB*$_{full}$.

- ***Daily-DA*** [64] is a recently introduced benchmark composed of videos from four domains: ARID (A) [60], HMDB51 (H) [28], Moments-in-Time (M) [38] and Kinetics (K) [7, 26]. It consists of 8 overlapping classes representing everyday actions, with a total volume of approximately 19K videos. Videos from the ARID domain present a unique challenge in this benchmark, as they were explicitly filmed under low-illumination conditions.
- ***Sports-DA*** [64] contains videos from three domains: Sports-1M (S) [25], UCF101 (U) [49] and Kinetics (K) [7, 26]. It includes 23 overlapping categories depicting various sporting activities, with a total volume of approximately 41K videos.
- ***UCF*↔*HMDB*$_{full}$** [8] contains approximately 3.2K videos across 12 action categories taken from the UCF101 (U) [49] and HMDB51 (K) [28] datasets.

5.2. Implementation Details

Frame Sampling. During training, we perform random uniform frame sampling [53] on each video. We divide the video into $T = 8$ uniform segments and randomly sample one frame from each. All frames are resized to 224 × 224 pixels, resulting in a 3 × 8 × 224 × 224 tensor for each video. During testing, we average the network predictions across 3 spatial crops and 4 temporal clips, where clip i is formed by splitting each of the T segments into 4 sub-segments and sampling a frame from the i-th sub-segment of each.

Network Architecture & Training. All of our experiments use a standard ViT-B/16 [12] architecture with a fixed sine and cosine positional encoding scheme. Our patch embedding projection operates on each frame independently, and self-attention is performed across both the space and time dimensions. All classification is performed using mean pooling of the final layer patch representations. We conduct our experiments using the PyTorch deep learning library [40] on 4 NVIDIA A5000 GPUs.

During Stage 1, we perform UMT training for 50 epochs on the target dataset using a batch size of 256 and a learning rate of 1.5e-5. During Stage 2, we use a batch size of 28, a learning rate of 2.5e-5, and apply a layer-wise learning rate decay of 0.65. During stage 3, we use a batch size of 40 (20 videos from each domain), a learning rate of 1e-5, a MatchOrConf confidence threshold of $\gamma = 0.1$ and a target domain loss weight $\lambda = 1$. Training in all stages uses the AdamW optimizer [35] and a weight decay of 0.05. Please refer to Sec. 8 in the Appendix for more training details for each stage.

Zero-Shot Classification with CLIP. In Stage 3 of UNITE, we use CLIP (ViT-B/16) in a zero-shot manner to refine the pseudolabels of target domain videos. We do so by passing each of the T video frames through the CLIP image encoder and computing the cosine similarity with a set of text representations produced by the CLIP text encoder. The inputs to the text encoder are the class names associated with the particular action recognition task, embedded in the following template: "*A video of a person* {class}." To form a video-level prediction, we simply average the softmax of the cosine similarity scores across the video frames. Refer to the Appendix (Sec. 11) for a listing of class names used for each task.

	Method	Target Domain Accuracy (Top-1%)												
		H→A	M→A	K→A	A→H	M→H	K→H	H→M	A→M	K→M	M→K	H→K	A→K	**Avg.**
	DALL-V [67] LB	17.5	34.7	15.6	14.6	44.6	47.9	25.5	15.5	35.7	61.6	45.1	17.8	31.3
	Source Only	40.4	**52.1**	36.5	49.6	68.3	57.9	41.5	36.3	43.3	79.3	48.0	41.7	49.6
ZS	CLIP (ViT-B/16) [44]	36.5	36.5	36.5	60.0	60.0	60.0	48.5	48.5	**48.5**	68.1	68.1	68.1	53.3
SFVUDA	ATCoN [62]	17.9	27.2	17.2	26.7	47.3	48.2	30.7	17.2	32.5	57.7	48.5	31.0	33.5
SFVUDA	EXTERN [63]	26.2	18.1	23.9	26.2	53.7	55.8	40.7	18.2	35.2	68.1	57.6	51.4	39.6
SFVUDA	DALL-V [67]	24.0	24.0	24.0	57.9	65.4	52.5	47.0	45.7	47.0	78.1	**76.7**	**75.0**	51.4
UDA	DANN [18]	14.2	22.8	21.2	20.1	43.3	37.5	29.5	19.7	21.7	58.8	38.2	27.0	29.5
UDA	MK-MMD [34]	20.3	21.0	21.7	18.7	50.4	36.2	25.7	18.0	24.0	58.5	33.8	26.1	29.5
UDA	TA^{3}N [8]	14.4	21.6	19.9	14.9	43.0	37.7	25.7	15.6	31.5	55.5	38.4	23.4	28.5
UDA	UNITE w/o CST	43.8	46.7	34.7	51.7	70.8	54.6	44.3	39.0	43.3	78.1	51.7	50.8	50.8
UDA	UNITE (Ours)	**48.0**	44.1	**37.5**	**67.9**	**74.2**	**65.8**	**51.8**	**50.0**	48.0	**89.9**	69.9	63.6	**59.2**
	DALL-V [67] UB	26.9	26.9	26.9	70.4	70.4	70.4	61.5	61.5	61.5	88.9	88.9	88.9	61.9
	Target Only	68.5	68.5	68.5	84.6	84.6	84.6	73.0	73.0	73.0	98.3	98.3	98.3	81.1

Table 1. UDA results on *Daily-DA*. Rows in color use our UMT pre-trained backbone, and rows without color are reported from [61].

	Method	Target Domain Accuracy (Top-1 %)						
		U→S	K→S	S→U	K→U	U→K	S→K	**Avg.**
	DALL-V [67] LB	64.3	79.5	84.4	85.4	67.2	78.2	76.5
	Source Only	67.6	86.8	97.9	98.9	79.9	89.1	86.7
ZS	CLIP (ViT-B/16) [44]	85.0	85.0	93.3	93.3	91.3	91.3	89.9
SFVUDA	ATCoN [62]	47.9	69.7	90.6	93.6	65.2	76.0	73.8
SFVUDA	EXTERN [63]	72.7	73.8	95.4	93.7	81.2	82.2	83.2
SFVUDA	DALL-V [67]	75.9	77.7	88.8	88.0	81.2	82.3	82.3
UDA	DANN [18]	55.1	75.0	85.7	88.0	65.9	73.4	73.8
UDA	MK-MMD [34]	55.6	67.9	90.9	90.2	66.1	73.6	74.0
UDA	TA^{3}N [8]	54.1	68.6	93.0	90.3	63.6	72.6	73.7
UDA	UNITE w/o CST	73.9	87.3	98.2	99.2	84.6	90.2	88.9
UDA	UNITE (Ours)	**86.0**	**90.2**	**98.7**	**99.8**	**94.2**	**95.3**	**94.0**
	DALL-V [67] UB	88.3	88.3	93.4	93.4	85.6	85.6	89.1
	Target Only	97.9	97.9	98.8	98.8	99.9	99.9	98.9

Table 2. UDA results on *Sports-DA*. Rows in color use our UMT pre-trained backbone, and rows without color are reported from [61].

5.3. Baselines

We compare results using UNITE with previously developed techniques in UDA, as well as in the closely related setting of SFUDA as reported by [67]. For *Daily-DA* and *Sports-DA* we compare against DANN [18], MK-MMD [34] and TA^{3}N [8] in the UDA setting, and ATCoN [62], EXTERN [63] and DALL-V [67] for SFUDA. For $UCF{\leftrightarrow}HMDB_{full}$, we additionally report accuracies from CO^{2}A [52] and UDAVT [10].

5.4. Benchmark Results

Tables 1, 2 and 3 show the results of our evaluation of UNITE on *Daily-DA*, *Sports-DA* and $UCF{\leftrightarrow}HMDB_{full}$. We include accuracies after Stage 2 of the pipeline (denoted as UNITE w/o CST) as well as after applying the full UNITE process. For reference, we include accuracies using our UMT pre-trained network with source domain supervised fine-tuning (Source Only) and with target domain supervised fine-tuning (Target Only). We also include the source only and target only accuracies (denoted in the tables as LB and UB) of DALL-V [67] (a CLIP-based ResNet-50 model) for added context. We additionally display the zero-shot classification accuracy of the CLIP teacher model that we utilize during Stage 1 and Stage 3 of UNITE.

An observation we make across all benchmarks is the enhanced baseline accuracy of the UMT pre-trained model, which speaks to the strength of the unsupervised masked distillation process and the ViT architecture. On the *Daily-DA* benchmark, we find that UNITE exceeds previously reported results on most domain shifts. On only one shift (M→A) we observe degradation in target accuracy after employing UNITE compared to source only, which is consistent with previous work.

UNITE exhibits even stronger performance on the larger-scale *Sports-DA* benchmark, where it exceeds previous results by a large margin. We also see more consistent and

	Method	Accuracy (%) H→U	U→H	Avg.
	DALL-V [67] LB	71.6	76.1	73.8
	Source Only	88.8	80.0	84.4
ZS	CLIP (ViT-B/16) [44]	88.8	91.7	90.3
SFVUDA	ATCoN [62]	85.3	79.7	82.5
	EXTERN [63]	91.9	88.9	90.4
	DALL-V [67]	93.1	88.9	91.0
UDA	DANN [18]	74.4	75.1	74.8
	MK-MMD [34]	74.7	79.7	77.2
	TA^3N [8]	78.1	84.8	81.5
	CO^2A [52]	95.8	87.8	91.8
	UDAVT [10]	**96.8**	92.3	**94.6**
	UNITE w/o CST	92.1	83.6	87.9
	UNITE (Ours)	92.5	**95.0**	93.8
	DALL-V [67] UB	93.7	91.4	92.6
	Target Only	99.7	96.7	98.2

Table 3. UDA results on *UCF↔HMDB*$_{full}$. Rows in color use our UMT backbone, and rows without color are reported from [61].

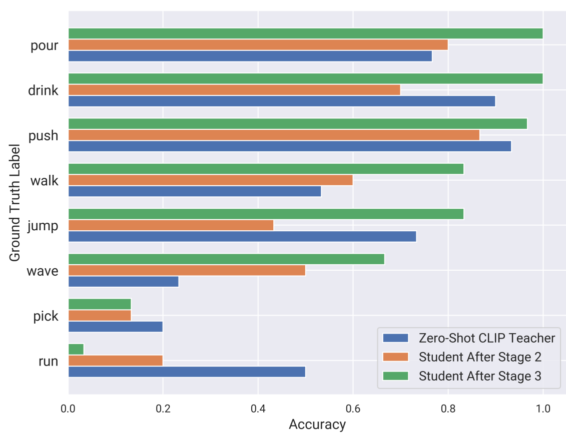

Figure 4. Class-wise accuracy on ARID→HMDB. Performances are shown for the CLIP teacher model and the student model before and after the collaborative self-training stage (Stage 3). For most classes, the Stage 3 student exceeds the accuracy of both the Stage 2 student and the CLIP teacher.

substantial improvement after the UMT target pre-training stage alone (UNITE w/o CST), suggesting that overtraining may be occurring on the smaller datasets in *Daily-DA*.

On *UCF↔HMDB*$_{full}$ we see state-of-the-art performance on U→H, while UDAVT outperforms on H→U. Nevertheless, UNITE still achieves strong overall performance on the benchmark.

6. Additional Analysis & Discussion

Student exceeds the teacher. We see in the benchmark results that the video student model exceeds the zero-shot accuracy of the CLIP teacher in virtually all domain shifts. This suggests that UNITE successfully leverages the spatial modeling capabilities of CLIP to train an even more powerful spatiotemporal model for the action recognition task.

Method	PT	CST	Accuracy (%) H→A	A→H
Source Only	✗	✗	40.4	49.6
+ Target UMT Pre-Training	✓	✗	43.8	51.7
+ Collaborative Self-Training	✗	✓	42.0	60.8
UNITE	✓	✓	**48.0**	**67.9**

Table 4. Ablation of stages in UNITE on ARID↔HMDB from *Daily-DA*. PT and CST denote UMT pre-training and collaborative self-training, respectively.

We can see from Figure 4 that the teacher model and the student model prior to collaborative self-training have differing capabilities. For example, the image teacher is stronger at recognizing actions from the *jump* class than the Stage 2 student. On the other hand, the video student is significantly better than the teacher at recognizing *wave* at the start of CST. In both cases, we see that the student model, through collaborative self-training, comes to exceeds both the Stage 2 student and the CLIP image teacher. With the MatchOrConf pseudolabeling scheme, the two models are able to work together to achieve higher accuracy in the target domain. We see accuracy on the *pick* and *run* classes declines after CST, as self-training tends to reinforce the accuracy of well-performing classes, sometimes at the expense of more difficult ones that are infrequently assigned pseudolabels. Modifying CST to account for the learning status of each class, like in [68], may help remedy this issue.

Impact of Stage 1 and 3. In Table 4 we ablate the stages in UNITE. While UMT pre-training and collaborative self-training each offer a benefit in the absence of the other, the largest gains in domain transfer performance are obtained when combining the two stages. For example, on the ARID→HMDB domain shift in *Daily-DA*, Stage 1 and Stage 3 alone improve upon source only accuracy by +1.8% and +11.2% respectively. Meanwhile, the full UNITE pipeline results in +18.3% accuracy improvement on the target domain, suggesting that the masked pre-training and masked self-training processes complement one another.

Impact of data domains during pre-training. In Table 5, we assess the impact of utilizing each of the two data domains during UMT pre-training on UDA performance. While the results vary across domain shifts, we observe the most consistent improvements when pre-training only on target domain videos. Although incorporating both domains during unsupervised pre-training has been successfully applied in image-based domain adaptation [27, 37, 43, 46], we find it to be suboptimal with UMT training on the VUDA tasks studied in this work.

Effect of masking during self-training. In Table 6, we see that enforcing the target domain CE loss on masked videos results in increased performance compared to using unmasked target videos, with smaller benefits for H→A and much larger benefits for A→H. We believe the benefits of

UMT Pre-Training Data	Accuracy (%) H→A	A→H
None	42.0	60.8
Source Only	**48.3**	60.0
Target Only	<u>48.0</u>	**67.9**
Source + Target	45.2	<u>67.1</u>

Table 5. Comparison of data domains used during UMT pre-training. Accuracies are reported after the self-training stage. **Bold** indicates best accuracy and <u>underline</u> indicates second best.

Data for Target CE Loss	Accuracy (%) H→A	A→H
Unmasked	46.6	54.6
Masked	**48.0**	**67.9**

Table 6. Effect of applying the target domain cross entropy loss on masked vs. unmasked target videos during the self-training stage.

Pseudolabeling Strategy	Accuracy (%) H→A	A→H
Cons ($k = 2$)	39.8	57.5
ConsOrConf ($k = 2, T = 0.5$) [43]	**48.6**	56.3
ConsAndConf ($k = 2, T = 0.5$)	40.1	57.1
Zero-Shot CLIP Only ($T = 0.5$)	34.7	51.3
MatchOrConf ($\gamma = 0.1$) [69]	<u>48.0</u>	**67.9**
MatchAndConsOrConf ($k = 2, \gamma = 0.1$)	47.4	<u>60.8</u>

Table 7. Comparison of pseudolabeling strategies.

masking arise from forcing the model to use different cues from the same video to predict the pseudolabel, resulting in more robust target domain recognition. As we saw in Table 4, the masked pre-training stage plays a key role in unlocking the benefits of masked self-training.

Comparison of pseudolabeling strategies. In Table 7, we explore the use of alternative pseudolabeling strategies to the MatchOrConf [69] scheme employed in UNITE. The first 3 rows of the table assess variations of the masked consistency pseudolabeling scheme proposed in PACMAC [43], where target samples are selected if there is agreement between predictions of k masked views. This criterion can also be combined with a constraint on the confidence exceeding threshold T. While we find the combination of these two constraints (ConsOrConf) to be effective for self-training, we find that MatchOrConf results in more consistent target accuracy improvements across domain shifts. We also consider combining MatchOrConf with a masked consistency constraint as a way to reduce cases where the student and teacher models agree on an incorrect prediction. This scheme, denoted as MatchAndConsOrConf, adds an additional constraint to the matching criterion by also requiring agreement between the student model predictions of k masked views. We find that this approach underperforms compared to the vanilla MatchOrConf scheme due to low utilization of target domain samples. Our results also indicate that using both models together for pseudolableing outperforms using only zero-shot CLIP predictions.

Loss During Self-Training	Accuracy (%) H→A	A→H
$\mathcal{L}_{\text{CE}}(m(\mathbf{x}^T), \tilde{y})$	33.7	42.1
$\mathcal{L}_{\text{CE}}(m(\mathbf{x}^T), \tilde{y}) + \mathcal{L}_{\text{CE}}(\mathbf{x}^S, y^S)$	**48.0**	**67.9**

Table 8. Effect of including a source domain classification loss in addition to the target domain loss during the self-training stage.

Confidence Threshold	Accuracy (%) H→A	A→H
0.1	<u>48.0</u>	**67.9**
0.3	44.1	55.4
0.5	**51.9**	64.6
0.7	42.6	<u>65.8</u>
0.9	40.7	62.5

Table 9. Effect of confidence threshold γ in MatchOrConf pseudolabeling scheme from Equation 3.

Effect of source classification during self-training. In Table 8, we study the importance of including a source domain classification loss while performing collaborative self-training on masked target videos. Indeed, we find that source classification is a crucial ingredient in this stage, as excluding it destabilizes training and results in performance degradation rather than improvement. We suspect that the ground truth labels on source data help to combat the effects of inaccurate target pseudolabels, which may be prevalent especially at the start of training.

Choice of pseudolabeling confidence threshold. In Table 9, we compare adaptation performance for various values of confidence threshold in the MatchOrConf pseudolabeling scheme. We observe significant variation across domain shifts, but find that $\gamma = 0.1$ leads to relatively consistent improvement. We leave for future work the development of a more principled approach to determining an appropriate confidence threshold for a given domain shift.

7. Conclusions

In this work, we presented UNITE, a three step process for unsupervised video domain adaptation. Our approach leverages a powerful spatial encoder to aid in the adaptation of a spatiotemporal network using unlabeled target domain videos— making significant progress towards bridging the domain gaps in various VUDA datasets. The advancements demonstrated by UNITE underscore the untapped potential of masked modeling techniques for video domain adaptation, which we hope will spur further research in this area.

Acknowledgements. This research was sponsored by the Army Research Laboratory under Cooperative Agreement W911NF-21-2-0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *ICCV*, 2021. 1

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael Rabbat, and Nicolas Ballas. Masked Siamese Networks for Label-Efficient Learning. In *ECCV*, 2022. 2

[3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. In *CVPR*, 2023.

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *ICML*, 2022. 2

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *ICML*, pages 813–824, 2021. 1

[6] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017. 1

[7] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A Short Note about Kinetics-600, 2018. 5

[8] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal Attentive Alignment for Large-Scale Video Domain Adaptation. In *ICCV*, 2019. 2, 5, 6, 7

[9] Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. 2020 ieee. In *CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2019. 1

[10] Victor G. Turrisi Da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. Unsupervised Domain Adaptation for Video Transformers in Action Recognition. In *ICPR*, pages 1258–1265, Montreal, QC, Canada, 2022. 2, 3, 6, 7

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021. 5

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. In *ICCV*, 2021. 1

[14] Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *CVPR*, 2020. 1

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *ICCV*, 2019. 1

[16] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners. In *NeurIPS*, 2022. 2

[17] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. Test-time training with masked autoencoders. In *NeurIPS*, 2022. 2, 4

[18] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2, 6, 7

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 2021. 2

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017. 5

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. In *NeurIPS*, 2014. 2

[22] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation. In *CVPR*, pages 11721–11732, Vancouver, BC, Canada, 2023. 2, 5

[23] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661. Springer, 2016. 1

[24] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to Hide from Your Students: Attention-Guided Masked Image Modeling. In *ECCV*, pages 300–318, 2022. 2

[25] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR*, pages 1725–1732, 2014. 5

[26] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset, 2017. 5

[27] Donghyun Kim, Kuniaki Saito, Tae-Hyun Oh, Bryan A Plummer, Stan Sclaroff, and Kate Saenko. Cds: Cross-domain self-supervised pre-training. In *ICCV*, pages 9123–9132, 2021. 7

[28] H Kuehne, H Jhuang, E Garrote, T Poggio, and T Serre. HMDB: A Large Video Database for Human Motion Recognition. In *ICCV*, 2011. 5

[29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Unlocking the potential of image vits for video understanding. In *ICCV*, pages 1632–1643, 2023. 3

[30] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked Teacher: Towards Training-Efficient Video Foundation Models. In *ICCV*, 2023. 2, 3

[31] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. MST: Masked Self-Supervised Transformer for Visual Representation. In *NeurIPS*, pages 13165–13176, 2021. 2

[32] Han Lin, Guangxing Han, Jiawei Ma, Shiyuan Huang, Xudong Lin, and Shih-Fu Chang. Supervised masked knowledge distillation for few-shot transformers. In *CVPR*, pages 19649–19659, 2023. 2

[33] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *CVPR*, pages 3192–3201, New Orleans, LA, USA, 2022. 1

[34] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105. PMLR, 2015. 2, 6, 7

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 5, 1

[36] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: End-to-End Multi-grained Contrastive Learning for Video-Text Retrieval. In *ACM MM*, 2022. 1

[37] Samarth Mishra, Kate Saenko, and Venkatesh Saligrama. Surprisingly Simple Semi-Supervised Domain Adaptation with Pretraining and Consistency, 2021. 7

[38] Mathew Monfort, Carl Vondrick, Aude Oliva, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, and Dan Gutfreund. Moments in Time Dataset: One Million Videos for Event Understanding. *PAMI*, 42(2):502–508, 2020. 5

[39] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video Transformer Network. In *ICCV*, 2021. 1

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 5

[41] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual Domain Adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. 1

[42] Aj Piergiovanni, Weicheng Kuo, and Anelia Angelova. Rethinking Video ViTs: Sparse Video Tubes for Joint Image and Video Learning. In *CVPR*, pages 2214–2224, Vancouver, BC, Canada, 2023. 1

[43] Viraj Prabhu, Sriram Yenamandra, Aaditya Singh, and Judy Hoffman. Adapting self-supervised vision transformers by probing attention-conditioned masking consistency. *NeurIPS*, 35:23271–23283, 2022. 2, 3, 4, 5, 7, 8, 1

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 1, 2, 4, 6, 7

[45] Aadarsh Sahoo, Rutav Shah, R. Panda, Kate Saenko, and Abir Das. Contrast and Mix: Temporal Contrastive Video Domain Adaptation with Background Mixing. In *NeurIPS*, 2021. 2

[46] Kendrick Shen, Robbie M. Jones, Ananya Kumar, Sang Michael Xie, Jeff Z. Haochen, Tengyu Ma, and Percy Liang. Connect, Not Collapse: Explaining Contrastive Learning for Unsupervised Domain Adaptation. In *ICML*, pages 19847–19878, 2022. 7

[47] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *NeurIPS*, 2014. 1

[48] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In *NeurIPS*, pages 596–608, 2020. 4

[49] K. Soomro, A. Zamir, and M. Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *ArXiv*, 2012. 5

[50] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *NeurIPS*, 2022. 2

[51] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning Spatiotemporal Features with 3D Convolutional Networks. In *ICCV*, pages 4489–4497, Santiago, Chile, 2015. 1

[52] 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*, pages 2234–2243, 2022. 2, 6, 7

[53] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks for Action Recognition in Videos. *PAMI*, 2017. 1, 5

[54] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *CVPR*, pages 14549–14560, Vancouver, BC, Canada, 2023. 2

[55] Mengmeng Wang, Jiazheng Xing, and Yong Liu. ActionCLIP: A New Paradigm for Video Action Recognition, 2021. 1

[56] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu–Gang Jiang. Masked Video Distillation: Rethinking Masked Feature Modeling for Self-supervised Video Representation Learning. In *CVPR*, pages 6312–6322, Vancouver, BC, Canada, 2023. 2

[57] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. InternVideo: General Video Foundation Models via Generative and Discriminative Learning, 2022. 1

[58] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling. In *CVPR*, 2022. 2

[59] Yuecong Xu, Haozhi Cao, Zhenghua Chen, Xiaoli Li, Lihua Xie, and Jianfei Yang. Video Unsupervised Domain Adaptation with Deep Learning: A Comprehensive Survey, 2022. 2

[60] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. ARID: A New Dataset for Recognizing Action in the Dark. In *Communications in Computer and Information Science*, 2022. 5

[61] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. Aligning Correlation Information for Domain Adaptation in Action Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2022. 6, 7

[62] 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*, pages 147–164, Cham, 2022. 2, 6, 7

[63] Yuecong Xu, Jianfei Yang, Min Wu, Xiaoli Li, Lihua Xie, and Zhenghua Chen. Extern: Leveraging endo-temporal regularization for black-box video domain adaptation. *arXiv preprint arXiv:2208.05187*, 2022. 2, 6, 7

[64] Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Min Wu, Zhengguo Li, and Zhenghua Chen. Multi-Source Video Domain Adaptation With Temporal Attentive Moment Alignment Network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3860–3871, 2023. 5

[65] Yuecong Xu, Jianfei Yang, Yunjiao Zhou, Zhenghua Chen, Min Wu, and Xiaoli Li. Augmenting and Aligning Snippets for Few-Shot Video Domain Adaptation. In *ICCV*, 2023. 3

[66] Hongwei Xue, Peng Gao, Hongyang Li, Yu Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. Stare at What You See: Masked Image Modeling without Reconstruction. In *CVPR*, pages 22732–22741, Vancouver, BC, Canada, 2023. 2

[67] Giacomo Zara, Alessandro Conti, Subhankar Roy, Stephane Lathuiliere, Paolo Rota, and Elisa Ricci. The Unreasonable Effectiveness of Large Language-Vision Models for Source-Free Video Domain Adaptation. In *ICCV*, 2023. 1, 2, 6, 7

[68] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *NeurIPS*, 34:18408–18419, 2021. 4, 7

[69] Wenyu Zhang, Li Shen, and Chuan-Sheng Foo. Rethinking the Role of Pre-Trained Networks in Source-Free Domain Adaptation. In *CVPR*, 2023. 4, 8

[70] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT Pre-training with Online Tokenizer. In *ICLR*, 2022. 2