Integrally Migrating Pre-trained Transformer Encoder-decoders for Visual Object Detection Supplementary Material

Feng Liu[1*] Xiaosong Zhang[1*] Zhiliang Peng[1] Zonghao Guo[1]
Fang Wan[1†] Xiangyang Ji[2] Qixiang Ye[1]
[1]University of Chinese Academy of Sciences [2]Tsinghua University
liufeng20@mails.ucas.ac.cn zhangxiaosong18@mails.ucas.ac.cn
pengzhiliang19@mails.ucas.ac.cn guozhonghao19@mails.ucas.ac.cn
wanfang@ucas.ac.cn xyji@tsinghua.edu.cn qxye@ucas.ac.cn

1. Details in MAE Pre-training

There is no experiment involving ViT-S in MAE [2], so we refer to the decoder design of ViT-B and scale the decoder settings of ViT-S proportionally. In our experiments, both the encoder and decoder of ViT-S are about one-fourth computational cost of the corresponding model in ViT-B. We give the details of ViT-S, ViT-B and ViT-L in Table 7. Since the ViT-S is mainly used to ablate the effectiveness of imTED, we only pre-train ViT-S for 800 epochs, while ViT-B and ViT-L are pre-trained for 1600 epochs as MAE [2] does.

Table 7: The details of encoder and decoder in MAE pre-training.

Model	Encoder			Decoder		
	Depth	Dim	Heads	Depth	Dim	Heads
ViT-S	12	384	6	4	256	8
ViT-B	12	768	12	8	512	16
ViT-L	24	1024	16	8	512	16

2. Single Object Detection

We give the details of single object detection in the motivation here. Kaggle held a ImageNet Object Localization Challenge, providing the original dataset for our single object detection task. Datasets can be download from the link[1]. It contains 544546 images and 50000 images in the training set and validation set separately, with some images containing multiple objects. We filter out the images with more than 2 objects, obtaining 494264 images and 38285 images in the training set and validation set separately. We choose the MAE pre-trained ViT-B as the model. The training lasts for 3×schedule (36 epochs with the learning rate decayed by 10 at epochs 27 and 33). The batch size is 128, distributed across 8 GPUs (16 images per GPU). A learing rate of 0.001, a layer-wise lr decay [1] of 0.75 and a drop path rate of 0.2 are also applied.

3. Detection Examples under Few-shot Setting

We show some detection examples under the few-shot setting in Fig. 7. The compared state-of-the-art detector (DeFRCN [3]) is observed to miss objects when multiple objects come together, which could cause occlusion and inferences between object. In contrast, the proposed imTED detector can successfully detect most of the objects, demonstrating the generalization capability to these complex scenarios.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[3] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *IEEE ICCV*, pages 8681–8690, 2021.

*Equal Contribution.
†Corresponding Author.
[1]https://www.kaggle.com/competitions/imagenet-object-localization-challenge/data

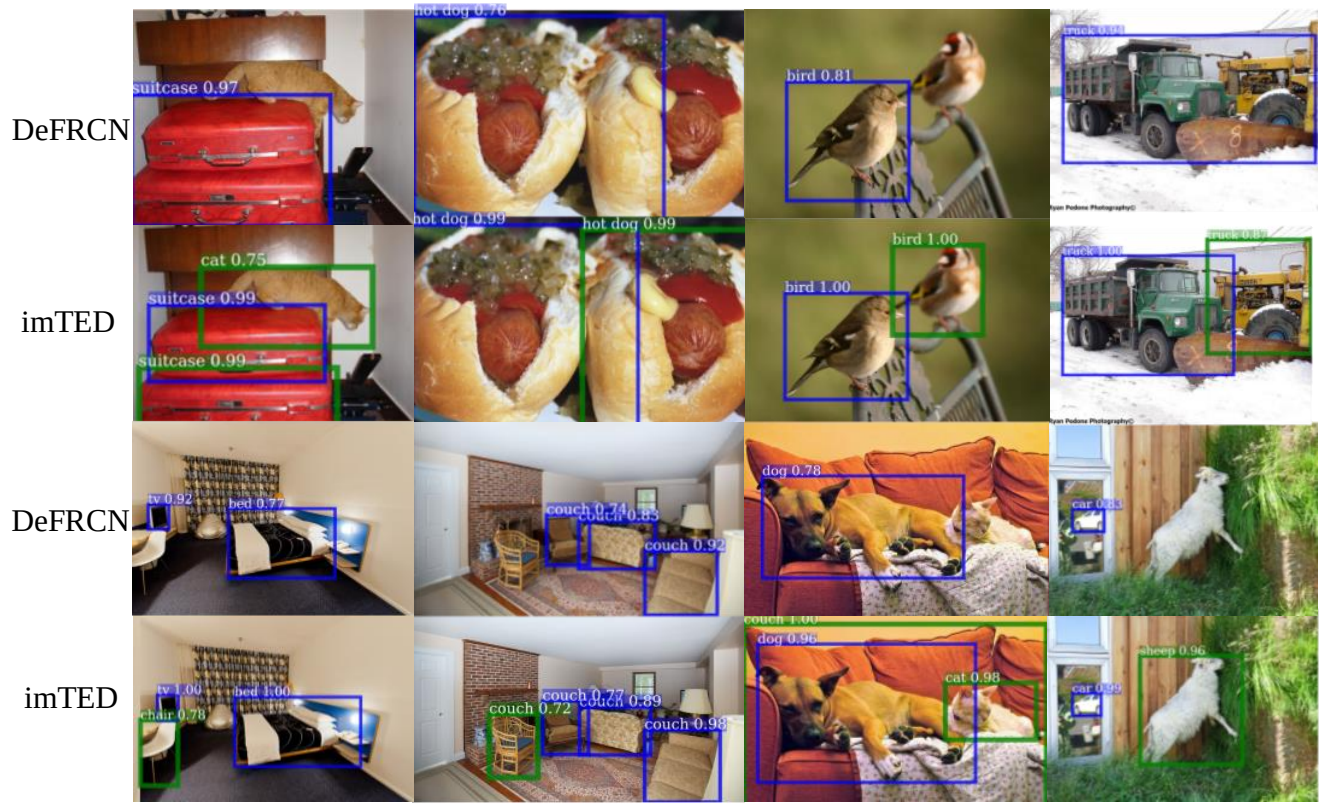

Figure 7: Comparison of detection examples under the few-shot setting. The first and third row present the results of DeFRCN [3] and the second and fourth row present the results of the proposed imTED detector. Objects in blue bounding boxes are detected by both detectors. Objects in green bounding boxes are missed by DeFRCN while correctly detected by imTED. (Best viewed in color)