Efficient Decoder-free Object Detection with Transformers

Peixian Chen[1†], Mengdan Zhang[1†], Yunhang Shen[1], Kekai Sheng[1], Yuting Gao[1], Xing Sun[1], Ke Li[1⋆], and Chunhua Shen[2]

[1]Tencent Youtu Lab, [2]Zhejiang University
{peixianchen, davinazhang, saulsheng, yutinggao, tristanli}@tencent.com
{shenyunhang01, winfred.sun, chhshen}@gmail.com

Fig. 1: Supplementary figures for Figure 6 in the main paper.

1 Supplementary Figures

Figures 1 and 2 illustrate more detection results in addition to Figure 6 in the main paper. As we can observe, our proposed TAE makes the best anchors for

⋆ Corresponding author
† Equal Contribution.

classification (red) and localization (orange) closer than YOLOF (blue boxes and centers indicate the ground truth). It shows that TAE can handle task conflicts in a coupled head and further generates task-aligned predictions in a single pass.

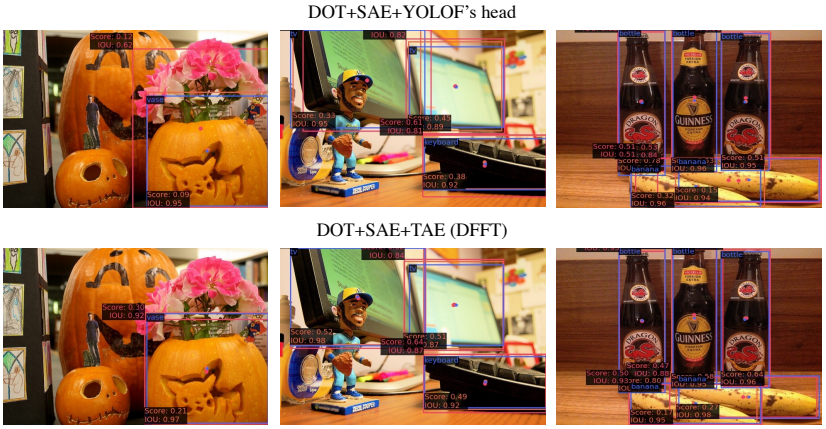

Fig. 2: Supplementary figures for Figure 6 in the main paper.

Table 1: The performance of different GCA position in DOT Block. We report the accuracy of the pretrained backbone on ImageNet and the detection AP with YOLOF after training on the MS COCO dataset.

Position	Results (%)	
	Accuracy	AP
w/o GCA	80.0	32.0
Top	80.5	33.8
Middle	80.7	30.0
Bottom	81.2	34.1

2 Supplementary Experiments

We provide a supplementary experiment for our detection-oriented transformer (DOT) backbone. In this experiment, we conduct ablation studies of the *position* of global channel-wise attention (GCA) within the DOT block. For this purpose, we use the YOLOF framework and replace the backbone network. The results are shown in Table 1. The results show that placing GCA at the bottom achieves

the best performance. This observation indicates GCA's strong capability of aggregating local features. As a result, placing GCA at the end of DOT blocks can aggregate local features from previous layers and extract corresponding global semantic information.

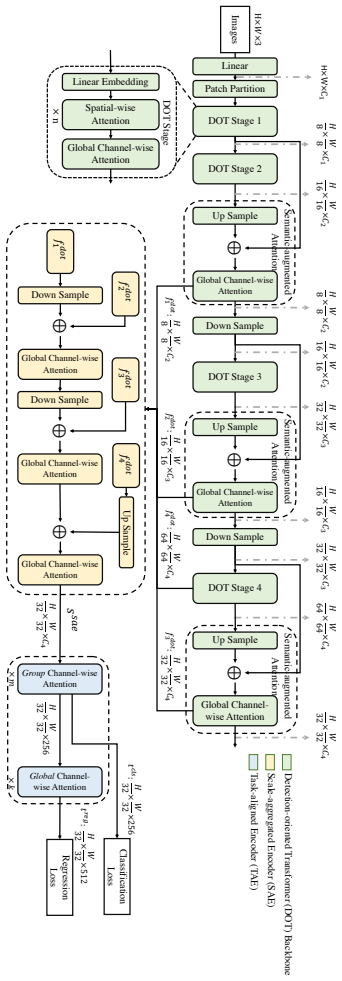

Fig. 3: The detailed architecture of DFFT, including how the feature shape changes during the forward propagation.