Supplementary Material for SVFormer: Semi-supervised Video Transformer for Action Recognition

This supplementary Appendix contains the following.

- Section A: The efficiency comparison between our SVFormer with previous methods.
- Section B: The comparison of different spatial-temporal attention mechanisms at SSL settings.
- Section C: Additional experiments supported for our methods.
- Section D: More visualizations of different mixing methods.

A. Efficiency Comparison

In this section, we compare the efficiency of SVFormer with previous state-of-the-art methods [3, 4] in Table 1. We present the Input modals, Input Frames, Training Epochs, Model Parameters and Floating point operations (Flops) here to show the efficiency of SVFormer. Note that we use the inference Flops for a single view here.

We can observe that both variants of our methods (*i.e.*, SVFormer-S and SVFormer-B) require only single RGB modal with 8 frames as input. Besides, our methods only need 30 training epochs, which alleviates the training cost significantly compared to previous methods [3, 4] requiring 200 or 600 training epochs. In particular, SVFormer-S outperforms the state-of-the-art [4] with the fewest model parameters and Flops. Though SVFormer-B requires more model parameters and Flops, the performances are improved significantly.

B. Attention Mechanisms

In this section, we compare three different attention mechanisms in TimeSformer [1] at semi-supervised settings in Table 2. Space-only refers to perform self-attention only in each individual frame, and average predictions for every frame. Joint Space-Time treats each token equally and performs global attention on each token which increasing the computation cost. Divided Space-Time is the default setting in our SVFormer, where temporal attention and spatial attention are separately applied one after the other.

We can observe that Space-only attention performs satisfactory results at 1% labeling ratio. This is mainly caused since the model may not be able to learn the temporal attention when the labels are extremely scarcity. However, when there are more labeled samples (10% labeling ratio), Divided Space-Time attention achieve the best superiority. We believe that the labeled samples are enough for the model to learn the temporal dynamic information in this case. For the consideration of accuracy and efficiency, we thus set Divided Space-Time attention as default, same as in TimeSformer [1].

C. Additional Experiments

In this section, we demonstrate additional experiments and ablation studies to support for our methods. First, it would be interesting to show the effectiveness of our method in more data regime such as 50% labeling ratio and even supervised(*i.e.*,100% labeling ratio) setting. Note in 100% case, the loss $\mathcal{L}_{un}$ and $\mathcal{L}_{mix}$ are applied to all data. The results in Table 3 indicate SVFormer can still improve the supervised baseline with more labeled data.

Besides, following MvPL [3], we warm up the training with only labeled data in the first few epochs, which ensures a stable start. As shown in Figure 1, though EMA performs lower in the beginning, it achieves higher results and is more stable in the end.

To investigate the effectiveness of different pretrain data, we do ablation study on UCF-101 datasets with different pretrain methods in Table 4. It can be clearly seen that the pre-training of ImageNet is crucial to performance. At the same time, if large-scale Kinetics-400 is used for pre-training, it will also greatly help improve the performance of low-labeled datasets.

D. More Visualizations

In this section, we present more visualization of the different mixing methods in Fig 2. We show the examples of Tube TokenMix strategy with three pixel-level mixing methods, CutMix [5], Mixup [6], PixMix [2], as well as the other two token-level mixing methods, *i.e.*, Frame TokenMix and Rand TokenMix.

Table 1. **Comparison of efficiency.** We show the efficiency comparison of our SVFormer and previous state-of-the-art methods. The results are reported on Kinetics-400 and UCF-101 with 1% labeling ratio.

Method	Backbone	Input	Frames	Epochs	Params (M)	Flops (G)	Infer. View	UCF-1%	Kinetics-1%
MvPL [3]	3D-R50	V+F+G	8	600	32.5	54.5	10×3	22.8	17.0
CMPL [4]	R50+R50-1/4	V	8+16	200	34.6	54.5	10×3	25.1	17.6
SVFormer-S	ViT-S	V	8	**30**	**30.7**	**50.8**	$\mathbf{5 \times 3}$	**31.4**	**32.6**
SVFormer-B	ViT-B	V	8	**30**	121.4	196.6	$\mathbf{5 \times 3}$	**46.3**	**49.1**

Table 2. **Comparison of different attention mechanisms.** The results are reported on Kinetics-400 and UCF-101 with 1% and 10% labeling ratios at SVFormer-S.

Attention Mechisams	Flops (G)	UCF-101		Kinetics-400	
		1%	10%	1%	10%
Space-only	38.6	30.5	77.6	31.1	58.6
Divided Space-Time	50.8	**31.4**	**79.1**	**32.6**	**61.6**
Joint Space-Time	58.5	28.4	78.5	27.3	60.9

Table 3. **Comparison of more labeled data.** The results are reported on UCF101 and HMDB51 with 50% and 100% labeling ratios. We show the Top-1 accuracy here.

	UCF-50%	UCF-100%	HMDB-50%	HMDB-100%
Supervised	78.4	85.2	53.7	60.8
SVFormer-S	82.3	86.4	58.2	61.6

Table 4. **Comparison of different pretrain data.** The results are reported on UCF-101 with 1% and 10% labeling ratios at SVFormer-S.

	pretrain	UCF-1%	UCF-10%
Supervised	-	5.3	27.5
Supervised	ImageNet	12.7	62.5
SVFormer-S	ImageNet	31.4	79.1
SVFormer-S	Kinetics-400	35.6	84.5

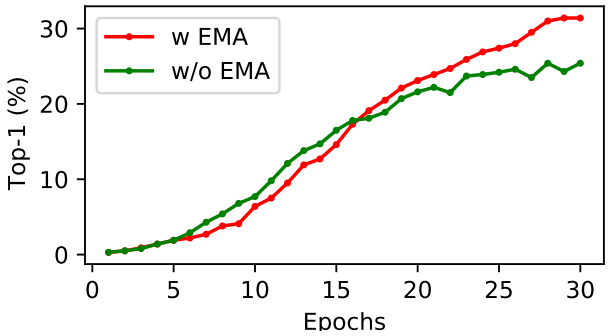

Figure 1. Training accuracy curves with or w/o EMA on UCF-1%. We show the Top-1 accuracy at SVFormer-S here.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 1

[2] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *CVPR*, 2022. 1

[3] Bo Xiong, Haoqi Fan, Kristen Grauman, and Christoph Feichtenhofer. Multiview pseudo-labeling for semi-supervised learning from video. In *ICCV*, 2021. 1, 2

[4] Yinghao Xu, Fangyun Wei, Xiao Sun, Ceyuan Yang, Yujun Shen, Bo Dai, Bolei Zhou, and Stephen Lin. Cross-model pseudo-labeling for semi-supervised action recognition. In *CVPR*, 2022. 1, 2

[5] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 1

[6] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1

Figure 2. Example of the traditional pixel-level mixing methods and our proposed token-level mixing.