Supplementary Materials for DETR Does Not Need Multi-Scale or Locality Design

Yutong Lin[1†] Yuhui Yuan[2†] Zheng Zhang[2†] Chen Li[1] Nanning Zheng[1] Han Hu[2†]

[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University

[2]Microsoft Research Asia

method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Plain DETR	46.5	70.2	50.0	26.3	50.2	65.7
Deformable DETR[9]	52.1	71.6	56.9	33.5	55.2	**69.0**
Ours	**53.8**	**73.4**	**58.9**	**35.9**	**57.0**	68.9

Table 1: **Comparison of the plain DETR baseline, Deformable DETR, and the improved plain DETR with a MIM pre-trained ViT-Base backbone.** Our plain DETR with global cross-attention improves the baseline by a huge margin and outperforms the Deformable DETR, which relies on multi-scale features and local cross attention.

A. More Plain ViT Results

Table 1 reports more comparison results based on the plain ViT. We use the default setup, described in Section 5.4 of the main text, to adopt a MAE [3] pre-trained ViT-Base as the backbone and train the model for $\sim$50 epochs. According to the results, we observe that (i) our method boosts the plain DETR baseline from 46.5 AP to 53.8 AP when only using a global cross-attention scheme to process single-scale feature maps; (ii) our approach outperforms the strong DETR-based object detector, e.g., Deformable DETR [9], which uses a local cross-attention scheme to exploit the benefits of multi-scale feature maps.

B. Runtime Comparison with Other Methods

We further analyze the runtime cost of different cross-attetnion modulations in Table 2. BoxRPB slightly increases runtime compared to standard cross-attention, while having comparable speed to other positional bias methods.

method	Training (min/epoch)	Inference (fps)
standard cross attn.	69	9.9
conditional cross att.	72	9.5
DAB cross attn.	73	9.3
SMCA cross attn.	79	9.6
Ours	75	9.5

Table 2: **Runtime comparison with local cross-attention scheme.** Global cross-attention with BoxRPB has comparable speed to other positional bias methods.

C. More Details of Local Attention Scheme

Figure 1 shows how our method differs from local cross-attention methods like deformable cross-attention [9], RoIAlign [4], RoI Sampling (fixed points in the Region of Interest), and box mask from [2]. Most local cross-attention methods need to construct a sparse key-value space with special sampling and interpolation mechanism. Our method uses all image positions as the key-value space and learns a box-to-pixel relative position bias term (gradient pink circular area in (e)) to adjust the attention weights. This makes our method more flexible and general than previous methods.

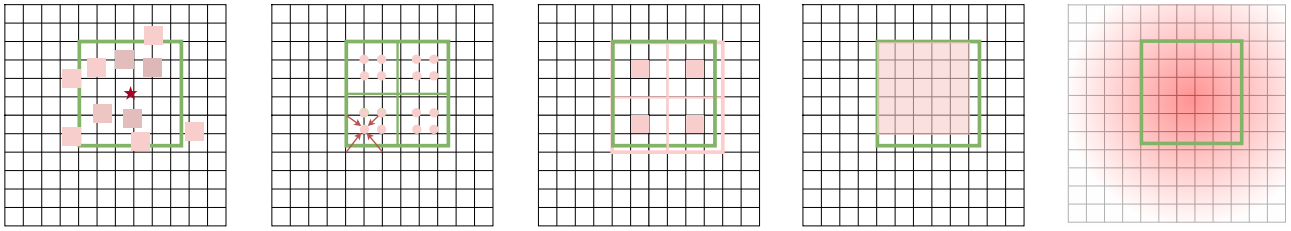
D. System-level Comparison on COCO val

Table 3 compares our method with previous state-of-the-art methods when using Swin-Large as the backbone. With 36 training epochs, our model achieves 59.8 AP on COCO val, outperforming DINO-DETR by +1.3 AP. With Objects365[7] pre-training, our method gets 63.8 AP, much higher than DINO-DETR. These results show that, with our approach, the improved plain DETR can achieve competitive performance without intrinsic limitations.

†Equal contribution. ✉ {yuhui.yuan, hanhu}@microsoft.com

method	framework	extra data	#params	#epoch	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Swin [6]	HTC	N/A	284M	72	57.1	75.6	62.5	42.4	60.7	71.1
Group-DETR [1]	DETR	N/A	≥218M	36	58.4	–	–	41.0	62.5	73.9
$\mathcal{H}$-Deformable-DETR [5]	DETR	N/A	218M	36	57.8	76.5	63.7	42.3	61.8	73.1
DINO-DETR [8]	DETR	N/A	218M	36	58.5	77.0	64.1	41.5	62.3	74.0
Ours*	DETR	N/A	228M	36	59.8	78.8	66.0	45.5	63.4	74.2
DINO-DETR [8]*	DETR	O365	218M	26 + 18	63.2	–	–	–	–	–
Ours*	DETR	O365	228M	24 + 24	**63.8**	**81.9**	**70.6**	**50.9**	**67.8**	**77.1**

Table 3: System-level comparisons with the state-of-the-art methods on COCO val. All methods adopt the Swin-Large backbone. The superscript $*$ marks the results with test time augmentation.



(a) Deformable cross-attn. (b) RoIAlign (c) RoI Sampling (d) Box Mask (e) Ours

Figure 1: Illustrating the comparisons between different local cross-attention mechanisms and our global cross-attention schema. We mark the sampling positions with pink color. The input image is represented by the black grid and the green box is the predicted bounding box from the previous layer. The red star marks the bounding box center. (a) Deformable cross-attention: It learns to sample important positions around the predicted bounding box for the key-value space. (b) RoIAlign: It uses bi-linear interpolation to compute the value of each sampling position in each RoI bin for the key-value space. (c) RoI Sampling: It quantizes the sampling positions to discrete bins and uses them as the key-value space. (d) Box mask: It selects all the positions within the green bounding box as the key-value space. (e) Our method: It improves global cross-attention with BoxRPB, which uses all the positions in the input image as the key-value space. The attention values are indicated by color intensity.

References

[1] Q. Chen, J. Wang, C. Han, S. Zhang, Z. Li, X. Chen, J. Chen, X. Wang, S. Han, G. Zhang, et al. Group detr v2: Strong object detector with encoder-decoder pretraining. *arXiv preprint arXiv:2211.03594*, 2022. 2

[2] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021. 1

[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 1

[5] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, and H. Hu. Detrs with hybrid matching. *arXiv preprint arXiv:2207.13080*, 2022. 2

[6] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2

[7] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 1

[8] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2

[9] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 1