CDFSL-V: Cross-Domain Few-Shot Learning for Videos (Supplementary)

1. Overview

In this Supplementary, we report additional results for our proposed method. In particular, this Supplementary includes the following sections: Section 2 provides extra results for our approach, when we utilize Kinetics-100 and UCF101 as the source datasets. In Section 3, we report results with different numbers of support samples which complements our results reported in the main text. Next, we provide results on image datasets for cross-domain few-shot learning in Section 4. Finally, in Section 5, we show a quantitative comparison of our method against the previous state-of-the-art.

2. Varying Source Datasets

We repeat the main experiments using Kinetics-100 and UCF101 as the source datasets. In comparison to Kinetics, UCF101 has higher scene bias [1] and temporally simpler actions. In this experiment, We compare our method to the other CDFSL methods with a reduced amount of classes in the source dataset. The number of source classes changes to 61 and 101 in this experiment, for Kinetics-100 and UCF101, respectively. We also evaluate how effectively features learned from UCF101 transfer to the target domains with different actions in nature. We follow the same experimental setup as main experiments.

The main result is that we similarly outperform STARTUP and Dynamic Distillation. We achieve a 2.1% improvement on average over Dynamic Distillation with Kinetics-100 as the source dataset, shown in Table 1. When using UCF101 as the source dataset, as in Table 2, We have an increase in performance over Dynamic Distillation of 4.8% on HMDB51 and 1.9% on RareACT. In summary, our experiments in this experimental setup, re-validate the effectiveness of our method compared to the other CDFSL methods in the video setting.

Method, Source Dataset: Kinetics-100	UCF101	HMDB51	SSV2	Diving48	RareAct	Average
Random Initialization	23.83	16.02	12.08	15.37	16.57	16.78
STARTUP	32.20	24.97	15.16	14.55	31.77	23.73
Dynamic Distillation	34.10	25.99	16.00	16.24	31.20	24.71
Ours	**36.53**	**29.80**	**17.21**	**16.37**	**33.91**	**26.82**

Table 1: 5-way 5-shot Accuracy using Kinetics-100 as the source dataset

Method, Source Dataset: UCF101	HMDB51	DIVING48	RareAct	Average
Random Initialization	21.69	14.48	26.98	21.05
STARTUP	23.56	14.84	31.31	23.24
Dynamic Distillation	24.06	**16.15**	32.00	24.07
Ours	**28.86**	16.07	**33.91**	**26.82**

Table 2: 5-way 5-shot Accuracy using UCF101 as the source dataset

3. Varying k for Few-Shot Classification

In this section, we include the results for different sizes of target support sets. This is complementary to the Kinetics-100 Experiment of the previous section, reported in Table 1. We follow the same experimental setup for training across all methods, only changing the number of k shots for the few-shot evaluation. Table 3 shows the performance of different approaches in 1-shot evaluation. We observe that even in this challenging setup our proposed method outperforms all the other methods with a noticeable margin ($> 1\%$). We observe an even bigger improvement when we increase the number of support samples to 20. We report the results for this evaluation in Table 4 and notice that our proposed method outperforms the previous state-of-the-art method, Dynamic Distillation [7], by a significant margin. This further validates the effectiveness of our method and shows that our method is more effective with a higher number of support samples.

4. Cross-domain Few-shot Learning on Image data

In addition to improving performance in the CDFSL-V problem setting, we demonstrate the effectiveness of our method on the Cross-Domain Few-Shot Learning problem for images. For these experiments we follow the BS-CDFSL [3] benchmark for CDFSL. The benchmark uses the 100-class miniImageNet dataset as the source dataset. Following [7], for the target datasets, we use CropDisease[8] – a plant disease dataset with both healthy and sick specimens, EuroSAT[6] – an aerial-view dataset of various land use and land cover types, and the ISIC Challenge dataset from 2018[2] – a dermoscopic dataset of various skin lesions relating to diseases. In comparison to the source

Method, Source Dataset: Kinetics-100	UCF101	HMDB51	SSV2	Diving48	RareAct	Average
Random Initialization	23.83	16.02	12.08	15.37	16.57	16.78
STARTUP	24.48	16.66	14.17	13.13	17.21	17.13
Dynamic Distillation	26.04	17.44	14.96	13.73	19.02	18.24
Ours	**27.78**	**18.59**	**16.01**	**14.11**	**20.06**	**19.31**

Table 3: 5-way 1-shot Accuracy using Kinetics-100 as the source dataset

Method, Source Dataset: Kinetics-100	UCF101	HMDB51	SSV2	Diving48	RareAct	Average
Random Initialization	32.33	27.97	15.12	15.83	33.53	24.96
STARTUP	34.02	30.48	17.15	17.30	38.45	27.48
Dynamic Distillation	36.72	33.09	17.56	17.33	39.97	28.93
Ours	**39.92**	**36.89**	**18.72**	**17.81**	**42.51**	**31.17**

Table 4: 5-way 20-shot Accuracy using Kinetics-100 as the source dataset

dataset the target datasets have significantly fewer classes, with 38, 10, and 5 for CropDisease, EuroSAT, and ISIC, respecitvely. Despite having the fewest classes, ISIC is the most difficult dataset in the CDFSL task, having both high few-shot difficulty and low domain similarity with the source[9].

Dynamic Distillation[7] normally uses a ResNet[5] as the backbone for the student encoder. For the following image experiments we compare our method against Dynamic Distillation using a VitMAE[4] backbone, to provide a fair comparison by keeping the architectures consistent. We denote this altered backbone Dynamic Distillation as Dynamic Distillation* in Table 5. From the reported results in Table 5, we notice that our proposed method even outperforms the previous state-of-the-art methods by a significant margin on all three datasets. On average, our proposed method improves over Dynamic Distillation* by more than 2%. In summary, these results demonstrate that our proposed method is effective in both image and video data.

5. Qualitative Analysis

In this section we show some video samples from SSV2 and RareAct datasets that our method classifies correctly, while Dynamic Distillation, the runner-up performing method, does not. Specifically, Fig.1 shows an example where Dynamic Distillation [7] confuses *pouring a liquid* with *spreading air* from SSV2. Fig. 2 shows another video where Dynamic Distillation [7] misclassifies *folding* with *lifting up*. Fig. 3 shows a video from RareAct dataset where our method successfully recognizes *hammering a can*, while Dynamic Distillation [7] misclassifies it as *peeling a corn*. Finally, Fig. 4 shows a video from RareAct where Dynamic Distillation [7] misclassifies *cutting a phone* with *dropping a fridge*.

Method, Source Dataset: miniImageNet	CropDisease	EuroSAT	ISIC	Average
Random Initialization	58.37	52.90	32.69	47.99
Supervised Pretraining	84.48	73.92	45.06	67.82
Dynamic Distillation*	86.68	77.11	47.60	70.46
Ours	**88.81**	**81.42**	**47.80**	**72.68**

Table 5: 5-way 5-shot Accuracy using miniImageNet as the source dataset

Figure 1: *Qualitative results. Dataset: SSV2.* Ground-Truth Label: Pouring [something] into [something] until it overflows. Dynamic Distillation [7] Output: Pretending to spread air onto [something]. Ours: Pouring [something] into [something] until it overflows.

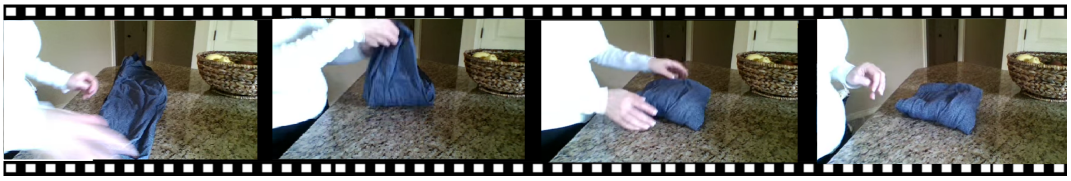

Figure 2: *Qualitative results. Dataset: SSV2.* Ground-Truth Label: Folding [something]. Dynamic Distillation [7] Output: Lifting up one end of [something] without letting it drop down. Ours: Folding [something].

Figure 3: *Qualitative results. Dataset: RareAct.* Ground-Truth Label: Hammer Can. Dynamic Distillation [7] Output: Peel Corn. Ours: Hammer Can.

Figure 4: *Qualitative results. Dataset: RareAct.* Ground-Truth Label: Cut Phone. Dynamic Distillation [7] Output: Drop Fridge. Ours: Cut Phone.

References

[1] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[2] 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. 1

[3] 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

[4] 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. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 2

[6] 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. 1

[7] 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

[8] Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7, 2016. 1

[9] 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. 2