Less is More: Focus Attention for Efficient DETR
Supplementary Material

Dehua Zheng[1,2] Wenhui Dong[2] Hailin Hu[2] Xinghao Chen[2] Yunhe Wang[2*]
[1]Huazhong University of Science and Technology [2]Huawei Noahs Ark Lab

dwardzheng@hust.edu.cn {wenhui.dong, hailin.hu, xinghao.chen, yunhe.wang}@huawei.com

A. More Implementation Details

A.1. Cascade Structure

In order to increase the fault tolerance of our model, we gradually reduce the scope of foreground regions through a cascade structure. As we show in Section 3.4, the computational complexity of deformable attention [6] is linear with the number of preserved tokens. Therefore, there is no significant difference in complexity between the even structures (e.g., {0.4,0.4,0.4,0.4,0.4,0.4} and the cascade structures(e.g.,{0.65,0.55,0.45,0.35,0.25,0.15}). Table 1 lists different average $keep - ratio$ and corresponding ratios of different layers designed in this paper.

Average $keep - ratio$	Ratios
0.1	{0.1, 0.1, 0.1, 0.1, 0.1, 0.1}
0.2	{0.3, 0.3, 0.2, 0.2, 0.1, 0.1}
0.3	{0.5, 0.4, 0.3, 0.3, 0.2, 0.1}
0.4	{0.65,0.55,0.45,0.35,0.25,0.15}
0.5	{0.75,0.65,0.55,0.45,0.35,0.25}

Table 1: Detailed cascade keep-ratio desiged by Focus-DETR.

A.2. Label Assignment

Unlike the traditional label assignment scheme for multiscale feature maps, the ranges are allowed to overlap between the two adjacent feature scales to enhance the prediction near the boundary. This strategy increases the number of foreground samples while ensuring that the multiscale feature map predicts object heterogeneity. Intuitively, we assign the interval boundaries to be a series of integer power of two. As shown in Table 2, our overlapping interval setting improves the detection accuracy of the model when compared to non-overlapping ones using similar interval boundaries.

Model	Interval	AP	AP_{50}	AP_{75}
non-overlapping	{[-1, 64], [64, 128], [128, 256], [256, ∞]}	50.2	68.2	54.9
	{[-1, 128], [128,256], [256,512], [512, ∞]}	50.2	68.1	54.8
overlapping	{[-1, 64], [64, 256], [128, 512], [256, ∞]}	50.4	68.5	55.0

Table 2: Effect of preset scale intervals of multi-scale feature maps on experimental performance. Interval represents different scale intervals of multi-scale feature maps and $\infty = 999999$ in experiments.

B. Supplementary Experiments

B.1. Using Swin Transformer as the Backbone

When using Swin Transformer [3] as the backbone, Focus-DETR also achieves excellent performance. As shown in the following table, when Focus-DETR uses Swin-T as the backbone, the AP reaches 51.9 and achieve 56.0AP using Swin-B-224-22K and 55.9AP using Swin-B-384-22K. Compared with Deformable DETR [6] and Sparse DETR [4], our model achieves significant performance improvements, as shown in Table 3.

B.2. Convergence Analysis

In order to better observe the changes in model performance with the training epoch, we measured the changes in Focus-DETR test indicators and compared them with DINO. Experimental results show that Focus-DETR outperforms DINO even at 12 epochs when using ResNet50 as the backbone, as shown in Table 4. In addition, we found that the Focus-DETR reached the optimal training state at 24 epochs due to special foreground selection and fine-grained feature enhancement.

B.3. Apply Dual Attention to Other Models

As we mentioned in Section 4.3 of the main text, a precise scoring mechanism is critical to the proposed dual attention. We add the experiments of applying the encoder with dual attention to those models equipped with Sparse

Model	Epochs	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params	GFLOPs	FPS
Deformable-DETR	Swin-T	50	48.0	68.0	52.0	30.3	51.4	63.7	41M	185	–
Sparse DETR	Swin-T	50	49.1	69.5	53.5	31.4	52.5	65.1	41M	129	18.9
Focus-DETR	Swin-T	36	51.9	69.8	56.7	34.8	55.2	67.1	49M	163	15.3
	Swin-B-224-22K	36	56.0	74.8	61.1	40.1	59.5	72.0	109M	368	15.3
	Swin-B-384-22K	36	55.9	74.7	60.9	39.6	59.5	73.0	109M	390	8.5

Table 3: Results for our Focus-DETR using Swin Transformer as the backbone. Herein, Swin-T indicates the tiny version pretrained on ImageNet-1K [1]. Swin-B-224-22K represents the base version pretrained on ImageNet-22K [1] and the resolution of training set is 224. All reported FPS are measured on a NVIDIA V100 GPU.

DETR, such as Deformable DETR [6], DN DETR [2] and DINO [5]. As shown in Table 5, the proposed dual attention for fine-grained tokens enhancement brings only +0.3AP in Deformable DETR(two-stage), 0.0AP in Deformable DETR(without two-stage), -0.1AP in DN-Deformable-DETR and +0.3 AP in DINO. Results show us that untrusted fine-grained tokens do not bring significant performance gains, which is still inefficient compared to Focus-DETR.

C. Visualization

As shown in Fig. 1, we visualize nine test images with diverse categories, complex backgrounds, overlapping targets, and different scales. We analyze the foreground features retained by different encoder layers. Visualization results show that foreground areas focus on a more refined area layer by layer in the encoder. Specifically, the result of Layer-6 captures a more accurate foreground with fewer tokens. The final test results of Focus-DETR are also presented, as shown in the first column.

In addition, we compare the differences of multi-scale feature maps retention object tokens due to our label assignment strategy. We also visualize Sparse DETR [4] to demonstrate the performance. As shown in first column of Fig. 2, Focus-DETR can obtain more precise foreground than Sparse DETR. According to the results of $\{f_1, f_2, f_3, f_4\}$, the multi-scale feature map of Focus-DETR can retain tokens according to different object scales, which further proves the advantages of our tag allocation and top-down score modulations strategy.

Model	Backbone	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DINO [5]	R-50	12	49.0	66.6	53.5	32.0	52.3	63.0
		24	50.4	68.3	54.8	33.3	53.7	64.8
		36	50.9	69.0	55.3	34.6	54.1	64.6
Focus-DETR	R-50	12	49.6	67.3	54.2	33.6	52.8	63.5
		24	50.3	68.4	55.1	33.9	53.5	64.4
		36	50.4	68.5	55.0	34.0	53.5	64.4
	R-101	12	51.1	69.4	55.5	33.5	54.8	65.3
		24	51.3	70.0	55.6	34.1	54.9	65.6
		36	51.4	70.0	55.6	34.2	55.0	65.5
	Swin-T	12	50.7	68.6	55.2	33.4	54.0	65.4
		24	51.8	69.7	56.6	34.6	55.2	66.9
		36	51.9	69.8	56.7	34.8	55.2	67.1

Table 4: Focus-DETR uses different backbones at different training epochs and provides comparison results with DINO [5]. R-50 and R-101 is ResNet backbone, Swin-T represents Swin Transformer of the tiny version.

Model	epoch	AP	GFLOPs	FPS
Deformable DETR (priori)	50	46.2	177	19
+ Sparse DETR ($\alpha = 0.3$)	50	46.0	121	23.2
+ Sparse DETR(dual attention) ($\alpha = 0.3$)	50	46.3	123	23.0
or + **Focus-DETR** ($\alpha = 0.3$)	50	46.6	123	23.0
Deformable DETR (learnable)	50	45.4	173	19
+ Sparse DETR ($\alpha = 0.3$)	50	43.5	118	24.2
+ Sparse DETR(dual attention) ($\alpha = 0.3$)	50	43.5	120	23.9
or + **Focus-DETR** ($\alpha = 0.3$)	50	45.2	120	23.9
DN-Deformable-DETR (learnable)	50	48.6	195	18.5
+ Sparse DETR ($\alpha = 0.3$)	50	47.4	137	23.9
+ Sparse DETR(dual attention) ($\alpha = 0.3$)	50	47.3	138	23.7
or + **Focus-DETR** ($\alpha = 0.3$)	50	48.5	138	23.6
DINO (mixed)	36	50.9	279	14.2
+ Sparse DETR ($\alpha = 0.3$)	36	48.2	152	20.2
+ Sparse DETR(dual attention) ($\alpha = 0.3$)	36	48.5	154	20.0
or + **Focus-DETR** ($\alpha = 0.3$)	36	50.4	154	20.0

Table 5: Apply dual attention to the classic models equipped with Sparse DETR and compare them with Focus-DETR.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[2] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Computer Vision and Pattern Recognition*, 2022. 1

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[4] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse DETR: Efficient end-to-end object detection with learnable sparsity. In *International Conference on Learning Representations*, 2022. 1, 2, 4

[5] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 1, 2

[6] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference*

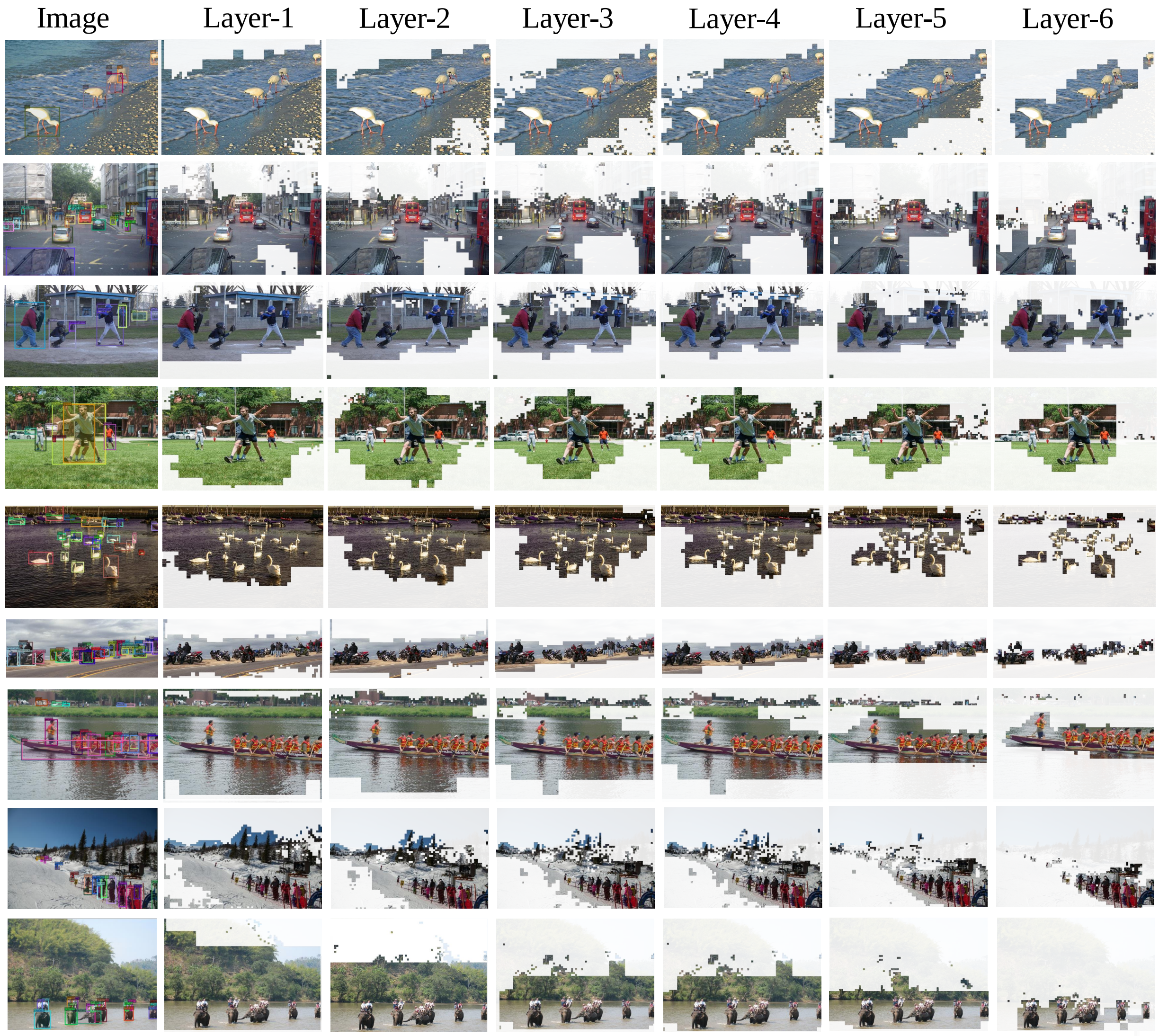

Figure 1: Visualization result of foreground tokens reserved at each encoder layer, and final detection results are provided. Layer-$\{1, 2, 3, ...\}$ indicates different encoder layers.

on Learning Representations, 2021. 1

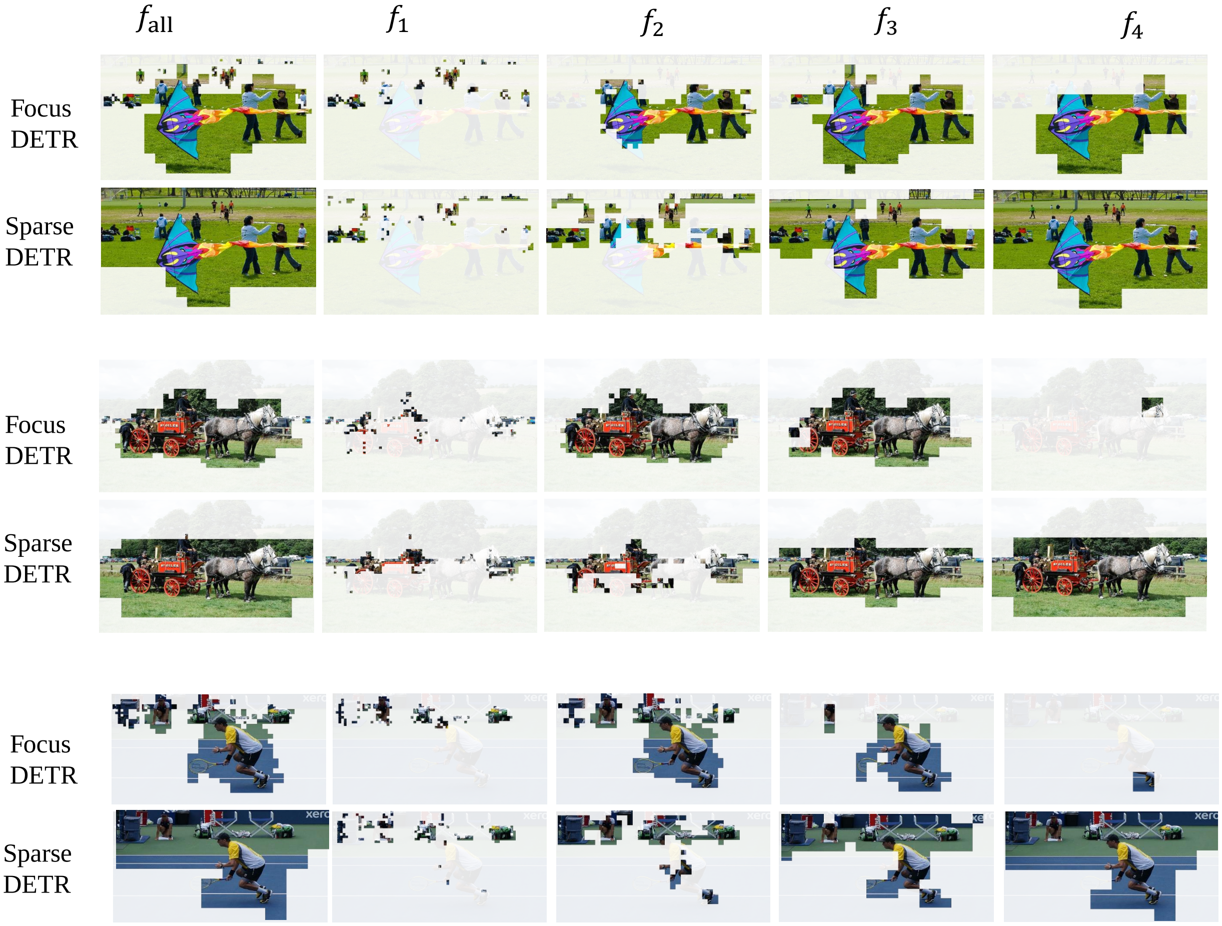

Figure 2: Visualized comparison result of foreground tokens reserved in different feature maps. We analyze the difference between Focus-DETR and Sparse DETR [4] by using three images with obvious object scale differences. f_{all} is the tokens retained by all feature maps, $\{f_1,\ f_2,\ f_3,\ f_4\}$ represents different feature maps.