LP-DETR: Layer-wise Progressive Relations for Object Detection

Zhengjian Kang[1,*],Ye Zhang[2,*], Xiaoyu Deng[3], Xintao Li[4] and Yongzhe Zhang[5]
[1]New York University, NY 10012, USA
[2]University of Pittsburgh, PA 15213, USA
[3]Fordham University, NY 10458, USA
[4]Georgia Institute of Technology, GA 30332, USA
[5]California Institute of Technology, CA 91125, USA
[1]zk299@nyu.edu, [2]yez12@pitt.edu, [3]xdeng24@fordham.edu, [4]xli3204@gatech.edu, [5]yongzhe@caltech.edu

Abstract—**This paper presents LP-DETR (Layer-wise Progressive DETR), a novel approach that enhances DETR-based object detection through multi-scale relation modeling. Our method introduces learnable spatial relationships between object queries through a relation-aware self-attention mechanism, which adaptively learns to balance different scales of relations (local, medium and global) across decoder layers. This progressive design enables the model to effectively capture evolving spatial dependencies throughout the detection pipeline. Extensive experiments on COCO 2017 dataset demonstrate that our method improves both convergence speed and detection accuracy compared to standard self-attention module. The proposed method achieves competitive results, reaching 52.3% AP with 12 epochs and 52.5% AP with 24 epochs using ResNet-50 backbone, and further improving to 58.0% AP with Swin-L backbone. Furthermore, our analysis reveals an interesting pattern: the model naturally learns to prioritize local spatial relations in early decoder layers while gradually shifting attention to broader contexts in deeper layers, providing valuable insights for future research in object detection.**

Index Terms—**object detection, detection transformer, relation network, self-attention**

I. Introduction

DEtection Transformers (DETRs) [1] have achieved great progress by proposing an end-to-end architecture for object detection. However, their low training efficacy remains a critical challenge. The root cause is the imbalanced supervision during training - DETR employs Hungarian algorithm to assign only one positive prediction to each ground-truth box, leaving the majority of predictions as negative samples. This insufficient positive supervision leads to slow and unstable convergence. While various approaches have been proposed to address this issue through different technical routes like multi-scale feature learning [2], denoising training [3], [4], hybrid matching strategies [5], [6] and loss alignment [7], [8], they primarily focus on local feature enhancement or query learning optimization, leaving the potential of relation modeling in self-attention not been fully explored.

In the vision community, modeling inter-object relationships has proven beneficial for detection performance. Previous approaches mainly focus on two aspects: co-occurrence patterns of object categories [9]–[12] and spatial relations using various criteria [13]–[15]. These methods have demonstrated that incorporating relation information can effectively enhance detection accuracy by capturing contextual dependencies between objects. However, in DETR field, few works have investigated the learnable relation between object queries in the self-attention, a key component in DETR decoders. Hao et al. [12] attempt to model class correlations using a learnable relation matrix in the decoder's self-attention, but their approach does not consider spatial information and requires mapping class-to-class relations back to query-to-query interactions. More recently, Relation-DETR [16] introduces explicit position relations between bounding boxes with cross-layer refinement. Motivated by their work but different from these approaches, we directly incorporate geometric relation weights into queries within each layer and propose layer-specific relation modeling to capture evolving spatial dependencies.

In this paper, we present LP-DETR (Layer-wise Progressive DETR), which enhances object detection through explicit modeling of multi-scale spatial relations across decoder layers. Our key insight is that object relations naturally evolve from local to global contexts through the detection pipeline, and different scales of spatial relations may play varying roles at different stages of the detection process. Based on this observation, we propose a progressive relation-aware self-attention module that adaptively learns to balance different scales of spatial relations at different decoder layers. This design allows the model to capture fine-grained local relationships in early layers while gradually incorporating broader relation information in deeper layers. The main contributions of our work are threefold:

- We introduce a relation-aware self-attention mechanism that explicitly models multi-scale spatial relationships between object queries.
- We propose a progressive refinement strategy that allows the model to adaptively adjust relation weights across decoder layers.
- We discover and validate an interesting pattern where spatial relations naturally progress from local to global contexts through decoder layers, providing valuable insights for future research.

[*]Equal contribution to this work.

Finally, we conduct extensive experiments on COCO 2017 dataset to demonstrate the effectiveness of our approach. LP-DETR achieves competitive results with 52.3% AP under 12-epoch training and 52.5% AP under 24-epoch training using ResNet-50 backbone. With Swin-L backbone, our method further improves to 58.0% AP. More importantly, our analysis reveals that the proposed progressive relation modeling contributes to both improved convergence and detection accuracy. These results validate our hypothesis about the importance of layer-wise relation modeling and suggest promising directions for future research in object detection.

II. Related Work

A. Transformer for Object Detection

DEtection TRansformer (DETR) [1] establishes a new paradigm for end-to-end object detection by eliminating hand-crafted post-processing steps such as Non-maximum Suppression (NMS). Its transformer-based architecture consists of two main components: an encoder that transforms flattened image features into enriched memory representations, and a decoder that converts a set of learnable object queries into final detection results. The decoder operates through two attention mechanisms: self-attention for modeling interactions among object queries, and cross-attention for capturing relationships between queries and encoded memory features.

However, DETR suffers from slow convergence during training, and various approaches have been proposed to address this issue from different methodological perspectives: (1) Enhanced Feature Learning: Deformable DETR [2] explores multi-scale features through deformable attention with sparse reference points, while Focus-DETR [17] and Salience-DETR [18] improve feature selection through salient token identification in the encoder. (2) Query Enhancement: DAB-DETR [19] decouples object queries into 4D anchor box coordinates for iterative refinement, while DN-DETR [3] and DINO [4] accelerate training through auxiliary denoising task and contrastive learning. (3) Better Supervision: Hybrid DETR [5] and Group DETR [6] adopt one-to-many matching to increase supervision signals, while Stable-DINO [7] and Align-DETR [8] propose specialized loss functions to align classification and localization. (4) Attention Mechanism: Recent works focus on improving attention mechanisms, where Cascade-DETR [20] enhances query-feature interactions through cross-attention, and Relation-DETR [16] learns explicit relation modeling between queries in self-attention.

B. Relation Network

Relation networks have emerged as a powerful approach for modeling inter-object relationships at instance level, which can be broadly categorized into two main directions: co-occurrence modeling and spatial relation modeling.

Co-occurrence approaches focus on capturing statistical dependencies between object categories. Some methods [9], [10] directly learn from category distribution patterns in large datasets, while others [11], [12] adaptively learn class relationships from annotations. However, these approaches either rely on fixed statistical priors or encounter with challenges in mapping between instances and categories [10].

Spatial relation approaches construct graph structures where object features serve as nodes and their spatial relationships as edges. Pioneering works like Relation Network [13] introduces geometric weights in attention modules to model spatial relations. Recent methods determine relation weights through various metrics, such as position-aware distance [21], [22], attention mechanisms [14], [15] or appearance similarity [23]. While these learnable relations offer greater flexibility compared to fixed priors, they typically require larger datasets and longer training time to effectively learn the relations from data.

III. Methodology

A. DETR Preliminaries

A DETR-style detector consists of a backbone network (e.g., ResNet [24], Swin Transformer [25]) and a transformer architecture with encoder and decoder modules. Given an input image, the backbone first extracts image features, which are then split into patch tokens. The transformer encoder processes these tokens through self-attention mechanisms and outputs enhanced feature representations, denoted as memories $\mathbf{Z} = \{z_1, ..., z_m\}$.

The transformer decoder takes a set of learnable object queries $\mathbf{Q} = \{q_1, ..., q_n\}$ as input. Recent works [4], [19] propose to decouple these queries into content queries $\mathbf{Q^c}$ for label embedding and position queries $\mathbf{Q^p}$ for bounding box prediction, enabling better relation alignment. The decoder consists of L stacked blocks, where each block contains three sequential components: a self-attention layer, a cross-attention layer, and a feed-forward network (FFN).

The self-attention layer enables communication between object queries, allowing each query to refine its prediction by considering other queries' predictions. The cross-attention layer facilitates interaction between object queries $\mathbf{Q}$ and encoded memories $\mathbf{Z}$, aggregating features for object localization and classification. Finally, the FFN transforms the query embeddings for prediction through parallel classification and regression heads.

B. Layer-wise Progressive Relation-Aware Attention

The high-level architecture of our proposed progressive relation-aware DETR model is presented in Fig. 1. Our proposed attention is applied into the self-attention in the decoder component. Let's consider object queries $\mathbf{Q}$ consists of content embedding queries $\mathbf{Q^c}$, reference box position queries $\mathbf{Q^p}$ (represented by (x, y, w, h)), and relation queries $\mathbf{Q^r}$. Given a set of N object queries, $q_i = \{q_i^c, q_i^p, q_i^r\}_{i=1}^N$, the i-th relation query q_i^r with respect to all the object queries can be calculated as the weighted sum of all the queries:

$$q_i^r = \sum_{j=1}^{N} w_{ij}^r \cdot (W_V \cdot q_j^c). \tag{1}$$

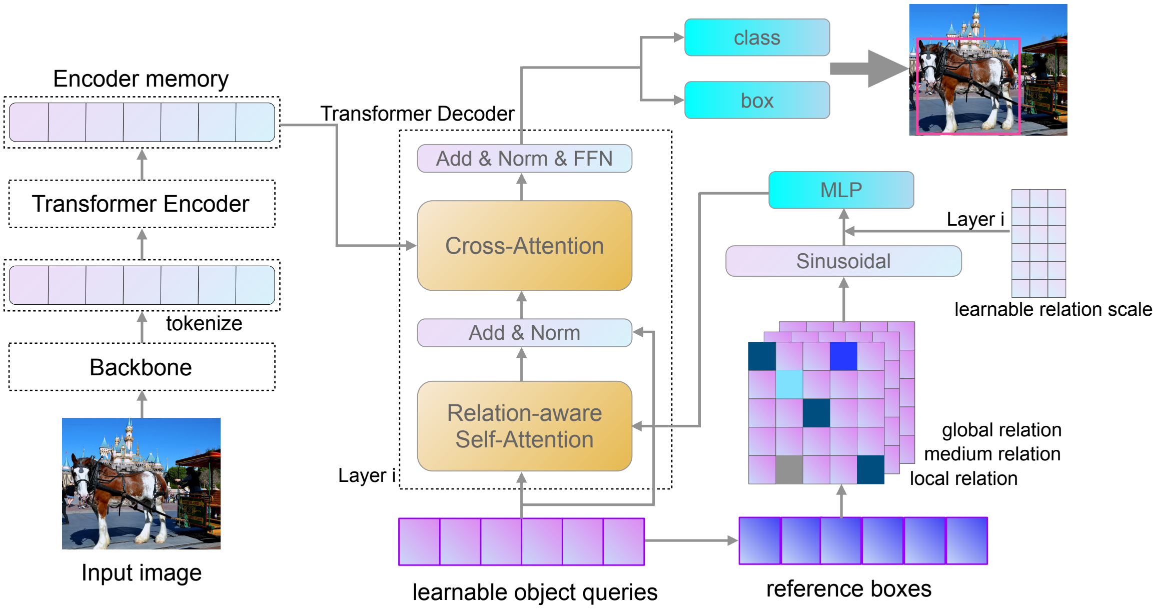

Fig. 1. DETR with layer-wise progressive relation pipeline. A learnable relation-aware self-attention mechanism that augments object queries with multi-scale spatial relations, which adaptively evolve from local to global contexts across decoder layers for progressive detection refinement.

The relation weight w_{ij}^r captures both geometric and content attention based relationships between queries, which is computed as:

$$w_{ij}^r = \frac{w_{ij}^p \cdot \exp\left(w_{ij}^c\right)}{\sum_{k=1}^N w_{ik}^p \cdot \exp\left(w_{ik}^c\right)}, \tag{2}$$

$$w_{ij}^c = \frac{(W_Q \cdot q_i^c) \cdot (W_K \cdot q_j^c)}{\sqrt{d_k}}, \tag{3}$$

where W_Q, W_K and W_V are learnable projection matrices for query, key and value in self-attention, d_k denotes the embedding size of $W_Q \cdot q_i^c$. The attention scores w_{ij}^c are normalized by geometric weights w_{ij}^p to obtain the final relation weights w_{ij}^r.

The geometric weight w_{ij}^p incorporates spatial relationships through:

$$w_{ij}^p = W_G \cdot E\Big(R(q_i^p, q_j^p)\Big), \tag{4}$$

$$\begin{aligned} R(q_i^p, q_j^p) = \Big(&\log(\frac{|x_i - x_j|}{w_i}), \log(\frac{|y_i - y_j|}{h_i}), \\ &\log(\frac{w_i}{w_j}), \log(\frac{h_i}{h_j}), \texttt{giou}(q_i^p, q_j^p)\Big), \end{aligned} \tag{5}$$

where the relation metric R captures spatial transformations in distances, scales and gIoU. The embedding function E maps these 5-D features to high-dimensional space using sinusoidal encoding [15], followed by a learnable projection W_G implemented as MLP with ReLU activation. Then the overall object query integrates information from multiple attention heads:

$$q_i^c = q_i^c + \texttt{MLP}\big(\texttt{Concat}(q_i^{r1}, ..., q_i^{rK})\big), \tag{6}$$

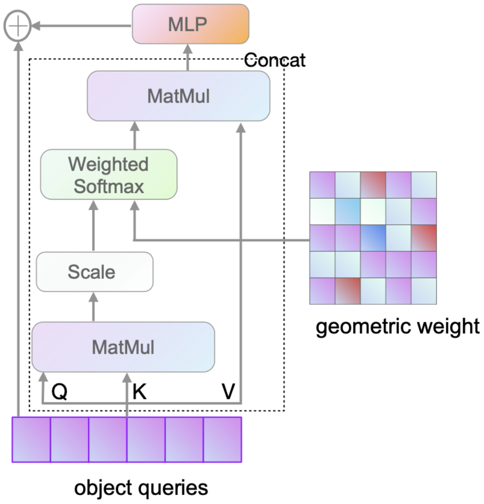

Fig. 2. Relation-aware self-attention module architecture. The module takes object queries and geometric weights as inputs and produces relationally-enhanced object queries through weighted self-attention mechanism.

where K denotes the number of relation-aware attention heads. The `Concat` operator aggregates the relation queries, and `MLP` enhances the object queries with learnable query-to-query relation weights, making them more sensitive to spatial relations during training. The details of relation-aware self-attention module is presented in Fig. 2.

To investigate how different scales of relations evolve across decoder layers, we propose three types of relation metrics:

local, medium and global relations. The local relation R_l uses the original metric in Eq. 5, emphasizing relative distances and scale variations. The medium relation R_m applies a scaling factor of $(1+2\times l/L)$, where l is l-th decoder layer, to reduce the steepness of the log function. The global relation R_g uses constant weights 1.0. Thus the geometric weight at the l-th decoder layer is formulated as:

$$w_{ij}^{p} = W_G \cdot \Lambda_l \cdot E\Big(R_l(q_i^p, q_j^p); R_m(q_i^p, q_j^p); R_g(q_i^p, q_j^p)\Big), \quad (7)$$

where $\Lambda = [\lambda_l; \lambda_m; \lambda_g]_{L\times 3}$ represents learnable weights that adaptively adjust the importance of different relation scales across decoder layers.

IV. Experiments

A. Experiment Settings

1) Dataset and backbone: We evaluate our Progressive Relation-Aware DETR on COCO 2017 [26], which contains 118k training images and 5k validation images across 80 object categories. The performance is evaluated on the validation set using standard COCO metrics: average precision (AP) at different IoU thresholds (IoU=0.5, 0.75, 0.5:0.95) and scales (small, medium, large). We implement our method with two backbone networks: ResNet50 [24] pretrained on ImageNet-1k and Swin-Large [25] pretrained on ImageNet-22k [27]. Both backbones are finetuned with an initial learning rate of 1×10^{-5}, which is decreased by a factor of 0.1 at later stages.

2) Implementation details: All experiments are conducted on NVIDIA RTX 3090 GPUs using AdamW optimizer [28] with a weight decay of 1×10^{-4} and a total batch size of 16. For the relation embedding module, we set the temperature $T = 10000$, scale $s = 100$, and position embedding dimension $d_{\text{pos}} = 16$ in the sinusoidal encoding. The position relations are constructed at three scales (local, medium and global) with equal initial weights 0.33 across all 6 decoder layers. Following standard practice in DETR-based methods, we apply common data augmentations including random resize, crop and flip during training. We report results under both 1x (12 epochs) and 2x (24 epochs) training schedules.

B. Experiment Results

We evaluate our model on COCO 2017 validation dataset using both ResNet-50 [24] and Swin-L [25] backbones. The results are summarized in Tables I and II. Using ResNet-50 backbone with 12-epoch training, our model achieves competitive results of 52.3% AP, 69.6% AP$_{50}$ and 56.8% AP$_{75}$. When compared to Relation-DETR [16], we observe consistent improvements across all metrics (+0.6% AP, +0.5% AP$_{50}$ and +0.5% AP$_{75}$). Analysis on objects of different scales shows that our method achieves 35.8% AP$_S$, 55.9% AP$_M$ and 66.6% AP$_L$, with notable gains on medium (+0.3% AP$_M$) and large objects (+0.5% AP$_L$) while maintaining competitive performance on small objects (-0.3% AP$_S$). These improvements become more evident with 24-epoch training, where our method further surpasses Relation-DETR by 0.4% AP, 0.3% AP$_{50}$ and 0.6% AP$_{75}$.

Furthermore, our approach demonstrates strong scalability to larger backbones. With Swin-L backbone under 12-epoch training, we achieve state-of-the-art performance of 58.0% AP, 76.4% AP$_{50}$ and 63.2% AP$_{75}$, improving upon the previous best results from Relation-DETR by 0.2% AP, 0.3% AP$_{50}$ and 0.3% AP$_{75}$.

C. Ablation study

Analysis of the number of relation heads. We examine how the number of relation heads affects model performance in our relation-aware self-attention module. Table III shows the results with varying numbers of relation heads while maintaining a total of 8 attention heads. Using no relation head (0 head) serves as our baseline, where the module functions as standard self-attention, achieving 51.1% AP, 68.6% AP$_{50}$ and 55.8% AP$_{75}$. The performance consistently improves as we increase the number of relation heads. With all 8 heads dedicated to relation-aware attention, our model achieves the best performance of 52.3% AP, showing an improvement of 1.2% AP over the baseline. Comparable gains are also observed in AP$_{50}$ (+1.0%) and AP$_{75}$ (+1.0%). These results demonstrate the benefit of incorporating relation-aware attention in the self-attention mechanism.

Analysis of different modules. We evaluate the effectiveness of progressive refinement (PR) in combination with our relation-aware self-attention module. In Table IV, the relation-aware self-attention module alone improves the detection performance by 0.9% AP, 0.6% AP$_{50}$ and 0.7% AP$_{75}$ compared to the baseline. Adding progressive refinement brings additional gains of 0.3% AP, 0.4% AP$_{50}$ and 0.3% AP$_{75}$. The improvements are also consistent across different object scales.

D. Progressive Refinement Analysis

To understand how different scales of position relations (local, medium and global) evolve through decoder layers, we visualize the learned relation weights across layers in Fig. 3. Our analysis reveals an interesting pattern in how the model balances different spatial relations. In the first layer, the weights among three scales remain relatively comparable, suggesting initial uncertainty in relation scale selection. From layer 2, the model develops a strong preference for local relations, with weights exceeding 0.9, indicating that early decoder layers focus primarily on establishing local relationships between queries. Moving to deeper layers, we observe a gradual transition in attention distribution: local relation weights decrease to 0.24, while medium and global relation weights steadily increase to around 0.4. By the final layer, the weights for medium and global relations surpass that of local relations. This progressive transition from local to broader spatial contexts aligns with the intuitive understanding that object detection requires hierarchical processing - from local to global query interactions. These findings suggest promising directions for future research in relation modeling, as the clear layer-wise progression from local to global relations indicates

TABLE I

EVALUATION ON COCO VAL2017 WITH STATE-OF-THE-ART METHODS USING RESNET-50 BACKBONE.

Method	Backbone	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Def-DETR [2]	ResNet-50	50	46.2	65.2	50.0	28.8	49.2	61.7
DAB-DETR [19]	ResNet-50	50	42.6	63.2	45.6	21.8	46.2	61.1
DN-Def-DETR [3]	ResNet-50	12	46.0	63.8	49.9	27.7	49.1	62.3
DINO [4]	ResNet-50	12	49.7	67.0	54.4	31.4	52.9	63.6
$\mathcal{H}$-Def-DETR [29]	ResNet-50	12	48.7	66.4	52.9	31.2	51.5	63.5
Cascade-DETR [20]	ResNet-50	12	49.7	67.1	54.1	32.4	53.5	65.1
$\mathcal{C}$o-Def-DETR [5]	ResNet-50	12	49.5	67.6	54.3	32.4	52.7	63.7
Align-DETR [8]	ResNet-50	12	50.2	67.8	54.4	32.9	53.3	65.0
Stable-DINO [7]	ResNet-50	12	50.4	67.4	55.0	32.9	54.0	65.5
DAC-DETR [30]	ResNet-50	12	50.0	67.6	54.7	32.9	53.1	64.4
Rank-DETR [31]	ResNet-50	12	50.4	67.9	55.2	33.6	53.8	64.2
MS-DETR [32]	ResNet-50	12	50.3	67.4	55.1	32.7	54.0	64.6
Relation-DETR [16]	ResNet-50	12	51.7	69.1	56.3	**36.1**	55.6	66.1
LP-DETR (ours)	ResNet-50	12	**52.3**	**69.6**	**56.8**	35.8	**55.9**	**66.6**
$\mathcal{H}$-Def-DETR [29]	ResNet-50	36	50.0	68.3	54.4	32.9	52.7	65.3
DINO [4]	ResNet-50	36	51.2	69.0	55.8	35.0	54.3	65.3
DINO [4]	ResNet-50	24	50.4	68.3	54.8	33.3	53.7	64.8
DDQ-DETR [33]	ResNet-50	24	52.0	69.5	**57.2**	35.2	54.9	65.9
Relation-DETR [16]	ResNet-50	24	52.1	69.7	56.6	36.1	56.0	66.5
LP-DETR (ours)	ResNet-50	24	**52.5**	**70.0**	**57.2**	**36.2**	**56.3**	**67.1**

TABLE II

EVALUATION ON COCO VAL2017 WITH STATE-OF-THE-ART METHODS USING SWIN-L BACKBONE.

Method	Backbone	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DINO [4]	Swin-L	12	56.8	75.4	62.3	41.1	60.6	73.5
$\mathcal{H}$-Def-DETR [29]	Swin-L	12	55.9	75.2	61.0	39.1	59.9	72.2
Rank-DETR [31]	Swin-L	12	57.3	75.9	62.9	40.8	61.3	73.2
Relation-DETR [16]	Swin-L	12	57.8	76.1	62.9	**41.2**	62.1	74.4
LP-DETR (ours)	Swin-L	12	**58.0**	**76.4**	**63.2**	41.0	**62.2**	**74.7**

TABLE III

PERFORMANCE COMPARISON OF DIFFERENT NUMBER OF RELATION HEADS IN RELATION-AWARE SELF-ATTENTION MODULE

# of heads	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
0	51.1	68.6	55.8	35.1	55.2	66.0
2	51.4	69.0	56.2	35.2	55.6	66.1
4	51.8	69.4	56.4	35.4	**56.2**	66.1
8	**52.3**	**69.6**	**56.8**	**35.8**	55.9	**66.6**

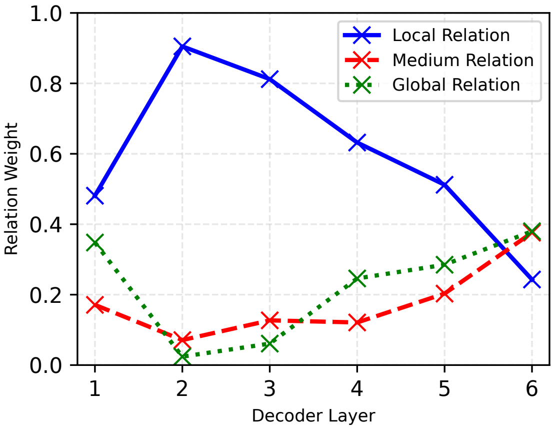

Fig. 3. Relation weight under local, medium and global relations in different decoder layers.

potential for more efficient architectures that explicitly leverage this hierarchical pattern.

E. Convergence comparison

Fig. 4 plots the convergence curves of different state-of-the-art methods with ResNet-50 backbone. Our model shows improved convergence behavior, benefiting from the learned layer-dependent multi-scale spatial relations between object queries. Although the absolute AP gain over Relation-DETR is moderate (+0.6% AP), our method consistently outperforms the baselines (DINO and Deformable-DETR) throughout the training process. This demonstrates that introducing layer-dependent multi-scale spatial relations can effectively refine the original relation modeling for object detection.

V. CONCLUSION

In this paper, we present a progressive relation-aware self-attention module that enhances DETR detector by incorporating learnable multi-scale spatial relationships between object queries. Our method adaptively adjusts relation weights across different scales and decoder layers, achieving competitive performance on standard benchmarks. Through extensive experiments, we demonstrate that our module improves both convergence speed and detection accuracy compared to standard self-attention. Our analysis reveals a pattern in how spatial relations evolve through the network: local relations dominate in early decoder layers, while global relations become increasingly important in deeper layers. Our findings open several promising directions for future research.

TABLE IV
PERFORMANCE COMPARISON ON DIFFERENT COMPONENTS USED IN THE DETECTOR.

Component	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
	51.1	68.6	55.8	35.1	55.2	66.0
Relation	52.0(**↑0.9**)	69.2(**↑0.6**)	56.5(**↑0.7**)	35.6(**↑0.5**)	55.5(**↑0.3**)	66.4(**↑0.4**)
Relation+PR	52.3(**↑0.3**)	69.6(**↑0.4**)	56.8(**↑0.3**)	35.8(**↑0.2**)	55.9(**↑0.4**)	66.6(**↑0.2**)

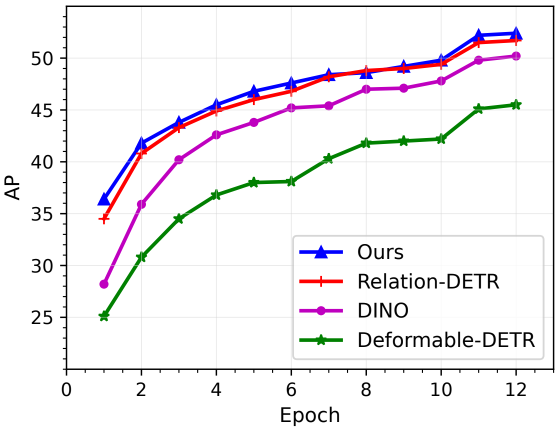

Fig. 4. Convergence comparison under different state-of-the-art methods with ResNet-50 backbone.

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision (ECCV)*. 2020, pp. 213–229, Springer.

[2] 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 on Learning Representations (ICLR)*, 2020.

[3] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13619–13627.

[4] 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," in *International Conference on Learning Representations (ICLR)*, 2022.

[5] Zykuang Zong, Guanglu Song, and Yu Liu, "DETRs with collaborative hybrid assignments training," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 6748–6758.

[6] Qiang Chen, Xiaokang Chen, Jian Wang, Shu Zhang, Kun Yao, Haocheng Feng, Junyu Han, Eric Ding, Gang Zeng, and Jingdong Wang, "Group DETR: Fast DETR training with group-wise one-to-many assignment," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 6633–6642.

[7] Shilong Liu, Tianhe Ren, Jinheng Chen, Zhengqiu Zeng, Hao Zhang, Feng Li, Hongyang Li, Jing Huang, Hang Su, Jun Zhu, et al., "Detection transformer with stable matching," *arXiv preprint arXiv:2304.04742*, 2023.

[8] Zhuokun Cai, Shilong Liu, Guanglu Wang, Zhirong Ge, Xinggang Zhang, and Dongjie Huang, "Align-DETR: Improving DETR with simple IoU-aware BCE loss," *arXiv preprint arXiv:2304.07527*, 2023.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, pp. 32–73, 2017.

[10] Hang Xu, Chenhan Jiang, Xiaodan Liang, Liang Lin, and Zhenguo Li, "Reasoning-RCNN: Unifying adaptive global reasoning into large-scale object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6419–6428.

[11] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta, "Iterative visual reasoning beyond convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7239–7248.

[12] Xin Hao, Dongjie Huang, Jimmy Lin, and Chin-Yew Lin, "Relation-enhanced DETR for component detection in graphic design reverse engineering," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2023, pp. 4785–4793.

[13] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei, "Relation networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3588–3597.

[14] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li, "Transformer-based dual relation graph for multi-label image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 163–172.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, Badong Chen, and Xuguang Lan, "Relation detr: Exploring explicit position relation prior for object detection," in *European conference on computer vision*. Springer, 2024.

[17] Dehua Zheng, Wenhui Dong, Hailin Hu, Xinghao Chen, and Yunhe Wang, "Less is more: Focus attention for efficient detr," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 6651–6660.

[18] Xin Hou, Ming Liu, Shiming Zhang, Pengchao Wei, and Biao Chen, "Salience detr: Enhancing detection transformer with hierarchical salience filtering refinement," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17574–17583.

[19] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xiaogang Qi, Hang Su, Jun Zhu, and Lei Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations (ICLR)*, 2021.

[20] Ming Ye, Lei Ke, Siyuan Li, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, and Fisher Yu, "Cascade-DETR: Delving into high-quality universal object detection," in *International Conference on Computer Vision (ICCV)*, 2023, pp. 6704–6714.

[21] Hengyue Bi, Canhui Xu, Cao Shi, Guozhu Liu, Yuteng Li, Honghong Zhang, and Jing Qu, "Srrv: A novel document object detector based on spatial-related relation and vision," *IEEE Transactions on Multimedia*, vol. 25, pp. 3788–3798, 2023.

[22] Jin Lin, Yanxiang Pan, Rong Lai, Xiaoqing Yang, Hongyang Chao, and Tao Yao, "Core-Text: Improving scene text detection with contrastive relational reasoning," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[23] Zhichao Li, Xiangyu Du, and Yang Cao, "Gar: Graph assisted reasoning for object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 1295–1304.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[25] 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, pp. 10012–10022.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[28] Diederik P. Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” 2017.

[29] Di Jia, Yuhui Yuan, Haodi He, Xiaokang Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu, “DETRs with hybrid matching,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19702–19712.

[30] Zhenghao Hu, Yueming Sun, Jiacheng Wang, and Yueming Yang, “DAC-DETR: Divide the attention layers and conquer,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[31] Yanjie Pu, Weicong Liang, Yanwei Hao, Yuhui Yuan, Yi Yang, Chao Zhang, Han Hu, and Gao Huang, “Rank-DETR for high quality object detection,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[32] Chuyang Zhao, Yueming Sun, Wenhan Wang, Qiang Chen, Eric Ding, Yueming Yang, and Jingdong Wang, “Ms-detr: Efficient detr training with mixed supervision,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17027–17036.

[33] Shihua Zhang, Xinggang Wang, Jian Wang, Jiangmiao Pang, Chen Lyu, Wenyu Zhang, Ping Luo, and Kai Chen, “Dense distinct query for end-to-end object detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7329–7338.