MUM : Mix Image Tiles and UnMix Feature Tiles for Semi-Supervised Object Detection Supplemental Materials

JongMok Kim[1,2] JooYoung Jang[1,2] Seunghyeon Seo[2] Jisoo Jeong[2] Jongkeun Na[1] Nojun Kwak[2]
[1]SNUAILAB [2]Seoul National University

A. Training details and stability

In this section, we provide training details of hyperparameters used in our experiments, as show in Table. 2. The most of parameters are from the Unbiased Teacher for the sake of fair comparison. since MUM is easy to add to any framework and prevent losing semantic information, our training process based on the Unbiased Teacher [27] was stable, and the training accuracy curve rose upward with slight variance (see Fig. 1). Thanks to this, we have conducted our experiments with the default hyperparameters in the Unbiased Teacher [27] and could verify that our MUM works well as an add-on to the baseline SSOD work and achieves better performance in fair comparison.

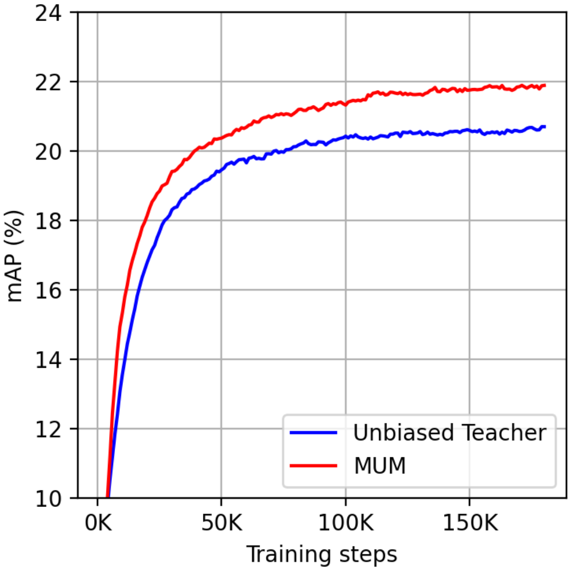

Figure 1. Training curves of the Unbiased Teacher and our MUM. MUM shows a better training curve with no more variance comparing the Unbiased Teacher.

B. Relation between N_T and foreground-image ratio

The number of tiles (N_T) correlates with the foreground-image ratio, because the degree of splitting and occluding the foreground object is simultaneously related to the foreground object size and tile size. In order to further investigate the correlation between foreground-image ratio and N_T, here we define a new parameter, N_O as the average number of tiles where a single foreground object lies. The relationship between the gain of mAP and N_O with various N_T and object size in the COCO validation dataset is summarized in Table 1 and Fig. 2

We find $1.2 \leq N_O \leq 2.5$ is an acceptable range for the AP gain. It is reasonable that too small N_O (< 1.2) means that augmentation is not enough, while too large N_O (> 2.5) tells that the foreground object is teared into too many pieces. This explains our choice of the tile size, N_T=4, was a reasonable. The above experiment implies that N_T should be adjusted for the image and foreground object's resolution depending on the N_O and MUM can be enhanced by sophisticatedly generating the mixing mask.

Table 1. The relationship between N_O and AP with varying N_T and object size.

	N_O / AP / AP gain		
N_T	Small	Medium	Large
1 (baseline)	1.00/8.93/0	1.00/21.85/0	1.00/28.07/0
2	1.13/9.33/+0.40	1.38/22.93/+1.08	2.48/28.60/+0.53
4	1.29/9.86/+0.93	1.96/23.66/+1.81	6.19/27.91/-0.16
8	1.65/9.72/+0.79	3.42/22.33/+0.48	17.9/27.12/-0.95

Table 2. Hyperparameters for various protocols

Hyperparameters	Description	COCO-Standard	COCO-Additional	VOC12	VOC12+COCO20cls	Swin
τ	Confidence threshold of pseudo label	0.7	0.7	0.7	0.7	0.7
λ_u	Unsupervised loss weight	4.0	2.0	4.0	4.0	4.0
δ	EMA decay rate	0.9996	0.9996	0.9996	0.9996	0.999
p	Percentage of applying MUM	1.0	1.0	1.0	1.0	1.0
Optimizer	Training optimizer	SGD	SGD	SGD	SGD	SGD
LR	Initial learning rate	0.01	0.01	0.01	0.01	0.01
Momentum	SGD momentum	0.9	0.9	0.9	0.9	0.9
Weight Decay	Weight decay	1e-4	1e-4	1e-4	1e-4	1e-4
Training Steps	Total training steps	180K	360K	45K	90K	60K
Batch Size	Batch size	32	32	32	32	16

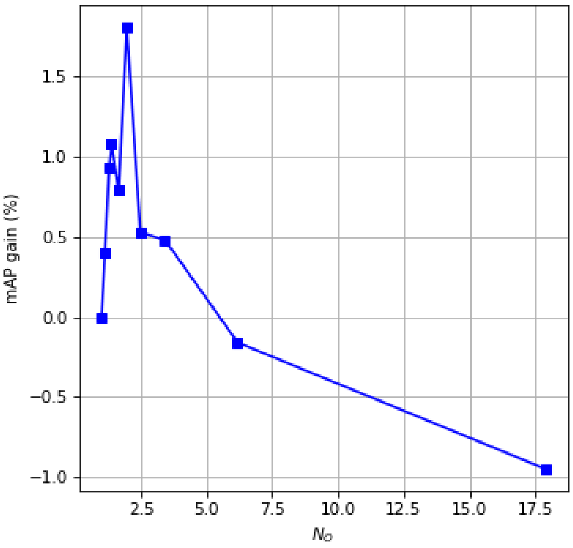

Figure 2. N_T vs. gain of mAP in the COCO validation dataset