Hybrid Proposal Refiner: Revisiting DETR Series from the Faster R-CNN Perspective Supplementary Materials

Method	Backbone	#Queries	#Epochs	AP
DiffusionDet	R50	500	-	46.8
Lite DETR	R50	900	36	49.5
Decoupled DETR	R50	300	50	47.0
Plain DETR	Swin-T	300	12	50.9
$\mathcal{C}$o-DETR	R50	900	12	52.1
Ours (w/ DDQ)	R50	300	12	**53.0**

Table 1. Comparison with the latest models.

Deformable Att.	Dynamic Conv.	Regional C.A.	AP
✓			47.8
✓	✓		48.9
✓		✓	48.4
✓	✓	✓	**49.3**

Table 2. Study on the integration of auxiliary proposal refiners (dynamic convolution and regional cross attention) into the primary proposal refiner (deformable attention). Refer to the supplementary materials for more results.

Primary	Auxiliary-1	Auxiliary-2	AP
Deformable Att.	-	-	47.8
Deformable Att.	Deformable Att.	Deformable Att.	48.5
Regional CA	Dynamic Conv.	Deformable Att.	48.9
Dynamic Conv.	Regional CA	Deformable Att.	48.8
Deformable Att.	Regional CA	Dynamic Conv.	**49.3**

Table 3. Ablation study on primary object refiners. Att.: attention. CA: cross attention. Conv.: convolution.

Loss Weight	AP	AP_l	AP_m	AP_s
1:1:1	49.1	63.8	51.7	32.5
2:1:1	**49.3**	62.8	52.4	32.6

Table 4. Ablation study on loss weight.

A. More Experiments

Unless otherwise specific, we utilize the enhanced Deformable DETR introduced by DINO [7], which includes a ResNet-50 backbone, 300 object queries, and a 12-epoch training schedule, serving as our base model. To this model, we apply HPR by incorporating two auxiliary refiners: one that utilizes regional cross attention and another that employs dynamic convolution. We conduct our experiments on the COCO benchmark.

Comparison with More Latest Models. We compare HPR with the latest models [2, 4, 5, 8, 10] in Table 1. HPR (w/ DDQ) achieves an AP of 53.0 surpassing all mentioned models.

Integration of Auxiliary Object Refiners into Primary Object Refiner. We adopt deformable attention as our primary proposal refiner. The performance improvements achieved by employing dynamic convolution, regional cross attention, and their combination as the auxiliary refiners are presented in Table 2. The inclusion of each auxiliary refiner enhances the effectiveness of using a solitary primary refiner.

Ablation Study on Primary Object Refiners. In our main paper, Figure 4 illustrates a scenario in which deformable attention is employed as the primary refiner, supported by dynamic convolution and regional cross attention as auxiliary refiners. In Table 3, we delve into alternative configurations, assigning the roles of primary refiners to both regional cross attention and dynamic convolution separately. We compare these setups against the original arrangement where deformable attention is the primary refiner. Additionally, we establish a baseline that utilizes deformable attention for the primary refiner, along with two auxiliary refiners. Our HPR amalgamates the strengths of diverse regional proposal refinement techniques, thereby surpassing the baseline that employs a singular type of proposal refinement strategy.

Ablation Study on Loss Weight. We perform an ablation study to examine how different loss weights between the primary and auxiliary refiners affect performance. Table 4 shows that our model achieves an AP of 49.3 under a loss weight distribution of 2:1:1.

Examination of Encoder and Decoder Number Variations. We explore the impact of varying the number of encoders and decoders on system performance. The number

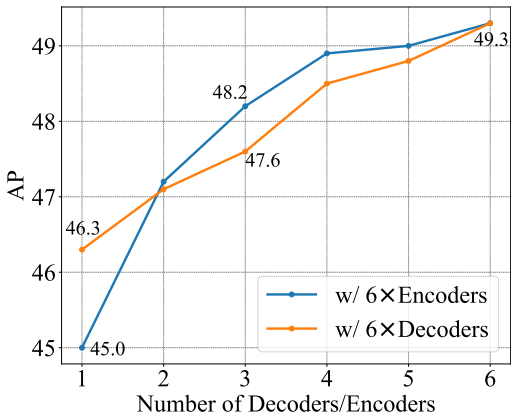

Figure 1. Ablation study on variations in the number of encoders (deformable encoders) and decoders (HPRs). Blue line: variation in the number of decoders within a model with 6× encoders. Orange line: variation in the number of encoders within a model with 6× decoders.

of decoders varies from 1 to 6 in a model with 6× encoders. Similarly, we apply this variation to a model with 6× decoders. The results are presented in Figure 1.

Operational Mechanisms of Various Proposal Refinement Strategies. In the realm of object feature utilization, the operational mechanisms of deformable attention, dynamic convolution, and regional cross attention exhibit distinct characteristics. Deformable attention predicts a sparse set of point features corresponding to each specific object feature. In contrast, dynamic convolution transforms object features into kernels—the generated kernels then slide over the RoI features to yield enhanced object features. Regional cross attention, meanwhile, operates by integrating object features with RoI features via a cross attention mechanism, wherein object features are treated as queries and RoI features as keys. In Figure 2, we visualize the activation maps for the three proposal refiners. It is evident that each refiner focuses on different areas and semantics of the object.

Additionally, we gather statistics on the cosine similarities of features extracted by two proposal refiners across object queries and throughout the images from the COCO `val` set. These statistics enable us to calculate an average cosine similarity s, which serves as a measure of the resemblance between the features extracted by the two refiners. A greater value of s suggests a higher degree of similarity in the features extracted by these refiners. We use the features derived from the first, the intermediate (third stage), and the last stages of HPR for similarity calculation. The visualizations are presented in Figure 3. It is evident that during the preliminary stages, specifically the first and third stages, the features encoded by various refiners exhibit considerable variance. In contrast, in the last stage, there is a notable increase in feature similarity, which is attributed to their convergence within a common latent space.

#Epochs	Data Re-aug.	LSJ	AP
12			49.3
	✓		50.3
		✓	49.3
	✓	✓	50.4
24			50.5
	✓		51.3
		✓	51.6
	✓	✓	52.8

Table 5. Ablation study on data re-augmentation and large-scale jitter (LSJ) augmentation.

Qualitative Study on Positive Sample Matching Strategies. In Figure 3 of the main paper, we visualize two activation maps generated by variants of Faster R-CNN using either Hungarian matching or IoU matching. We show more visualizations in Figure 4.

Training Curve Analysis. Figure 5 illustrates a comparative analysis of the training progression for the Align DETR [1] equipped with our HPR, alongside its original version and two other DETR variations, namely DINO [7] and Deformable DETR [9]. The incorporation of our HPR significantly accelerates the training convergence.

Ablation Study on Data Augmentations. We verify the effects of the proposed data re-augmentation and large-scale jitter augmentation in Table 5. It is observed that these two data augmentation strategies demonstrate compatibility in their application.

B. More Implementation Details

Data Augmentations. We summarize the normal (DETR-style), strong (used in our data re-augmentation), and large-scale jitter (LSJ) [3] data augmentations in Table 6. We apply the LSJ data augmentation to the image batch that has been processed with the proposed data re-augmentation.

Hyper-Parameters. All hyper-parameters used in our model are presented in Table 7.

C. Formulation of Proposal Refiners.

As described in Section 3.2 of the main paper, we use $\{\mathcal{P}_l\}$ to denote the feature maps encoded by the neck network (deformable encoder). Let $\boldsymbol{b}_i$ represent the i-th bounding box generated by the RPN. We use $\boldsymbol{p}_i$ and $\boldsymbol{r}_i$ to denote its object feature (point feature) and RoI feature, respectively. The enhanced object feature is represented by $\boldsymbol{p}'_i$. Below, we provide a formal formulation for each object refiner. For the sake of simplicity, we omit activation layers in our formulations.

R-CNN. It [6] employs a stack of convolutional layers to

Figure 2. Visualizations of the activation maps for deformable attention (the second row), dynamic convolution (the third row), and regional cross attention (the last row).

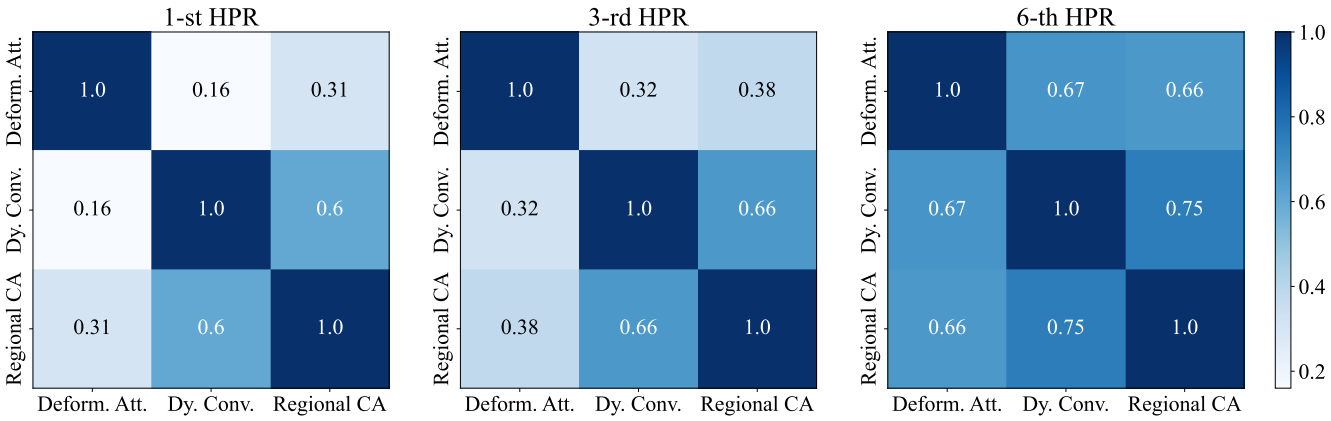

Figure 3. Visualizations for cosine similarities of various proposal refiners in distinct HPR stages.

refine the RoI features $\{\boldsymbol{r}_i\}$. This process can be formulated as:

$$\boldsymbol{p}_i' = \mathrm{FC}(\mathrm{Conv}(\boldsymbol{r}_i)).$$

Object Feature Refiner. This strategy directly processes the object features $\{\boldsymbol{p}_i\}$ using several FC layers:

$$\boldsymbol{p}_i' = \mathrm{FC}(\boldsymbol{p}_i).$$

Dynamic Convolution. This strategy facilitates interaction between $\boldsymbol{p}_i$ and $\boldsymbol{r}_i$. $\boldsymbol{p}_i$ is first used to generate convolution kernels through FC layers, and the convolution is subsequently applied to $\boldsymbol{r}_i$. The formulation is presented as

Figure 4. Visualizations of the activation maps generated by variants of Faster R-CNN using either IoU matching (the second row) or Hungarian matching (the third row).

Normal Augmentation	Random Flip, Random Resize, Random Crop	
Strong Augmentation	Geometric	Random Erasing, Rotate, Shear X, Shear Y, Translate X, Translate Y
	Appearance	Color Transform, Auto Contrast, Equalize, Sharpness, Posterize, Solarize, Color Balance, Contrast, Brightness
	Random Erasing	
LSJ Augmentation	Random Resize*, Random Crop*, Random Flip, Pad, Copy-Paste	

Table 6. Summary of various data augmentations applied in our model. $*$: the use of a larger augmentation factor.

follows:

$$\begin{aligned} \boldsymbol{K}_1 &= \mathrm{FC}(\boldsymbol{p}_i), \\ \boldsymbol{K}_2 &= \mathrm{FC}(\boldsymbol{p}_i), \\ \boldsymbol{p}'_i &= \mathrm{FC}(\mathrm{Conv}_{\boldsymbol{K}_2}(\mathrm{Conv}_{\boldsymbol{K}_1}(\boldsymbol{r}_i))), \end{aligned}$$

where $\mathrm{Conv}_{\boldsymbol{K}}$ denotes the convolution operator with kernel $\boldsymbol{K}$.

Regional Cross Attention. It applies cross attention between $\boldsymbol{p}_i$ and $\boldsymbol{r}_i$. $\boldsymbol{p}_i$ and $\boldsymbol{r}_i$ serve as queries and keys, respectively. We formulate the process as follows:

$$\begin{aligned} \hat{\boldsymbol{p}}_i^m &= \mathrm{FC}_m(\boldsymbol{p}_i),\ 1 \le m \le 5 \\ \{\boldsymbol{p}_i'^m\}_{m=1}^5 &= \mathrm{CrossAttention}(\{\hat{\boldsymbol{p}}_i^m\}_{m=1}^5, \boldsymbol{r}_i), \\ \boldsymbol{p}'_i &= \mathrm{Concatenation}(\{\boldsymbol{p}_i'^m\}_{m=1}^5). \end{aligned}$$

Deformable Attention. It uses several linear layers to predict a set of reference points with offsets Δ and the corresponding attention weights $\boldsymbol{A}$ for each $\boldsymbol{p}_i$. The entire process can be formulated as:

$$\begin{aligned} \Delta &= \mathrm{FC}(\boldsymbol{p}_i), \\ \boldsymbol{A} &= \mathrm{FC}(\boldsymbol{p}_i), \\ \boldsymbol{p}'_i &= \mathrm{DeformableAttention}(\{\mathcal{P}_l\}, \Delta, \boldsymbol{b}_i, \boldsymbol{A}), \end{aligned}$$

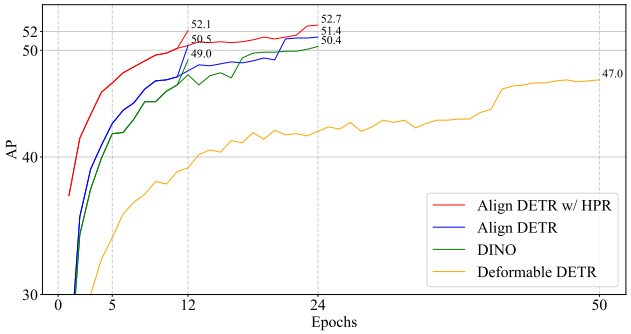

Figure 5. Training curves for AlignDETR equipped with our HPR, the original AlignDETR, DINO, and Deformable DETR.

Hyper-Patameter	Value
Backbone Features	(Res3, Res4, Res5)
Freeze Batchnorm	Truth
Neck Features	($\mathcal{P}_3$, $\mathcal{P}_4$, $\mathcal{P}_5$, $\mathcal{P}_6$)
Query Number	900
Loss Weight	2:1:1
Position Embedding Offset	-0.5
Position Embedding Temperature	10000
Encoder Number	6
Decoder Number	6
Embedding Dimension	256
Head Number	8
FFN Dimension	2048
HPR Integration Weight	Learnable
HPR Integration Type	Vector
HPR Integration Initialization	1:1:1
RoI Resolution	7×7
Dynamic Conv. Feature Dimension	64
Denoising Query Number	100
Classification Cost (Hungarian)	2.0
Bbox Cost (Hungarian)	5.0
GIoU Cost (Hungarian)	2.0
Classification Loss	Cross Entropy
Loss Weight (Classification)	1.0
Loss Weight (Bbox)	5.0
Loss Weight (GIoU)	2.0
gamma (Align DETR)	2.0
tau (Align DETR)	1.5
alpha (Align DETR)	0.25
Repeat GT Number (Align DETR)	2

Table 7. Summary of hyper-parameters.

where $\mathcal{P}_l$ denotes the l-th feature map generated by the deformable encoder and $\boldsymbol{b}_i$ represents the bounding box associated with $\boldsymbol{p}_i$.

Global Cross Attention. For this mechanism, each object feature $\boldsymbol{p}_i$ (query) interacts with $\mathcal{P}_5$ (keys) through a cross attention operation, which is formulated as:

$$\boldsymbol{p}_i' = \text{CrossAttention}(\boldsymbol{p}_i, \mathcal{P}_5).$$

Note that in the original DETR, the object features are randomly initialized, learnable object queries.

References

[1] Zhi Cai, Songtao Liu, Guodong Wang, Zheng Ge, Xiangyu Zhang, and Di Huang. Align-detr: Improving detr with simple iou-aware bce loss. *arXiv preprint arXiv:2304.07527*, 2023. 2

[2] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023. 1

[3] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 2

[4] Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, and Lionel M Ni. Lite detr: An interleaved multi-scale encoder for efficient detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18558–18567, 2023. 1

[5] Yutong Lin, Yuhui Yuan, Zheng Zhang, Chen Li, Nanning Zheng, and Han Hu. Detr does not need multi-scale or locality design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6545–6554, 2023. 1

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2

[7] 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. *arXiv preprint arXiv:2203.03605*, 2022. 1, 2

[8] Manyuan Zhang, Guanglu Song, Yu Liu, and Hongsheng Li. Decoupled detr: Spatially disentangling localization and classification for improved end-to-end object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6601–6610, 2023. 1

[9] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

[10] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6748–6758, 2023. 1