Supplementary Material for
Mutual Supervision for Dense Object Detection

Ziteng Gao Limin Wang Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, China

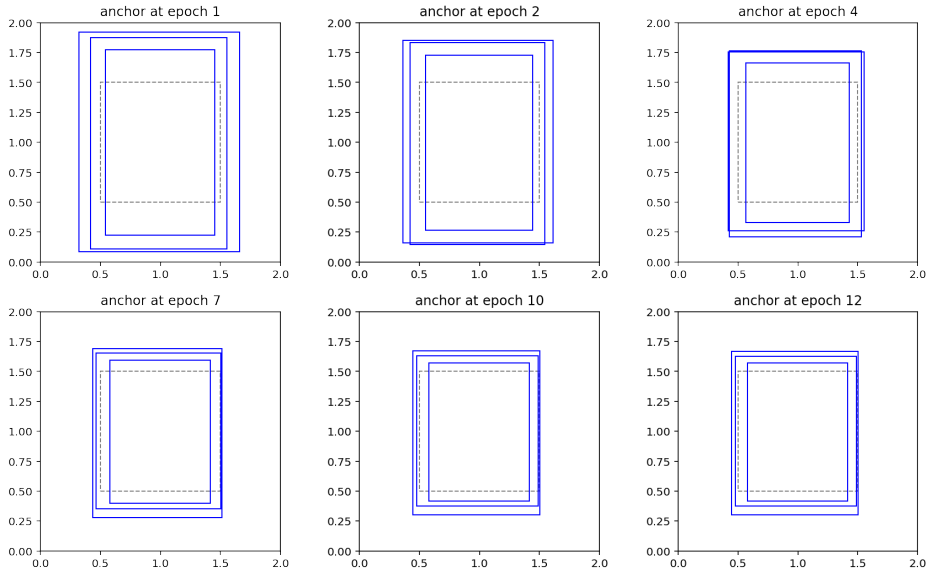

Figure 1: Multiple anchors visualized by the bias parameter of the last convolution to produce bounding boxes during the training procedure by the MuSu. Dashed boxes stand for basic spatial units on the feature map. One can see, when converged, anchors are specific to different scales and aspect ratios even though during training (*e.g.*, at the epoch 4) several anchors are similar. Another key ingredient of multiple anchor settings, the weight parameter of the last convolution layer to bounding boxes, also can lead to specific preference to different scales and aspect ratios.

I. Details of Learned Anchors

In the paper, we showed that multiple anchor settings benefit the MuSu supervision scheme and MuSu actually enabes the network to group anchors into different scales and aspect ratios. Here, we show details of learned anchors when $\#A = 3$ trained by MuSu in Figure 1.

II. Computational Complexity

It is worth noting that our MuSu models with 1 anchor per location ($\#A = 1$) on the feature map share the **exact same** number of parameters and FLOPs with the FCOS models since we do not add any new modules for MuSu models. For the multiple anchor variants, the parameters and FLOPs increase moderately when adding anchors, as shown in Table 1. The FPS drops slightly mainly since multiple anchors incur inevitable more items processed in the time-costing NMS procedure.

Models	#Params	FLOPs	FPS
FCOS	32.07M	205.3G	14.6
MuSu ($\#A = 1$)	32.07M	205.3G	14.6
MuSu ($\#A = 2$)	32.27M	209.5G	13.2
MuSu ($\#A = 3$)	32.46M	213.7G	12.5
MuSu ($\#A = 4$)	32.66M	217.8G	11.7
MuSu ($\#A = 5$)	32.85M	222.0G	11.0

Table 1: Comparisons of parameters and FLOPs with the backbone ResNet-50. The FLOPs of models are calculated with the input shape $(1333, 800)$. The FPS is measured by a single Titan Xp.

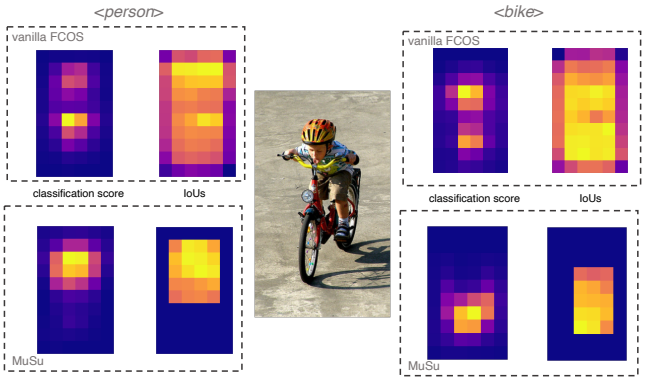

Figure 2: More visualizations of the consistency between heads. Note that the vanilla FCOS also faces the entanglement of classification scores of different classes, which our MuSu is exempt from.

III. Visualization

Consistency. As shown in Figure 1 in the main paper, our proposed MuSu alleviates the inconsistency between the classification and regression head suffered from the FCOS detector. Here, we provide more visualizations about this, shown in Figure 2.

Visualization of detection results. We present the detection results of the MuSu model with $\#A = 3$ and R-50 backbone in Figure 3. We can find that the MuSu-trained detector tends to relate the classification score to how well the predicted bounding box localizes (especially in multiple people cases).

Figure 3: Visualization of detection results of the MuSu model with $\#A = 3$ and R-50 backbone on COCO `minival` split. Better zoom in.