UP-DETR: Unsupervised Pre-training for Object Detection with Transformers
(Supplementary Material)

Appendix

1. More ablations

1.1. Single-Query Patch *vs*. Multi-Query Patches

We pre-train two UP-DETR models with single-query patch ($M = 1$) and multi-query patches ($M = 10$). The other hyper-parameters are set as mentioned in the paper.

Table 1 shows the results of single-query patch and multi-query patches. Compared with DETR, UP-DETR surpasses it in all AP metrics by a large margin no matter with single-query patch or multi-query patches. When pre-training UP-DETR with the different number of query patches, UP-DETR ($M = 10$) performs better than UP-DETR ($M = 1$) on the fine-tuning task, although there are about 2.3 instances per image on VOC. Therefore, we adopt the same UP-DETR with $M = 10$ for both VOC and COCO instead of varying M for different downstream tasks.

Model	AP	AP_{50}	AP_{75}
DETR	49.9	74.5	53.1
UP-DETR (M=1)	53.1 (+3.2)	77.2 (+2.7)	57.4
UP-DETR (M=10)	**54.9** (+5.0)	**78.7** (+4.2)	**59.1**

Table 1: The ablation results of pre-training models with single-query patch and multi-query patches on PASCAL VOC. The values in the brackets are the gaps compared to the DETR with the same training schedule.

1.2. Attention Mask

After downstream task fine-tuning, we find that there is no noticeable difference between the UP-DETR pre-trained w/ and w/o attention mask. So, we plot the loss curves in the pretext task to illustrate the effectiveness of attention mask.

As shown in Fig. 1, at the early training stage, UP-DETR without attention mask has a lower loss. However, as the model converging, UP-DETR with attention mask overtakes it with a lower loss. It is reasonable because the loss is calculated by the optimal bipartite matching. During the early training stage, the model is not converged, and the model without attention mask takes more object queries into attention. Intuitively, the model is easier to be optimized due to introducing more object queries. However, there is a mismatching between the query patch and the ground truth for the model without attention mask. As the model converging, the attention mask gradually takes effect, which masks the unrelated query patches and leads to a lower loss.

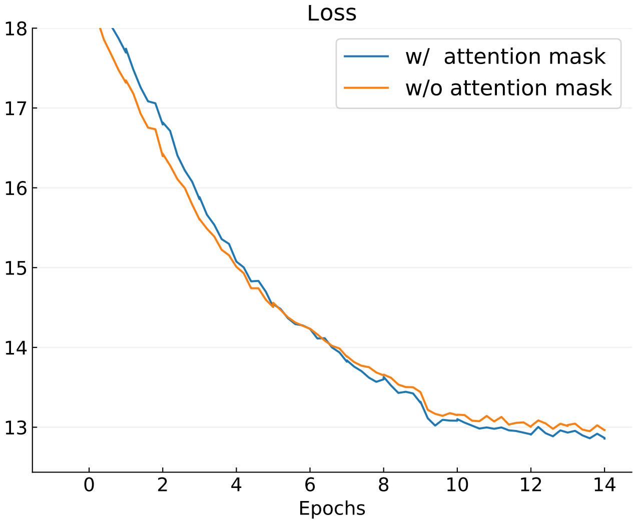

Figure 1: The loss curves of pre-training procedure for UP-DETR w/ and w/o the attention mask.

1.3. Object Query Shuffle

Without object query shuffle, the groups of object queries are assigned fixedly during the pre-training. However, for the downstream object detection tasks, there is no explicit group assignment between object queries. So, we design the object query shuffle to simulate implicit grouping between object queries.

The motivation of object query shuffle is clear, however, we find that object query shuffle is not helpful. In the pre-training and fine-tuning phase, the model w/o object query shuffle converges faster. Fig. 2 shows the fine-tuning result of COCO w/ and w/o object query shuffle. As seen, without object query shuffle, the model converges faster and achieves **43.1** AP (higher than 42.8 AP with ob-

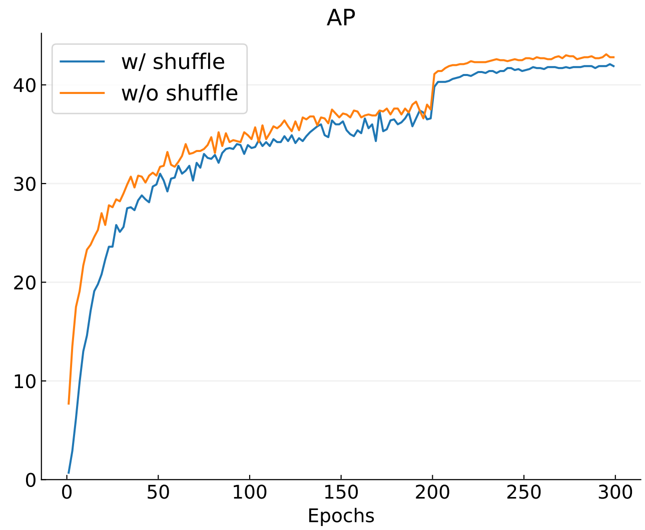

Figure 2: The AP curves of COCO fine-tuning procedure for UP-DETR w/ and w/o the object query shuffle. The learning rate is reduced at 200 epochs.

ject query shuffle pre-training). The result indicates that fixed group is beneficial for training object queries. Shuffle may disturb the spatial preference learning. Therefore, in our open-source code (https://github.com/dddzg/up-detr), we upload the pre-training model without object query shuffle.