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

Supplementary Material

8. Additional Training Details

In Tables 10, 11 and 12 we provide a more detailed account of the training configurations used in each of the three stages of UNITE. The ViT-B network we use has 87M parameters, all of which are trained in each of the 3 stages of UNITE. Stage 2, which is analogous to standard fine-tuning, took roughly 6 hours for K→S (the largest task we evaluate on) whereas our additional stages, Stage 1 and Stage 3, took 1.5 hours and 6 hours respectively on 4 NVIDIA A5000 GPUs.

Setting	Value
Learning Rate Schedule	Cosine
Base Learning Rate	1.5e-4
Batch Size	256
Warmup Epochs (Linear)	10
Total Epochs	50
Optimizer	AdamW [35]
Optimizer Betas	$\beta_1 = 0.9, \beta_2 = 0.95$
Weight Decay	0.05
Drop Path [23]	0.1
Horizontal Flip	Yes
Random Resize Scales	[0.66, 0.75, 0.875, 1]
Masking Ratio	0.8

Table 10. Training configuration for UMT pre-training stage of UNITE (Stage 1).

Setting	Value
Learning Rate Schedule	Cosine
Base Learning Rate	2.5e-5
Batch Size	28
Warmup Iterations (Linear)	4,000
Total Iterations	20,000
Optimizer	AdamW [35]
Optimizer Betas	$\beta_1 = 0.9, \beta_2 = 0.999$
Weight Decay	0.05
Drop Path [23]	0.1
Layer-Wise LR Decay	0.65
Random Erase	0.25
RandAug [9]	$M = 7, N = 4$

Table 11. Training configuration for source domain fine-tuning stage of UNITE (Stage 2).

9. Masked Consistency Implementation

In Table 7, we investigated alternative pseudolabeling strategies from the MatchOrConf scheme that we employ in UNITE. Rows 1, 2, 3 and 5 include a masked consistency constraint in the style of PACMAC [43]. Here, we provide more detail on the implementation of this masked consistency constraint.

A video is said to satisfy the masked consistency constraint if the class prediction of f_a on the full (*i.e.* unmasked) video is consistent with the class predictions of each of the k masked versions. Like [43], we form the masked views using an attention-guided, greedy round-robin assignment process. Because f_a uses mean pooling instead of a CLS token for classification, we use the attention map of the CLIP teacher image encoder to create a mask for each video frame, with a masking ratio of $r = 0.8$ (identical to the masking process used in the UMT pre-training stage of UNITE). The result is k disjoint masks for each video frame, which are then applied to the video to create the k masked views for consistency assessment.

Data	Setting	Value
Common	LR Schedule	Cosine
	Learning Rate	1e-5
	Warmup Iterations	4,000
	Total Iterations	20,000
	Optimizer	AdamW [35]
	Optimizer Betas	$\beta_1, \beta_2 = 0.9, 0.95$
	Weight Decay	0.05
	Drop Path [23]	0.1
	Layer-Wise LR Decay	0.75
	MatchOrConf Threshold (γ)	0.1
Source	Batch Size	20
	Random Erase	0.25
	RandAug [9]	$M = 7, N = 4$
Target	Batch Size	20
	Data Transform	CenterCrop
Masked Target	Batch Size	20
	Loss Coeff. (λ)	1

Table 12. Training configuration for collaborative self-training stage of UNITE (Stage 3). In order to enhance pseudolabel accuracy, unmasked target domain videos do not undergo data augmentation.

10. Supervised Kinetics-400 Initialization

In Sec. 3.2, we discussed the motivation behind initializing our student network from self-supervised UMT pre-training on Kinetics-710 rather than the supervised Kinetics-400 initialization that has become common in the video DA literature. In Tables 13 and 14 we provide a comparison of source only and target only baselines on *Daily-DA* and *Sports-DA* using both initializations. As expected, we observe significantly higher baseline accuracies when using a supervised Kinetics initialization.

Initialization	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.
UMT K710	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
	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
+ K400 Sup.	Source Only	57.2	74.7	54.0	70.8	71.7	58.3	57.8	51.8	44.3	91.2	65.4	61.7	63.2
	Target Only	81.5	81.5	81.5	90.4	90.4	90.4	80.0	80.0	80.0	99.6	99.6	99.6	87.9

Table 13. *Daily-DA* source only and target only baselines using self-supervised vs. supervised Kinetics pre-trained weights for initialization. "UMT K710" denotes self-supervised UMT pre-training on Kinetics-710, while "+ K400 Sup." denotes additional supervised fine-tuning on Kinetics-400.

Initialization	Method	Target Domain Accuracy (Top-1 %)						
		U→S	K→S	S→U	K→U	U→K	S→K	Avg.
UMT K710	Source Only	67.6	86.8	97.9	98.9	79.9	89.1	86.7
	Target Only	97.9	97.9	98.8	98.8	99.9	99.9	98.9
+ K400 Sup.	Source Only	83.1	86.9	99.3	99.9	95.9	95.7	93.5
	Target Only	96.4	96.4	99.5	99.5	98.7	98.7	98.2

Table 14. *Sports-DA* source only and target only baselines using self-supervised vs. supervised pre-trained weights for initialization. "UMT K710" denotes self-supervised UMT pre-training on Kinetics-710, while "+ K400 Sup." denotes additional supervised fine-tuning on Kinetics-400.

11. Zero-Shot Classification with CLIP

The process we use to perform zero-shot image classification using CLIP [44] is described in Sec. 5.2. In Table 15, we provide the exact class names used to form the inputs to the CLIP text encoder. Classification is performed using a single template: "*A video of a person* {class}."

Daily-DA	*Sports-DA*	*UCF↔HMDB*$_{full}$
drink	archery	climb
jump	baseball	fencing
pick	basketball	golf
pour	biking	soccer
push	bowling	pullup
run	swimming	boxing
walk	diving	pushup
wave	fencing	riding bike
	field hockey	horse riding
	gymnastics	basketball
	golf	archery
	horse riding	walking
	kayaking	
	rock climbing	
	climbing rope	
	skateboarding	
	skiing	
	sumo wrestling	
	surfing	
	tai chi	
	tennis	
	trampoline jumping	
	volleyball	

Table 15. Class names used in zero-shot CLIP classification for each VUDA benchmark.