Temporal Action Detection with Multi-level Supervision: Appendix

1. Experimental Settings and Additional Results

1.1. Hyperparameters

The weights of completeness loss and regression loss in SSN are $\alpha_c^U = \alpha_r^U = 0.1$. The weight of unsupervised loss is $\alpha^U = 0.3$ for Mean Teacher and Fixmatch, and 10 for Mixmatch. The weight of weakly-supervised loss is $\alpha^W = 1$. The loss weight for UFA and IB are 0.003 and 0.01, respectively.

We implemented our method with PyTorch [1] and trained on four NVIDIA Titan RTX GPUs. We load 16 videos in each mini-batch and sample 8 proposals from each video. The whole network is optimized using SGD with the learning rate of 0.0003 for THUMOS14 and 0.0001 for ActivityNet1.2.

1.2. Estimation of Annotation Cost

We estimate the annotation cost for full and weak supervision of THUMOS14 based on a previous user study. THUMOS14 contains only 200 training videos collected from Youtube, which may be too few to obtain an accurate estimation of the annotation cost. In that user study, we collect similar videos from Youtube in a larger number of $\sim 7,000$, ask the users to give full or weak annotation, and record the annotation time. On average, it costs 31.2h to fully annotate 800 videos and 29.8h to weakly annotate 6,000 videos. Then the annotation costs for full and weak supervision are 140.4 sec/vid and 17.88 sec/vid separately, with the ratio of about $8:1$.

1.3. Optimal Annotation Strategy under Different Ratios of Annotation Cost

In the paper we show the advantage of using multi-level supervision over single supervision when there is a fixed annotation budget. The result depends on the ratio of the annotation cost for different levels of supervision, which is estimated from our user study (Sec. 1.2). To demonstrate that our result is consistent under different cost ratios, we also test on the ratio of $4:1$ (full:weak) and show the results in Table 1. We can observe a similar trend that a mixed annotation strategy with multi-level supervision is more efficient.

Table 1: Annotation policies under cost ratio (full:weak) of 4:1.

Policy	$\lvert\mathcal{S}\rvert$ (%)	$\lvert\mathcal{W}\rvert$ (%)	$\lvert\mathcal{U}\rvert$ (%)	0.3	0.5	0.7
Full	30%	0%	70%	38.90	18.97	5.04
Mixed	20%	40%	40%	**39.38**	**20.41**	**5.90**
	10%	80%	10%	34.39	15.26	3.76
Weak	7%	93%	0%	33.57	15.44	3.38

1.4. Ablation on Video Augmentation

The ablation on different augmentations of video data is shown in Table 2. We test the SSAD baseline with Mixmatch on THUMOS14, with 50% labeled data and 10% labeled data. In 10% labeled case, using all the five augmentations gives the best results, while in 50% labeled case, the results are less sensitive to the different augmentations. In all our experiments, we use the five augmentations together by default.

1.5. Qualitative Results of UFA

Fig. 1 shows additional visualization of the attention from the proposed UFA module.

Table 2: Ablations of video augmentations.

Random Noise	Horizontal Flip	Temporal Resampling	FPS	Temporal Flip	50%	10%
✓	-	-	-	-	24.74	7.29
✓	✓	-	-	-	24.72	7.56
✓	✓	✓	-	-	24.79	7.93
✓	✓	✓	✓	-	24.75	8.02
✓	✓	✓	✓	✓	24.79	8.18

References

[1] 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. In *Advances in neural information processing systems*, pages 8026–8037, 2019. 1