CDFSL-V: Cross-Domain Few-Shot Learning for Videos

Sarinda Samarasinghe
sarinda.samarasinghe@ucf.edu

Mamshad Nayeem Rizve
mamshadnayeem.rizve@ucf.edu

Navid Kardan
nkardan@cs.ucf.edu

Mubarak Shah
shah@crcv.ucf.edu

Center for Research in Computer Vision
University of Central Florida, Orlando, Florida, USA

Abstract

Few-shot video action recognition is an effective approach to recognizing new categories with only a few labeled examples, thereby reducing the challenges associated with collecting and annotating large-scale video datasets. Existing methods in video action recognition rely on large labeled datasets from the same domain. However, this setup is not realistic as novel categories may come from different data domains that may have different spatial and temporal characteristics. This dissimilarity between the source and target domains can pose a significant challenge, rendering traditional few-shot action recognition techniques ineffective. To address this issue, in this work, we propose a novel cross-domain few-shot video action recognition method that leverages self-supervised learning and curriculum learning to balance the information from the source and target domains. To be particular, our method employs a masked autoencoder-based self-supervised training objective to learn from both source and target data in a self-supervised manner. Then a progressive curriculum balances learning the discriminative information from the source dataset with the generic information learned from the target domain. Initially, our curriculum utilizes supervised learning to learn class discriminative features from the source data. As the training progresses, we transition to learning target-domain-specific features. We propose a progressive curriculum to encourage the emergence of rich features in the target domain based on class discriminative supervised features in the source domain. We evaluate our method on several challenging benchmark datasets and demonstrate that our approach outperforms existing cross-domain few-shot learning techniques. Our code is available at https://github.com/Sarinda251/CDFSL-V

1. Introduction

Even though deep learning is inspired by the biological brain, in sharp contrast to humans, current deep models rely on large reservoirs of data to learn. The few-shot learning problem [41] is introduced to close this gap, where a learning model should generalize solely based on a handful of training data. In traditional few-shot learning [6], the learning model is initially exposed to an annotated *base dataset*, to learn generic features for the domain of interest. Then, this model is fine-tuned on a few labeled examples (support samples) of the test dataset and consequently evaluated on unlabeled test examples (query samples). However, this classic pipeline assumes the base and test datasets are from the same domain, thus closely related [35].

To mitigate this shortcoming, cross-domain few-shot learning (CDFSL) was proposed in [12], where the base dataset is from a different domain than the test data. Interestingly, it is shown in [25] that standard transfer learning— consisting of pre-training on the base dataset and fine-tuning on test data— can significantly outperform few-shot learning methods in the cross-domain few-shot learning problem. Recently, extra unlabeled test examples were incorporated in addition to the base dataset in [28, 16] . Their approaches push forward cross-domain few-shot learning performance. In this paper, we follow this recent adaptation of CDFSL.

While few-shot learning is widely studied in the computer vision community [26], video few-shot learning is less explored [4]. To the best of our knowledge, current methods in cross-domain few-shot learning are solely focused on image data. In this work, for the first time, we study cross-domain few-shot learning in the video domain. A common scheme in video few-shot learning [45] utilizes an implicit assumption about video data, such as: a common mode of variation, similar temporal dynamics, or class distinctive features. However, in cross-domain few-shot learning, the base dataset can be drastically different from the

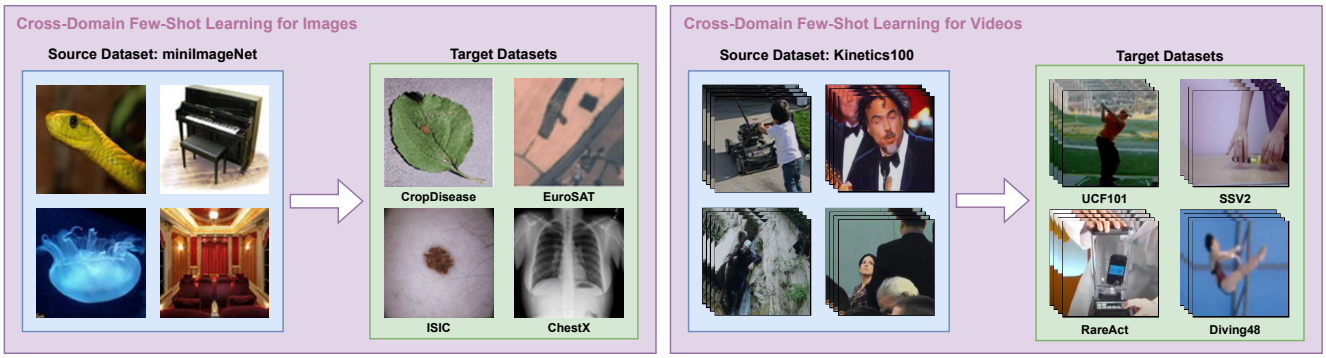


Figure 1: On the left, we have the existing benchmark for CDFSL in the image domain. On the right, we present our proposed benchmark for CDFSL in the video domain. Our benchmark includes tasks from diverse target datasets, which require recognizing novel actions from different data distributions (UCF101, HMDB51), strong temporal reasoning (SSV2), atypical action understanding (RareAct), and fine-grained temporal understanding (Diving48).

target data, For instance, the RareAct dataset [22] contains atypical actions which significantly deviate from the common actions present in the standard video datasets in terms of spatio-temporal dynamics, and the Diving48 dataset [21] contains temporally fine-grained actions which have very similar spatial layout. Therefore, it is challenging to apply standard video few-shot learning methods to these datasets.

In the context of cross-domain few-shot learning, *supervised* pre-training on the source dataset has emerged as a common first step for most techniques [28, 16]. This is because a strong source backbone can significantly contribute to the overall performance of the model [25]. However, simply relying on supervised pre-training may not be sufficient, especially when the target domain is substantially different from the source domain. To address this, in this work we propose to perform *self-supervised* pre-training on both source and target data to learn generic features. To be particular, we use recently proposed masked auto-encoder based [36] feature learning to learn generic features which are highly scalable and show better generalization performance. Nevertheless, the challenge remains on how to balance the learning of generic features (from source and target domain) and class discriminative features from the source dataset.

To this end, we propose a curriculum learning scheme by designing a progressive curriculum that balances learning the discriminative information from the source dataset with the generic information learned from the target domain. In the initial phase of the training, our curriculum utilizes supervised cross-entropy loss to learn class discriminative features from the source data. As the training progresses, we strive to transition to the target domain through learning discriminative features in the target domain. To achieve this, we devise a schedule that increases the weight of a consistency loss to help with this transition. We conduct extensive experiments to demonstrate the effectiveness of our proposed approach on various benchmark datasets. Our experiments show significant improvements in cross-domain few-shot action recognition performance.

In summary, our work makes the following major contributions,

- We propose a new, challenging, and realistic problem called cross-domain few-shot learning in videos (CDFSL-V).
- We propose a novel solution based on self-supervised feature learning and curriculum learning for this challenging problem, which can address the difficulties associated with CDFSL-V by striking a balance between learning generic and class-discriminative features.
- We conduct extensive experimentation on multiple benchmark datasets. Our proposed method outperforms the existing methods in cross-domain few-shot learning, as well as, strong baselines based on transfer learning.

2. Related Work

Few-Shot Classification Few-shot Learning methods can be split into two main categories: Meta-Learning and Transfer Learning [27]. Meta-Learning [32, 31] framework provides a very common technique for few-shot learning algorithms, where the training procedure mimics the evaluation procedure. Just as few-shot evaluation consists of multiple few-shot episodes on the target test set, meta-learning techniques train a model in an episodic fashion on a meta-train set. In meta-learning this is done to encourage fast adaptation on the meta-test set. The other main approach in few-shot learning is Transfer Learning, where a model is pretrained on the source dataset before being fine-tuned on

the target data for few-shot evaluation [35, 19, 40, 6]. Methods that use transfer learning aim to leverage as much information as possible from the source dataset in order to produce easily transferrable features to be adapted to the target dataset. Both methods assume some degree of similarity between the source and target datasets, hinging on the idea that features that can discriminate classes in the source domain can also discriminate classes in the target domain. When moving from images to videos, the introduction of temporal information adds to the difficulty of the task. OTAM [4] uses temporal alignment to improve few-shot classification for videos, using a distance metric to compare frames of the queries and the support set. STRM [34] introduces a spatio-temporal enrichment module to look at visual and temporal context at the patch and frame level. HYSRM [39] uses a hybrid relation model to learn relations within and across videos in a given few-shot episode. Our method focuses on training an encoder with generalizable features by leveraging unlabeled target data during training through both self-supervised learning and enforcing a consistency loss moderated by curriculum learning.

Self-Supervised Learning Self-supervised learning has been shown to improve performance when combined with supervised learning by creating more transferable features [36]. These more generalizable features are extremely important in the cross-domain few-shot classification task, due to the domain gap and scarcity of labels. For self-supervised video classification, existing methods use contrastive learning to improve learning visual representation, at the cost of increased data augmentation and batch sizes [37, 23, 43, 1]. Masked auto-encoders [14] mask patches of an image and attempt to reconstruct the missing parts. VideoMAE [36] extends this to video by adding space-time attention via a ViT backbone, providing a data efficient solution so the self-supervised video pretraining. We use VideoMAE as the backbone of our method.

Curriculum Learning Curriculum learning involves prioritizing easier samples (or tasks) during training before increasing the weights of the more difficult samples [2]. Typically, training examples are sorted by a difficulty metric, and used to create mini-batches of increasing difficulty for training the model [13]. This method has shown success when applied in computer vision, specifically when used with transfer learning [42].

For our problem setup we work with two datasets,the labeled source dataset and the unlabeled target data (for which we generate pseudo-labels), simultaneously during training. In our method we leverage curriculum learning such that we focus on the large labeled source dataset at the beginning of training, and eventually shifting towards equal weighting of the source and target losses.

Cross-Domain Few-Shot Learning Similar to open-world semi-supervised learning [3, 11, 30, 29] that allows semi-supervised learning methods to perform on loosely related domains, the cross domain few shot learning framework permits base and test data that belong to different domains. BS-CDFSL [12] introduces a benchmark for the Cross-Domain Few-Shot problem for images. It consists of miniImageNet as the source dataset, and four target datasets of increasing difficulty: CropDisease [24], EuroSAT [15], ISIC [7], and ChestX [38]. STARTUP [28] attempts to solve this problem by learning a teacher model on the source dataset that is applied to generate pseudo-labels for the target dataset. Eventually, a new model on both the labeled source set and pseudolabeled target set is trained. Dynamic Distillation [16] improves upon this by updating the teacher model as a moving average of the student's weights. Both of these methods exhibit redundancy in the supervised training across their stages that we strive to eliminate in our approach.

While source-target dataset pairs such as UCF-HMDB51 from the SDAI Action II dataset [9] and the UCF-OlympicSport datasets [17] have been proposed [8], these dataset pairs share classes across domains, which is not representative of the CDFSL problem. We take inspiration from the BS-CDFSL benchmark and use Kinetics-100 [44] as our source, with UCF101 [33], HMDB51 [20], Something-SomethingV2 [10], Diving48 [21], and RareAct [22] as our datasets. We ensure that we remove any class overlap between the source and target datasets.

3. Methodology

This section elaborates on our approach to tackle the CDFSL problem in the video domain. At the core of our method, we learn features from the source and target data in a supervised and self-supervised fashion, respectively. Furthermore, we propose a progressive curriculum to encourage the emergence of rich features in the target domain based on class discriminative supervised features in the source domain. In the following, first, we discuss our problem formulation (Sec. 3.1). After that, we present our approach involving self-supervised feature learning and curriculum learning (Sec. 3.2).

3.1. Problem Formulation

The Cross-Domain Few-Shot Video Classification task requires the classification of an unlabeled query video belonging to the target dataset $\mathcal{D}_T$. A large labeled source dataset $\mathcal{D}_S$ is available during training. $\mathcal{D}_S$ and $\mathcal{D}_T$ have no shared classes, and usually have a significant domain gap. The unlabeled training split of $\mathcal{D}_T$ is leveraged during training, denoted as $\mathcal{D}_{T_U}$. For evaluation, multiple Few-Shot episodes are sampled from the testing split of $\mathcal{D}_T$. These episodes consist of a small labeled support set $\mathcal{S} \subset D_T$, consisting of a few labeled samples of each target class in the episode, and a disjoint query set $\mathcal{Q} \subset D_T$ to be classi-

fied. In the N-way K-shot classification setting, $\mathcal{Q}$ and $\mathcal{S}$ share the same N classes sampled from $\mathcal{D}_T$ with $\mathcal{S}$ having K labeled examples for each class.

3.2. Approach

3.2.1 Self-Supervised Feature Learning

A fundamental challenge in solving few-shot problems is learning generalizable representations. A successful representation learning method is based on self-supervised learning, therefore, it has been readily applied to few-shot learning problem. Even though, it has yet to be applied in CDFS learning. Following the success of VideoMAE, to extract strong representations from video data we apply VideoMAE model in our Pretraining phase. To this end, a rich set of features are extracted from a combination of the source and unlabeled-set of target dataset $\mathcal{D}_S \bigcup \mathcal{D}_{T_U}$. After this step, the encoder model from VideoMAE, f, is utilized as our primary feature extractor.

3.2.2 Curriculum Learning

Next, in our framework, we further improve the quality of the extracted features with the help of the ground-truth labels of the source data. To this end, we train a classifier g on top of f, where this classifier outputs the number of classes equal to the classes in the source domain. Training a classifier in such a supervised manner makes the self-supervised representation more compact and class discriminative, particularly in the source domain. Ideally, we want to achieve the same in the target domain. However, doing such is difficult without accessing the ground-truth labels in the target domain. To overcome this challenge and to better utilize the target data, we minimize a consistency loss for the unlabeled target samples. This consistency loss is minimized at the output space of the source domain where pseudo-labels are generated using a teacher network.

Supervised Representation Learning To extract the discriminative features from the source dataset, we first train a student model f_s based on a supervised loss on the labeled source data. We use the commonly used cross-entropy loss as the supervised loss, $\mathcal{L}_{sup}$, defined in the following,

$$\begin{aligned}\mathcal{L}_{sup} =& \mathcal{L}_{CE}(\text{Softmax}(f_s(\mathbf{x}_i)), \mathbf{y}_i) \\ =& -\sum_{i=1}^{M} \mathbf{y}_i \log(\text{Softmax}(f_s(\mathbf{x}_i))),\end{aligned} \tag{1}$$

where, $\mathbf{x}_i \in \mathcal{D}_S$, $M = |\mathcal{D}_S|$, and $\mathbf{y}_i$ is the ground-truth label. The learned discriminative features provides us with more generalizable features to the target domain.

Unsupervised Representation Learning For the unlabeled data from target domain, we apply pseudo-labels to increase generalizability of the learned features in an unsupervised fashion. To this end, after obtaining the pseudo-labels we compute a consistency loss. The consistency loss ensures that the representations from the student model match the representations from a teacher network. We create a teacher model f_t by taking an exponential moving average of the student model in the following manner,

$$f_t^{(i+1)} = \alpha f_t^{(i)} + (1-\alpha) f_s^{(i+1)}, \tag{2}$$

where, α is the exponential decay parameter, i refers to ith iteration.

This consistency loss ensures that the f_s predictions for unlabeled target data match with the pseudo-labels generated from f_t. Additionally, following the success of DINO [5] we want to extract features that can learn a local-to-global relationship between data. To this end, each batch of unlabeled target data $\mathbf{X} \in \mathcal{D}_{T_U}$, is transformed into two separate sets to make strong and weak augmented copies of the batch: $\mathbf{X}_{str}$ and $\mathbf{X}_{weak}$. To be specific, we use temporally consistent RandomResizeCrop and $\text{RandomHorizontalFlip}$ as a set of weak augmentations, while the set of strong augmentations consists of temporally consistent RandomColorJitter, RandomGreyscale, and $\text{RandomGaussianBlur}$ in addition to the set of weak augmentations.

To compute the consistency loss, first the weakly augmented unlabeled target data is passed through the teacher model to get the teacher outputs $f_t(\mathbf{X}_{weak})$. These outputs are then sharpened by a temperature τ to form pseudo-labels for the target data after performing the Softmax operation. The consistency loss is a cross-entropy loss between the student outputs of the strongly augmented videos $f_s(\mathbf{X}_{str})$ and the sharpened teacher outputs which is defined in the following,

$$\mathcal{L}_{con} = -\sum \hat{\mathbf{Y}} \log(\text{Softmax}(f_s(\mathbf{X}_{str}))), \tag{3}$$

where, $\hat{\mathbf{Y}} = \text{Softmax}(f_t(\mathbf{X}_{weak})/\tau)$.

The overall training objective for updating the parameters of the student network is a weighted average of the Supervised and the Consistency losses, defined in the following,

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{con}, \tag{4}$$

where, the consistency loss scaling parameter λ controls the relative contribution of consistency loss to the total loss.

While previous CDFS methods have applied both supervised loss and consistency loss, they applied them in separate stages [16, 28]. One of the unique characteristics of our approach is to combine these losses through curriculum

learning, which not only simplifies the training pipeline but also improves performance.

In our curriculum, we adjust the difficulty of the consistency through tuning its scaling parameter λ following a pre-defined curriculum. In particular, at the beginning of training, we set the consistency loss scaling parameter, λ, to a very low value. This makes the beginning of the training similar to performing supervised training solely on the source dataset. As the training progress, we emphasize the importance of consistency by increasing λ over the course of the training, which encourages the emergence of local-to-global features that can potentially generalize better in the target domain. Additionally, to facilitate the transition from the source domain to the target domain, we decay the learning rate of the classifier in the student model over the course of training. Initially, this classifier is trained at the same rate as the rest of the student model. This learning rate is decreased over the course of training to emulate freezing the classifier after supervised training on the source data. Once the training is complete, the student model is kept and the classifier is discarded. Using the labeled support set of the target data, a new logistic regression layer c' is learned on top of the student model. The model can now be used for inference on the target query images. The entire procedure is summarized in Algorithm 1.

Algorithm 1 Curriculum Learning for CDFSL-V

f_s, f_t: student, teacher model with parameters θ_s, θ_t.
τ: teacher temperature
α: momentum rate to update teacher
for $(\mathbf{x}_s, \mathbf{y}_s)$, $\mathbf{x}_t$ in loader **do**
  sample $\mathbf{x}_s, \mathbf{y}_s$ from base data
  sample $\mathbf{x}_t$ from target data
  $\mathcal{L}_{sup} = \mathcal{L}_{CE}(f_s(\mathbf{x}_s), \mathbf{y}_s)$
  $\mathbf{x}_{weak}, \mathbf{x}_{str} = \mathrm{WeakAug}(\mathbf{x}_t), \mathrm{StrongAug}(\mathbf{x}_t)$
  $out_t, p_s = f_t(\mathbf{x}_{weak}), \mathrm{Softmax}(f_s(\mathbf{x}_{str}))$ ▷ teacher logits and student pseudo-labels
  $p_t = \mathrm{Softmax}(out_t/\tau, dim = -1).detach()$ ▷ sharpen + stop-grad
  $\mathcal{L}_{con} = \mathcal{L}_{CE}(p_s, p_t)$ ▷ consistency loss
  $\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda\mathcal{L}_{con}$
  $\theta_s \leftarrow \theta_s + \beta\nabla_{\theta_s}\mathcal{L}_{total}$ ▷ update student
  $\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s$ ▷ update teacher
end for

4. Experiments

In this section we evaluate our proposed method with strong transfer learning baselines and the recent techniques applied in cross-domain few-shot learning. For a thorough comparison, we utilize a variety of target domains to capture performance of different methods when encountering a variety of cross-domain scenarios. Our main result is that our approach outperforms existing sate-of-the-art cross-domain few-shot learning techniques. Finally, we conduct an ablation and analyse the significance of different components of our approach.

4.1. Datasets

We use the Kinetics-100 [44] train split as our Source dataset. It contains 100 of the original dataset further split into 64, 12, and 24 class subsets for train, validation, and test, respectively for few-shot action recognition. We also conduct experiments on the larger Kinetics-400 [18] (Table 1). Due to class overlap between Kinetics and two of our target datasets, UCF101 and HMDB51, we remove the overlapping classes from the source dataset. Without this removal, the supervised training on shared classes between the source and target datasets would be an unfair representation of the Cross-Domain Few-Shot problem setting. The target datasets in the order of increasing difficulty are: UCF101, RareAct, HMDB51, Something-SomethingV2, and Diving48. UCF101 and HMDB51 are most similar to Kinetics datasets in terms of domain gap. They even have overlapping classes that needed to be removed in order to make them appropriate target datasets However, that is not the case for the other target datasets. For instance, the Something-SomethingV2 dataset has 87 classes, consisting of actions doing 'something' to 'something'. This dataset primarily contains zoomed-in videos focusing on the object instead of the person which is generally not the case for actions present in the actor-centric Kinetics dataset. Diving48 on the other hand is a dataset for fine-grained action recognition with 48 different dives, each comprised of different sequences of complex sub-actions. The RareAct is very different from all other source and target datasets since it contains unusual actions like 'blend phone' and is generally used for evaluating few/zero-shot action compositionality. For evaluation, we compute the 5-way 5-shot accuracy on the test-split for each target dataset.

4.2. Experiment Details

We use the encoder network from VideoMAE with a ViT-S backbone for our feature extraction. For videos we sample 16 frames at a 112×112 resolution. We train on the combined training data of both the source and target datasets without labels for 400 epochs at a batch size of 32 using SGD optimizer at a learning rate of 0.1. After initializing the student and teacher models using the VideoMAE encoder, the student is trained for 200 epochs on the combined supervised and consistency losses. The student is updated directly using SGD with a learning rate of 0.01, and the teacher is updated as a moving average of the student weights with a momentum of 0.9. The teacher output is sharpened at a temperature of 0.1 to be used as pseudo-labels for the student output on the unlabeled target data.

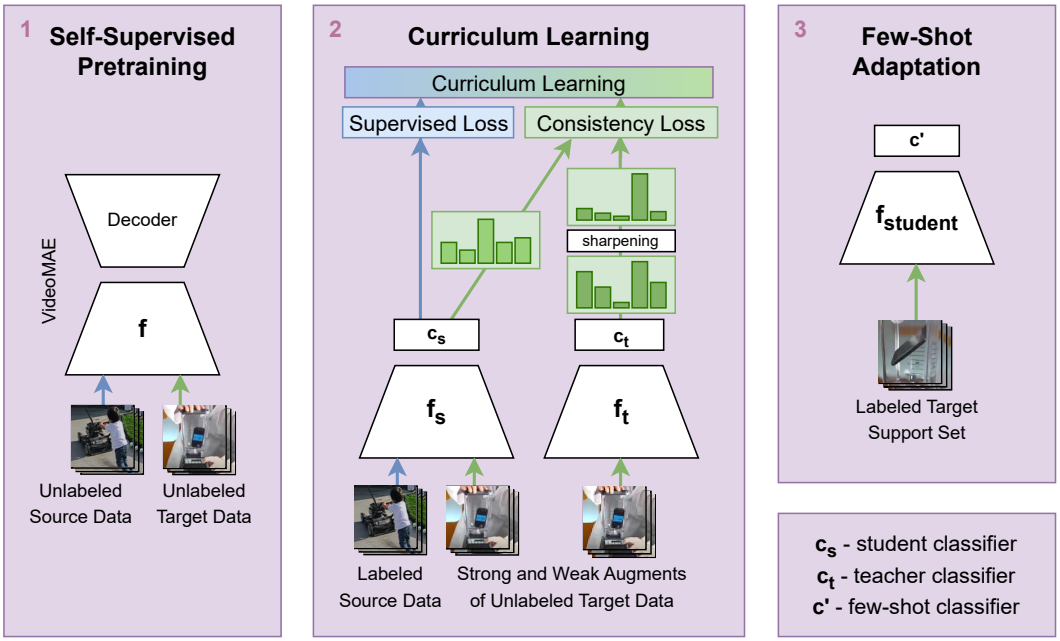


Figure 2: Our goal is to solve the cross-domain few-shot learning task for the target dataset, leveraging the labeled base dataset alongside unlabeled target data. Our method three three stages: **1**: Self-supervised pretraining of an autoencoder on both the source and target data without labels is performed. **2**: The encoder is used to initialize a student and teacher model for curriculum learning. We compute a supervised loss on the labeled source data. For the consistency loss, we generate pseudo-labels using the sharpened teacher output for weakly augmented target images. The pseudo-labels are then used with the student output on strong augmentations of the same images to calculate the consistency loss. The supervised and consistency losses are both used to directly update the student, while the teacher is updated as a moving average of the student's weights. **3**: for few-shot evaluation, the student classifier is replaced with a few-shot classifier that is fine-tuned on the labeled target support set. This classifier can then be used to classify the target query images.

The batch size used for the curriculum learning stage is 16. Over the course of training, we set the consistency loss scaling parameter, λ as:

$$\lambda_{cons} = \frac{\arctan(10(x - .5))}{\pi} + .5 \qquad (5)$$

where x is the ratio of current epoch to total epochs in training. This reduces the weight of the consistency loss significantly in the start of training, while making it on par with the supervised loss towards the end. Similarly, the learning rate for the student classifier head (the classifier layer following f_s) decayed according to $\lambda_{cls} = \frac{\arctan(-10(x-.5))}{\pi} + .5$. so that the classifier head learns primarily from the supervised loss early on and effectively freezes towards the end of training.

For few-shot adaptation, the student encoder is retained and the student classifier head is discarded. We then learn a new logistic regression classifier on top of the encoder using a sampled 5-way 5-shot support set from the target testing data. We report the accuracy on the remaining testing data for the selected classes.

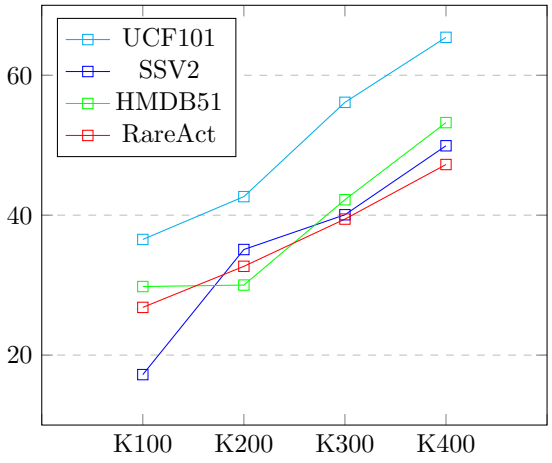


Figure 3: Results with varying size of source data.

4.3. Kinetics-400 Experiment

Random initialization is used as the baseline for this experiment and entails learning a logistic regression classi-

Method	UCF101	SSV2	HMDB51	Diving48	RareAct	Average
Random Initialization	23.83	16.02	12.08	15.37	16.57	16.78
STARTUP++	60.82	39.60	44.71	14.92	45.22	41.05
Dynamic Distillation++	63.26	44.50	48.04	16.23	47.01	43.81
STRM	42.33	35.01	24.98	16.69	39.01	31.60
HYRSM	45.65	40.09	29.81	17.57	44.27	35.49
Ours	**65.42**	**49.92**	**53.23**	**17.84**	**49.80**	**47.24**

Table 1: 5-way 5-shot Accuracy using Kinetics-400 as the source dataset. We use STARTUP++ and Dynamic Distillation++ to denote that these methods include self-supervised pretraining, despite being used in their original papers.

Method, Source Dataset: Kinetics-100	UCF101	HMDB51	SSV2	Diving48	RareAct	Average
Equal Loss Weighting	32.02	27.39	15.34	16.07	33.67	24.90
No Temperature Sharpening	34.01	28.18	15.21	16.77	33.80	25.59
Self-Supervised Training	37.54	25.09	16.21	17.14	29.58	25.11
Supervised Training	32.06	23.86	14.40	16.16	31.15	23.53
Ours	**36.53**	**29.80**	**17.21**	**16.37**	**33.91**	**26.82**

Table 2: The effect of removing different components of our proposed method.

fier on top of an *untrained* VideoMAE encoder. We compare our method to two Cross-Domain Few-Shot methods for images, as no other methods exist to solve the CDFSL problem for videos. For this experiment, we include self-supervised pre-training for Dynamic Distillation and STARTUP, denoting them as Dynamic Distillation++ and STARTUP++, respectively. In addition, we compare our method to two Few-Shot methods for Videos: STRM [34] and HYRSM[39]. Our method outperforms the previous state-of-the-art method, Dynamic Distillation [16], across all 5 target datasets while using the Kinetics-400 dataset as the source. Additionally, the absolute improvement in classification performance is consistent with the aforementioned relative difficulty of each of the target datasets, with Diving48 improving the least.

Our main result is that we do better than existing CDFSL methods for images, as well as the Few-Shot methods for videos. As shown in Table [18], We outperform Dynamic Distillation by 2.2% on UCF101, 5.2% on HMDB51, 5.4% on SSV2, 2.7% on RareAct, and 2.8% on Diving48, averaging to a 3.4% increase. Compared to STARTUP, STRM, and HYSRM our method outperforms by 6.19%, 15.64%, and 11.8%, respectively. Interestingly, even our modified image baselines (STARTUP++ and Dynamic Distillation++) outperform these video few-shot methods, highlighting the inadequacy of traditional video few-shot approaches for this challenging Cross-Domain Few-Shot problem.

4.4. Kinetics-100/200/300 Experiments

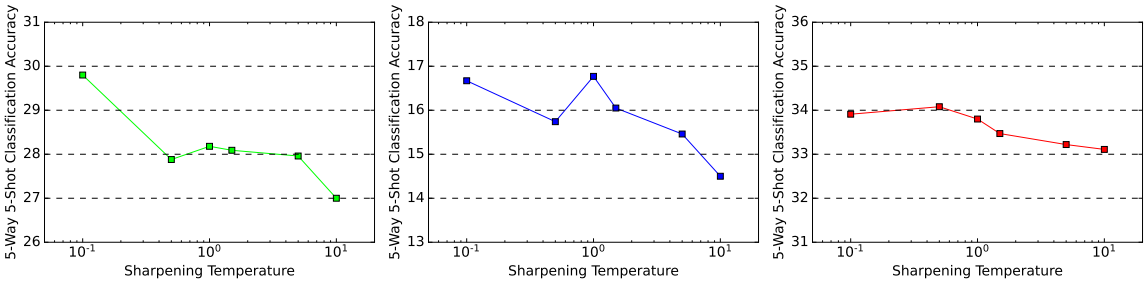
We repeat the experiments using Kinetics-100, Kinetics-200, and Kinetics-300 as the source datasets. We compare our method's performance across these varying source datasets. In this experiment, we evaluate how the increase in the number of classes in the source dataset impacts performance. As shown in Fig 3, increasing the size of the source dataset consistently improves performance on all datasets.

4.5. Ablation and Analysis

In this section we analyze the importance of different components of our approach. Particularly we study the effect of increasing the size of the source dataset, the effect of sharpening temperature, and the impact of curriculum learning on the performance of the method.

Increasing the size of the source dataset In few-shot learning and transfer learning literature, it is common to utilize a source domain with a significantly larger number of classes than the target domain [12]. A dataset with a larger number of classes can capture a more diverse set of features which facilitates its application on less diverse datasets. For example, in the BS-CDFSL benchmark for images, miniImageNet, the source dataset, has 100 classes. The target image datasets: CropDisease, EuroSAT, ISIC, and ChestX have 38, 10, 5, and 15 classes, respectively. In that setup, the source dataset has over double the amount of classes of the largest target dataset. In comparison, the source dataset in our video benchmark has 61 classes, which is less than two of the target datasets: UCF101 with 101 and Something-SomethingV2 with 87. In this experiment, we explore the impact of the size of the source dataset in CDFSL. To be more specific, we apply the larger Kinetics-400 dataset instead of Kinetics-100 as the source.

Both STARTUP and Dynamic Distillation make use of supervised pretraining on the source dataset. For the experiments on Kinetics-400, we supplement the pretraining stages of both methods with self-supervised pretraining as well to highlight the effect of our curriculum-based schedule. In Table 1, We observe a drastic improvement in the performance when we utilize a more diverse source dataset.



(a) HMDB51 (b) Diving48 (c) RareAct

Figure 4: Temperature parameter experiments. We use Kinetics-100 as the source dataset and vary the sharpening temperature for the teacher pseudo-labels. As the temperature increases (and sharpness decreases) the performance tends to decrease.

Interestingly, we further notice an increase in the relative performance of our method to Dynamic Distillation.

Temperature Sharpening Analysis As in STARTUP [28], we want to leverage the unlabeled target data during training by using a consistency loss. We use the teacher model to create the ground truth for this loss, dividing the teacher output by the temperature parameter T to sharpen it and use as pseudo-labels. Similarly to Dynamic Distillation [16], sharpening of the labels is used to develop low-entropy predictions from the student.

We study the impact of temperature sharpening by setting the temperature parameter to 1 (with the default value taken from Dynamic Distillation being 0.1), making the teacher output completely unsharpened. As shown in second row of Table 2, removing the temperature sharpening reduces performance in almost all datasets (the exception being Diving48 with a 0.4% increase) with an average decrease in performance of 1% compared to our original method. We can see that temperature sharpening has a slight but positive impact when used with our CDFSL problem setup for videos.

Pretraining Baselines It has been shown that pretraining contributes a significant part to few-shot learning [25]. To examine how much of the performance is attributed to this, we compare some established transfer learning baselines with multiple pretraining configurations followed by few-shot adaptation. Self-supervised training refers to only self-supervised training on the combined source and target datasets without labels, and supervised training is simply training on the labeled source dataset. In rows 3 and 4 of Table 2, we can see the contribution of each of the pretraining techniques. For most of the datasets, only using self-supervised pretraining outperforms using supervised pretraining, with the exception being RareAct. On average, our method performs 1.7% over the self-supervised baseline and 3.3% over the supervised baseline.

Impact of Curriculum Learning The motivation behind curriculum learning is to ease the training of the model by focusing more on easier data first. For our problem setup where we leverage unlabeled target data alongside the labeled source, we begin with focusing more on the supervised source loss as it is an easier task than matching target videos to pseudo-labels in the source domain. Once the model has sufficiently learned relationships from the source dataset, the importance of the target consistency loss can increase to help improve the adaptation.

We use λ_{cons} to scale the consistency loss during training, as shown in Eq. 5. To analyze the effect of enforcing the curriculum scaling, we compare keeping λ_{cons} at 1 for the entirety of training and making both supervised and consistency losses weighted equally the whole time. Additionally, we train our model at temperatures of 0.5, 1.5, 5, and 10 as shown in Figure 4. Weighting both losses equally results in an average drop in performance of 1.6%. We see that using curriculum learning improves the performance.

5. Conclusion

In this paper, we addressed the problem of cross-domain few-shot action recognition in videos, which is a challenging and realistic problem with several practical applications in fields such as robotics. We proposed a novel approach based on self-supervised feature learning and curriculum learning to address the challenges associated with this problem. Our approach strikes a balance between learning generic and class-discriminative features, which significantly improves the few-shot action recognition performance. We conducted extensive experiments on various benchmark datasets, where our proposed method outperforms current cross-domain few-shot learning methods in the image domain and few-shot learning methods in the video domain. Our work contributes to the computer vision community by introducing a new problem and providing a novel solution to address it. We hope that this work will inspire further research in this direction and help advance the state-of-the-art in few-shot action recognition.

6. Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (IARPA) via 2022-21102100001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. *CoRR*, abs/2004.06130, 2020. 3

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. 3

[3] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021. 3

[4] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. *CoRR*, abs/1906.11415, 2019. 1, 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021. 4

[6] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *CoRR*, abs/1904.04232, 2019. 1, 3

[7] Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1902.03368, 2019. 3

[8] Zan Gao, Leming Guo, Weili Guan, An-An Liu, Tongwei Ren, and Shengyong Chen. A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-r2. *IEEE Transactions on Image Processing*, 30:767–782, 2021. 3

[9] Zan Gao, Yibo Zhao, Hua Zhang, Da Chen, An-An Liu, and Shengyong Chen. A novel multiple-view adversarial learning network for unsupervised domain adaptation action recognition. *IEEE Transactions on Cybernetics*, 52(12):13197–13211, 2022. 3

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. *CoRR*, abs/1706.04261, 2017. 3

[11] Lan-Zhe Guo, Yi-Ge Zhang, Zhi-Fan Wu, Jie-Jing Shao, and Yu-Feng Li. Robust semi-supervised learning when not all classes have labels. *Advances in Neural Information Processing Systems*, 35:3305–3317, 2022. 3

[12] Yunhui Guo, Noel C. F. Codella, Leonid Karlinsky, John R. Smith, Tajana Rosing, and Rogério Schmidt Feris. A new benchmark for evaluation of cross-domain few-shot learning. *CoRR*, abs/1912.07200, 2019. 1, 3, 7

[13] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. *CoRR*, abs/1904.03626, 2019. 3

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 3

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *CoRR*, abs/1709.00029, 2017. 3

[16] Ashraful Islam, Chun-Fu Chen, Rameswar Panda, Leonid Karlinsky, Rogério Feris, and Richard J. Radke. Dynamic distillation network for cross-domain few-shot recognition with unlabeled data. *CoRR*, abs/2106.07807, 2021. 1, 2, 3, 4, 7, 8

[17] Arshad Jamal, Vinay P. Namboodiri, Dipti Deodhare, and K. S. Venkatesh. Deep domain adaptation in action space. In *British Machine Vision Conference*, 2018. 3

[18] Will Kay, João 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. *CoRR*, abs/1705.06950, 2017. 5, 7

[19] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Large scale learning of general visual representations for transfer. *CoRR*, abs/1912.11370, 2019. 3

[20] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011. 3

[21] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 3

[22] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Rareact: A video dataset of unusual interactions. *CoRR*, abs/2008.01018, 2020. 2, 3

[23] Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Unsupervised learning using sequential verification for action recognition. *CoRR*, abs/1603.08561, 2016. 3

[24] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 2016. 3

[25] Jaehoon Oh, Sungnyun Kim, Namgyu Ho, Jin-Hwa Kim, Hwanjun Song, and Se-Young Yun. Understanding cross-

domain few-shot learning: An experimental study. *CoRR*, abs/2202.01339, 2022. 1, 2, 8

[26] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning, 2022. 1

[27] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning, 2022. 2

[28] Cheng Perng Phoo and Bharath Hariharan. Self-training for few-shot transfer across extreme task differences. *CoRR*, abs/2010.07734, 2020. 1, 2, 3, 4, 8

[29] Mamshad Nayeem Rizve, Navid Kardan, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Openldn: Learning to discover novel classes for open-world semi-supervised learning. In *European Conference on Computer Vision*, pages 382–401. Springer, 2022. 3

[30] Mamshad Nayeem Rizve, Navid Kardan, and Mubarak Shah. Towards realistic semi-supervised learning. In *European Conference on Computer Vision*, pages 437–455. Springer, 2022. 3

[31] T. Schaul and J. Schmidhuber. Metalearning. *Scholarpedia*, 5(6):4650, 2010. revision #91489. 2

[32] Jurgen Schmidhuber. Evolutionary principles in self-referential learning. on learning now to learn: The meta-meta-meta...-hook. Diploma thesis, Technische Universitat Munchen, Germany, 14 May 1987. 2

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 3

[34] Anirudh Thatipelli, Sanath Narayan, Salman Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Bernard Ghanem. Spatio-temporal relation modeling for few-shot action recognition. In *CVPR*, 2022. 3, 7

[35] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B. Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *CoRR*, abs/2003.11539, 2020. 1, 3

[36] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. 2, 3

[37] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 3

[38] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315, 2017. 3

[39] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Mingqian Tang, Zhengrong Zuo, Changxin Gao, Rong Jin, and Nong Sang. Hybrid relation guided set matching for few-shot action recognition, 2022. 3, 7

[40] Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *CoRR*, abs/1911.04623, 2019. 3

[41] Yaqing Wang and Quanming Yao. Few-shot learning: A survey. *CoRR*, abs/1904.05046, 2019. 1

[42] Daphna Weinshall and Gad Cohen. Curriculum learning by transfer learning: Theory and experiments with deep networks. *CoRR*, abs/1802.03796, 2018. 3

[43] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10326–10335, 2019. 3

[44] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *European Conference on Computer Vision*, 2018. 3, 5

[45] Zhenxi Zhu, Limin Wang, Sheng Guo, and Gangshan Wu. A closer look at few-shot video classification: A new baseline and benchmark. *CoRR*, abs/2110.12358, 2021. 1