AugDETR: Improving Multi-scale Learning for Detection Transformer Supplementary Material

Jinpeng Dong, Yutong Lin, Chen Li, Sanping Zhou, Nanning Zheng*

National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
{djp1235a, yutonglin, edward82}@stu.xjtu.edu.cn
{spzhou, nnzheng}@xjtu.edu.cn

In the supplementary material, we conduct additional ablation experiments and present the visual analysis of detection results.

1 More Ablations

Table 1: Comparisons of DINO with our lite variants.

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Flops
DINO [1]	49.0	66.6	53.5	32.0	52.3	63.0	245G
AugDETR	50.2	67.8	55.0	32.3	53.2	64.7	287G
Lite AugDETR	50.0	67.7	54.7	31.9	53.0	64.8	257G

The results of lite variants. To reduce the computational complexity of our method, we explore the lite designs of our AugDETR. The results are shown in Table 1. Our lite AugDETR reduces the Flops significantly when decreasing the performance by 0.2 AP. The lite AugDETR is implemented by feeding only the first and last stage encoders into the EMCA module.

More quantitative analysis of EMCA. We first introduce the scale intervals for analyzing the weights of objects at different scales in the main paper. Size #1 means $0 \sim 32 \times 32$, Size #2 means $32 \times 32 \sim 128 \times 128$, Size #3 means $128 \times 128 \sim 224 \times 224$, Size #4 means $224 \times 224 \sim 320 \times 320$, Size #5 means $320 \times 320 \sim 416 \times 416$, Size #6 means $416 \times 416 \sim \infty$. Then we analyze the weights of the different heads. We counted the head weights of objects at different scales and the head weights of different encoders. As shown in Figure 1, different heads have different weights. The head weights also have different distributions for different objects and different encoders.

* Corresponding author.

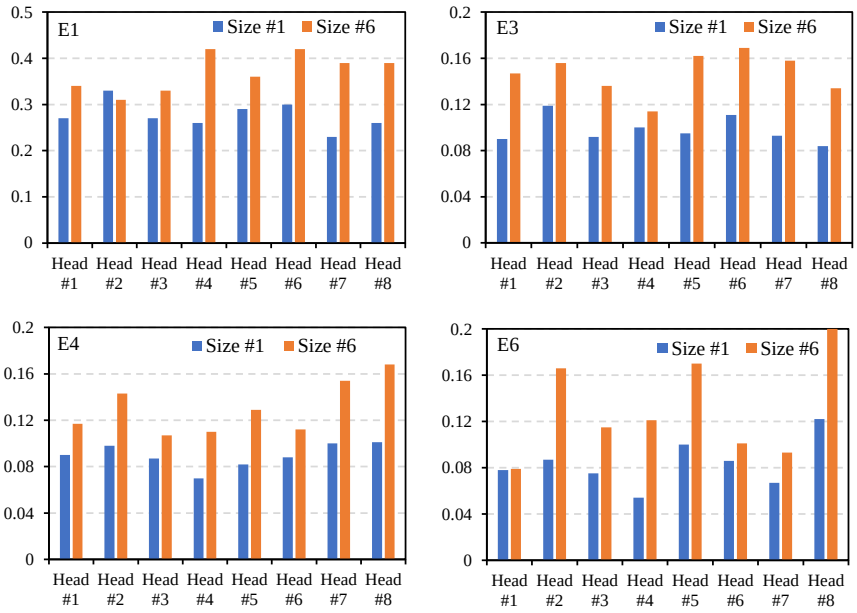

Fig. 1: Statistical fusion weights for different heads under different objects and encoders.

2 Visualization of Detection Results.

As shown in Figure 2 and Figure 3, we provide comparisons of object detection results with the score threshold of 0.5 between the baseline and our method on COCO val2017. It can be seen that our method can detect the large objects overlooked by the baseline methods.

References

1. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In: The Eleventh International Conference on Learning Representations (2022)

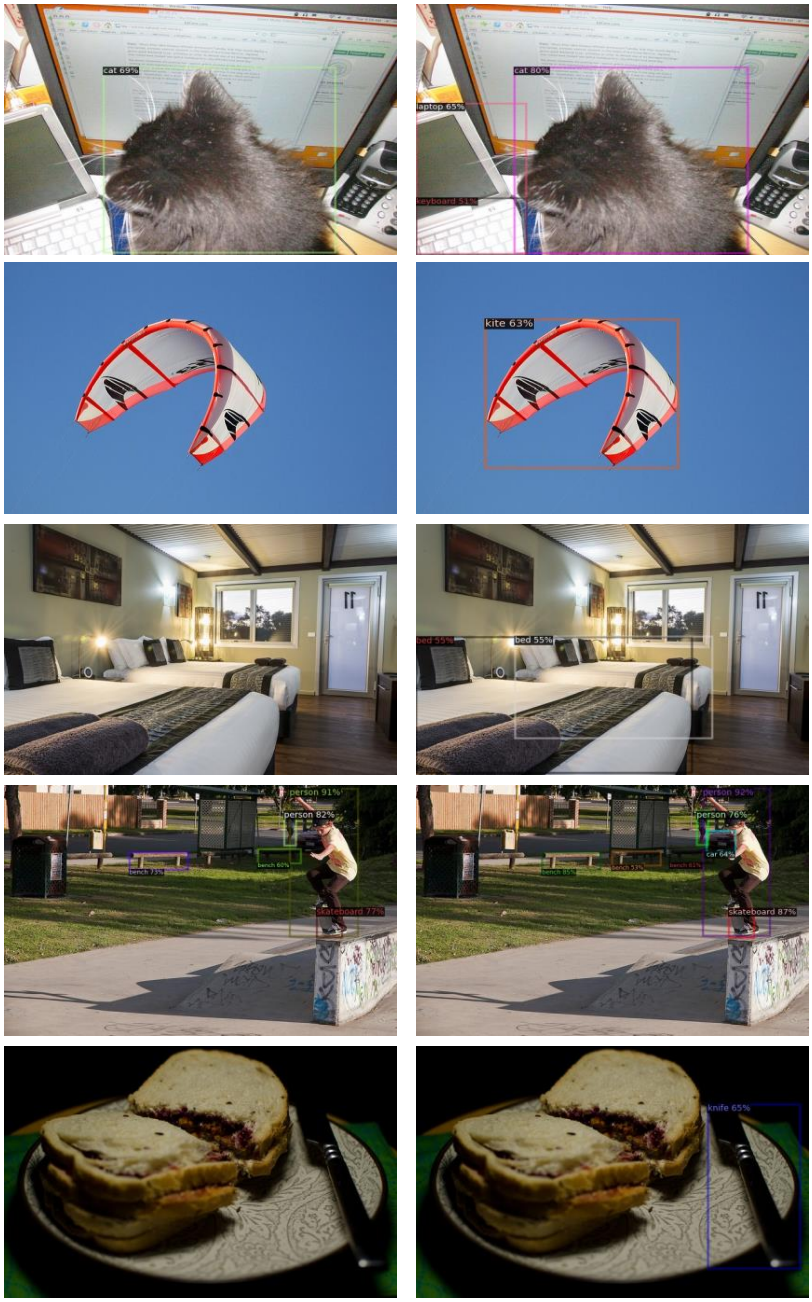

Fig. 2: Comparison of detection results between our method and baseline with ResNet-50 on COCO val2017. The left images are the results of the baseline. The right images are the results of our method. The score threshold is set to 0.5.

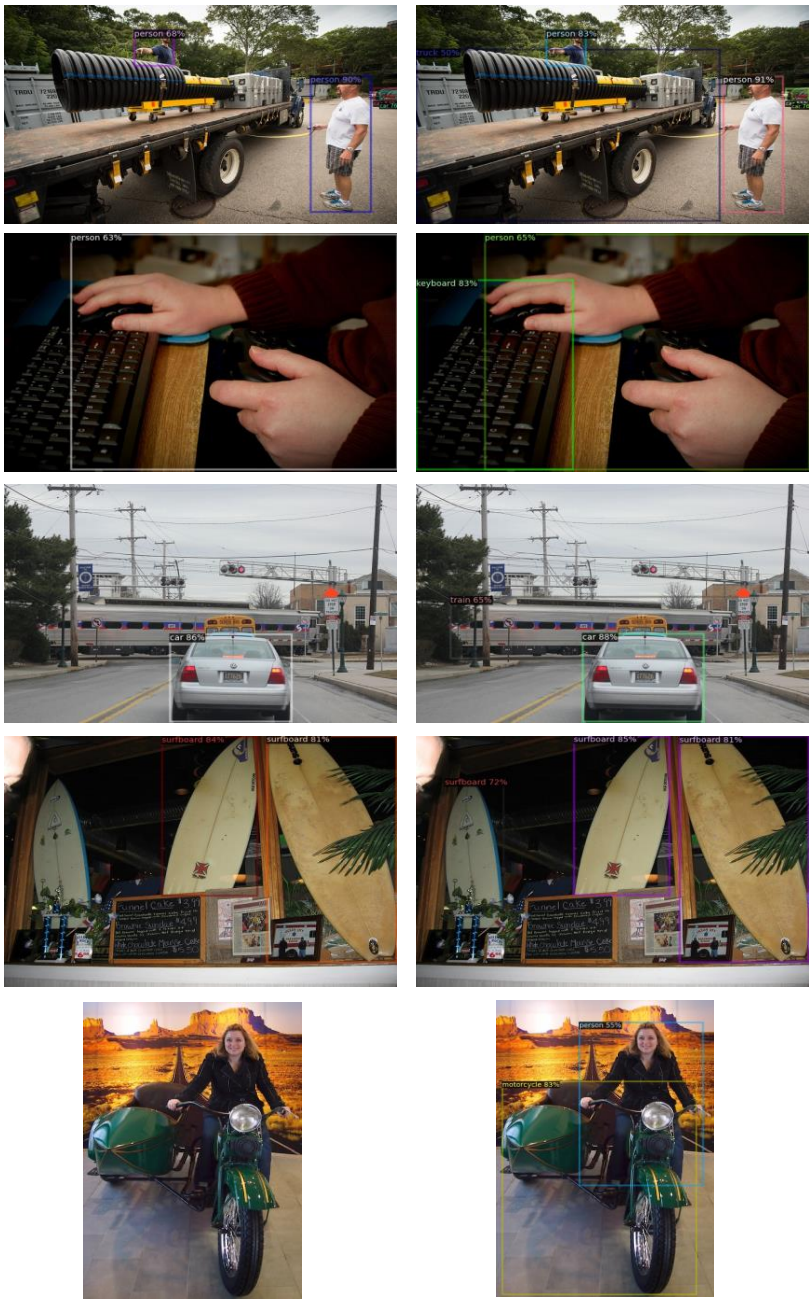

Fig. 3: Comparison of detection results between our method and baseline with ResNet-50 on COCO val2017. The left images are the results of the baseline. The right images are the results of our method. The score threshold is set to 0.5.