LEARNING DYNAMIC QUERY COMBINATIONS FOR TRANSFORMER-BASED OBJECT DETECTION AND SEGMENTATION SUPPLEMENTARY

Anonymous authors
Paper under double-blind review

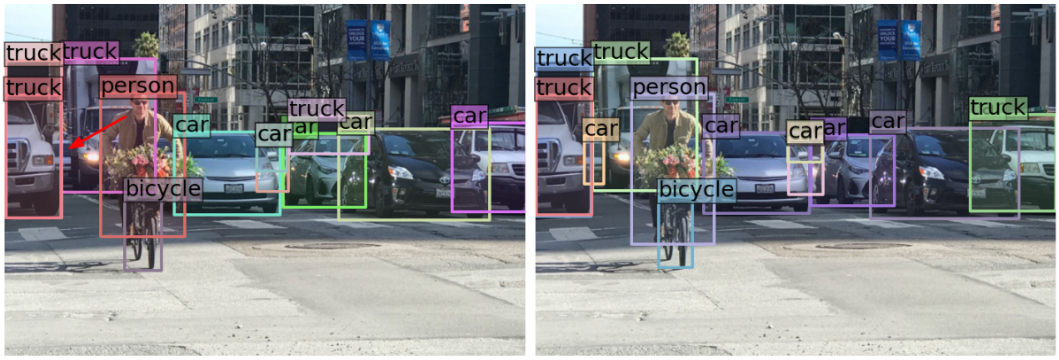
1 TRAINING DETAILS

For a fair comparison, all the models with and without our proposed methods are trained with identical experimental setups. All the models are trained on 8 Nvidia A100 GPUs. For the training details like hyperparameters and training pipelines, we follow the original models. For example, for the video instance segmentation task, all the models are first pretrained on the MS COCO (Lin et al., 2014) benchmark and then finetune on the YouTube-VIS (Yang et al., 2019) dataset. For SeqFormer (Wu et al., 2021), the models are jointly trained on both MS COCO and YouTube-VIS datasets, following their pipelines. All the models are trained on the training split and then evaluated on the validation split.

2 VISUALIZATION

Object detection. We use DAB-DETR (Liu et al., 2022) as an example to visualize and compare the performance of the object detection task on MS COCO `val` benchmark (Lin et al., 2014) as Figure 1. For better visualization, we use red arrows to point out the objects that are not detected correctly: one car is not detected. In contrast, our DQ-DAB-DETR can accurately detect those objects.

Instance/Panoptic segmentation. We visualize several examples and compare them between Mask2Former (Cheng et al., 2021) and DQ-Mask2Former as Figure 2. For better visualization, we use red rectangles to highlight where the original Mask2Former does not perform an accurate prediction. From the figure, our DQ-Mask2Former can effectively generate higher-quality masks for instance/panoptic segmentation, compared to the original Mask2Former model. For instance segmentation, our DQ-Mask2Former can generate smooth and complete masks with high qualities compared to the original Mask2Former, as shown in the red rectangles in Figure 2. For panoptic segmentation, the improvement of our DQ-Mask2Former is more obvious. As shown in Figure 2, the original Mask2Former cannot correctly detect and generate accurate masks for the category of tree compared with our DQ-Mask2Former.



DAB-DETR DQ-DAB-DETR

Figure 1: Quantitative comparison examples of DAB-DETR/DAB-Deformable-DETR and DQ-DAB-DETR/DQ-DAB-Deformable-DETR on the object detection task on MS COCO (Lin et al., 2014) benchmark with ResNet-50 as the backbone.

Mask2Former DQ-Mask2Former

(a) Instance Segmentation

Mask2Former DQ-Mask2Former

(b) Panoptic Segmentation

Figure 2: Quantitative comparison examples of Mask2Former(Cheng et al., 2021) and DQ-Mask2Former on the instance/panoptic segmentation task on MS COCO (Lin et al., 2014) benchmark with ResNet-50 as the backbone.

REFERENCES

Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oMI9PjOb9Jl.

Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2021.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5188–5197, 2019.