ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10887994

Precision in Pathology: PMA-DETR Elevates Tumor Lesion Detection

Yecheng Zhao
School of Computer Science and Engineering
Southeast University
NanJing, China
zhaoyecheng@seu.edu.cn

ZiHan Zhou
School of Computer Science and Engineering
Southeast University
NanJing, China
zhzhou@seu.edu.cn

Lei Qi
School of Computer Science and Engineering
Southeast University
NanJing, China
qilei@seu.edu.cn

Hui Xue
School of Computer Science and Engineering
Southeast University
NanJing, China
hxue@seu.edu.cn

***Abstract*—The DETR series, known for its end-to-end object detection models, has gained significant attention for its performance. RT-DETR excels with higher accuracy and faster real-time inference. However, applying these models to medical imaging poses challenges, such as low-contrast and complex lesion structures, which can reduce effectiveness. When detecting tumors, models may overfit due to the distinct differences and variability between different cases, affecting generalization and accuracy. To address these challenges, we propose a multi-view parallel feature extraction module, specifically for tumor detection. This module includes adaptive preprocessing, joint axial and channel attention, multi-pooling angular attention to enhance relevant features and reduce redundancy. Additionally, axial dynamic deformable convolution is used to improve adaptability and robustness. The resulting PMA-DETR architecture achieves state-of-the-art tumor detection while maintaining real-time processing.**

***Index Terms*—medical images, tumor detection, attention improvements**

I. Introduction

Tumor detection in ultrasound is complex due to many missed and misleading tumor-like lesions [1]. Tumors are classified as benign or malignant. Benign tumors cause minimal impact unless large, while malignant tumors are highly harmful. Deep learning-based object detection is pivotal in this process [2]. Effective detection models can help to mitigate this task in an efficient and automated manner.

Existing object detection models, trained on high-quality datasets like COCO [3], handle well-separated targets. However, tumors data faces challenges such as low resolution, indistinct boundaries, and significant morphological variations. There is a lack of suitable models for it. Although RT-DETR [4] shows superior performance in general detection tasks, defeats famous target detection methods such as YOLO series [5], its adaptation to medical images particularly for tumors still requires further optimization despite its advantages [6].

To address the challenges, we design a new **PMA-DETR(Parallel Multi-view scale convolution and multi-level contextual Attention Tumor DETR)**, we implement the following enhancements:

- We introduce a **Parallel Multi-view Scale Convolution and Dilated Convolution Module (PMCD)** for tumor preprocessing. It combines attention mechanisms and multi-residual links to effectively capture semantic features across different shapes of tumors.
- We develop a novel **Integrated Multi-level Contextual Attention Aggregation Module (IMCA)**. These innovations are designed to suppress redundant information and enhance the representation of relevant features. This ensures that subsequent features are more precise in characterization.
- Additionally, considering the irregular and complex nature of tumors, we design a new **Multi-axis Deformable Convolution Module (MDC)**. This enhancement incorporates offset learning, allowing convolution to better adapt to local deformations and capture the unique characteristics of tumors.

These refinements enhance our object detector's capability to effectively detect tumors in medical images by improving feature extraction, attention mechanisms, and convolution operations tailored to the specific challenges.

II. METHOD

Distinguishing from existing natural image detectors, we design a joint **Parallel Multi-view Scale Convolution and Dilated Convolution Module** to address the challenges of tumor detection in medical images. This helps the model focus on effective lesion areas early. **Itegrated Multi-level Attention Aggregation Module** enhances the representation of extracted features, while the **Multi-axis Deformable Convolution Module** adapts the detector to the variability of tumor

features. Replacing the standard convolution in DETR with MDC significantly improves the model's robustness.

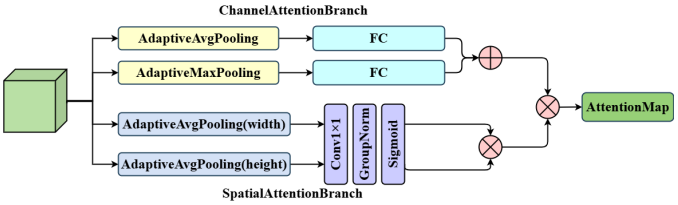

Fig. 1. The Axis Adaptive-Channel Aggregation Attention Block processes the input feature map using maximum and average pooling to extract important information. It facilitates complementarity between spatial and channel attention.

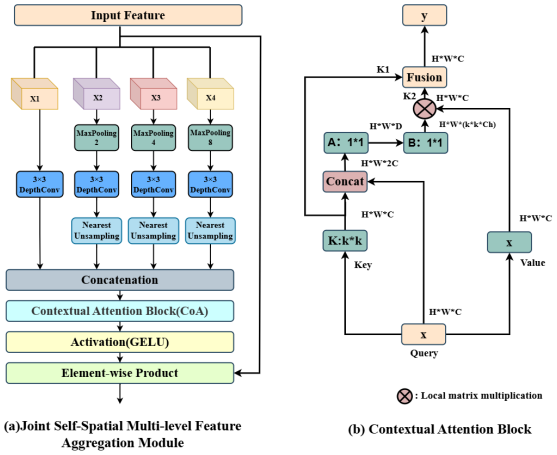

Fig. 2. The Joint Self-spatial Multi-level Feature Aggregation Block uses nearest concatenation upsampling and concatenation on different features. Contextual Attention utilizes contextual information between input keys to guide the learning of dynamic attention matrices.

A. Parallel Multi-view Scale Convolution and Dilated Convolution Module

In order to adapt to multi-scale target detection in medical images, before feature extraction backbone, four parallel convolutions in PMA-DETR are used to extract multi-scale features. How to strike a balance between receptive field and computational effort? Using dilated convolutions with different dilation rates in parallel. Using multiple sizes, it extracts features across various scales addressing variability in tumor shapes. Integrating dilated convolutions with different dilation rates expands the receptive field, allowing the model to capture irregularity. The structure composed of this module and subsequent integrated attention is shown in Figure 3.

B. Integrated Multi-level Attention Module

Two sub-attention modules Axis Adaptative-Channel Aggregation Attention Block (ACA Block) and Joint Self-spatial Multi-level Aggregation Block (JSMF Block), are designed to be cascaded in PMA-DETR. Realizing new spatial-channel attention, and effective complementarity of picture information contexts. By performing multi-view attention computation,

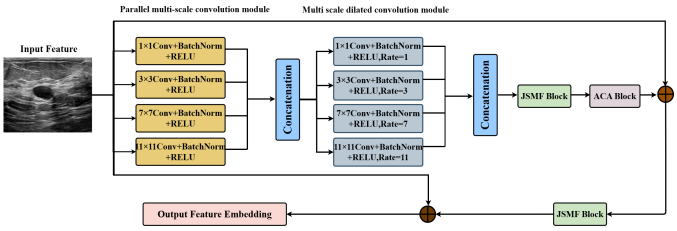

Fig. 3. Parallel Multi-view Scale Convolution and Dilated Convolution Module and Integrated Multi-level Attention Aggregation Module. Inherits multi-layer hybrid convolution, JSMF attention and ACA attention module.

features in the effective region are enhanced and redundant part are suppressed.

1) Axis Adaptative-Channel Aggregation Attention Block: Ultrasound images often contain noise and artifacts, which can be mitigated through channel attention mechanisms improving model robustness. As shown in Figure 1. We propose a Axis Adaptative-Channel Aggregation Attention to enhance the network's ability to process features across both channel and spatial dimensions. Additionally, adaptive average pooling is performed on the horizontal and vertical axes. The final output is element product of horizontal and vertical attention maps, expressed as the following formula:

$$\mathbf{SA_h} = Sigmoid\left(GroupNorm\left(Conv1d\left(H_{pool}\right)\right)\right), \quad (1)$$

$$\mathbf{SA_w} = Sigmoid\left(GroupNorm\left(Conv1d\left(W_{pool}\right)\right)\right), \quad (2)$$

$$\mathbf{SA_{out}} = \mathbf{SA_h} \times \mathbf{SA_w}. \quad (3)$$

In the design of channel attention, adaptive pooling are performed on the channels respectively to extract channel statistical information, and channel attention map is obtained after redistribution. The formula is as follows:

$$\mathbf{CA_{out}} = Sigmoid(Conv2d(AvgPool(X))) + Sigmoid(Conv2d(MaxPool(X))), \quad (4)$$

input X is element-wise multiplied with channel attention map and spatial to obtain final map. The final output is:

$$\mathbf{Out} = \mathbf{CA_{out}} \times \mathbf{X} \times \mathbf{SA_{out}}. \quad (5)$$

2) Joint Self-spatial Multi-level Aggregation Block: Global information aids models in understanding overall content and long-term dependencies, but global multi-head attention can introduce high computational overhead. Lesion areas are typically small and calculating attention across entire image can introduce noise, diluting key local details.

As shown in Figure 2,the Joint Self-spatial Multi-level Feature Aggregation Block aims to achieve global interaction similar to attention mechanisms without exce ssive cost. Long-distance interactions are enabled through multi-scale feature fusion. The input feature X is split along the channel dimension, and different spatial down-sampling and up-sampling processes are applied. The above process is expressed as the following formula:

$$\mathbf{S_i} = AdaptiveMaxPool\left(X_c\left[i\right], \left(\frac{H}{2^i}, \frac{W}{2^i}\right)\right), \quad (6)$$

Smaller feature maps have larger receptive fields, so convolution with the same kernel size produces different receptive fields across scales. We introduce global residual connection to learn high-frequency details and fast reconstruction. Additionally, the Contextual Attention Block (CoA) utilizes contextual information between input keys to guide the learning of dynamic attention matrices. Specifically, the CoA block encodes input keys via convolutions generating a static contextual representation. CoA performs context encoding on all adjacent keys within the grid and concatenate the context $K1$ and Q. The expressions are as follows:

$$\mathbf{F} \in \mathbb{R}^{H\times W\times 2C} = Concat\left(\mathbf{K}1, \mathbf{Q}\right), \tag{7}$$

$$\mathbf{M} \in \mathbb{R}^{H\times W\times (k\times k\times Ch)} = \mathbf{F}\mathbf{W}_A\mathbf{W}_B, \tag{8}$$

$K1$ reflects the static context information of local adjacent positions, $\mathbf{W}_A \in \mathbb{R}^{2C\times D}, \mathbf{W}_B \in \mathbb{R}^{D\times(k\times k\times Ch)}$ represents the parameters of two consecutive pointwise convolutions.

C. Multi-axis Deformable Convolution Module

Due to the small proportion of tumor lesions, the model inevitably loses its perception of the corresponding structure, and the convolutional nucleus is completely free from the target. Therefore, we hope to design specific feature extraction based on the characteristics of the lesion structure, so as to lead the model to focus on the key core features. Benign and malignant tumors, as well as different stages of the same tumor, present diverse morphologies. This variability causes models to overfit to previous features and struggle with novel shapes, weakening generalization and accuracy.

Inspired by Dynamic Snake Convolution (DSConv) [7], we design a new Multi-axis Deformable Convolution to replace traditional convolutions. This adaptation aims to capture the central structure of tumors more flexibly while maintaining alignment with target structure under constraints.

The Multi-axis Deformable Convolution Module begins with offset learning via a standard 2D convolution layer for each kernel position. Subsequently, input feature map undergoes coordinate mapping, where new feature coordinates are generated based on learned offsets. This process relies on kernel size,deformation range and so on. It can be described:

$$offset = tanh(BatchNorm\left(Conv2d\left(X\right)\right)), \tag{9}$$

$$x_{new} = x_{center} + x_{grid} + offset \cdot \Delta S, \tag{10}$$

$$y_{new} = y_{center} + y_{grid} + offset \cdot \Delta S, \tag{11}$$

x_{new} and y_{new} represent coordinates after convolution kernel position, ΔS means extend scope. This operation adjusts kernel's position on the feature map to respond to local deformations, improving the model's adaptability to complex scale changes in lesions. In addition to x and y axes, kernel positions are adjusted in 45-degree and 135-degree directions. By deforming, model can better adapt to morphological characteristics of tumor targets.

III. EXPERIMENTS

A. Dataset

Breast Ultrasound Image (BUSI). We chose the BUSI dataset which data collected includes breast ultrasound images of women aged between 25 and 75 years old. This data was collected in 2018. The number of patients is 600 female patients. This dataset consists of 780 images, with an average image size of 500 * 500 pixels. These images are in PNG format. The tumor image (mask) is presented together with the original image. These images are classified into three categories, namely normal (tumor free), benign benign, and malignant. We will convert the corresponding lesion areas of this dataset into txt format labels.

Brain Tumor v2.0. We also select the Brain Tumor Detection dataset publicly available on Roboflow by Yousef Ghanem. This dataset collected 9900 image data from brain CT imaging, with three corresponding labels identifying the tumor area of the lesion. We divided the dataset images into 7920 training sets, 990 validation sets, and 990 testing sets.These two datasets are both tumor detection task data in medical image scenes, which can to some extent prove the effectiveness and robustness of the method proposed in this paper.

B. Experimental Process and Results

We use the RT-DETR-r18 version as baseline network with the AdamW optimizer. The initial learning rate is set to 0.0001, with a learning rate decay factor of 1.0, momentum of 0.9, and weight decay of 0.0001. The images size are fixed at 640x640, and the batch size is 8. The model is trained on an NVIDIA Tesla V100 32GB GPU.

We test the effectiveness of the proposed module on the BUSI and Brain Tumor v2.0 datasets before integrating it into the feature extraction backbone and replacing the axial deformable convolution module. Ablation studies and comparisons with other object detection algorithms are conducted.

Under same experimental parameters and data augmentations, proposed methods demonstrating effectiveness. IOU measures the overlap between the ground truth and predicted boxes, while mAP (mean Average Precision) averages precision across multiple categories. The performance of PMA-DETR on BUSI dataset is shown in Table I, performance on the Brain Tumor v2.0 dataset is shown in Table II.

TABLE I
COMPARATIVE EXPERIMENTS ON BUSI DATASET USED ONE-STAGE AND DETR SERIES DETECTORS. PMA-DETR OUTPERFORM OTHERS WITH THE SAME BACKBONE SCALE, EFFECTIVELY CAPTURES THE CORE FEATURES OF TUMOR LESIONS. YOLO SERIES USE THE DEFAULT NORMAL SCALE, WHILE OTHERS INCLUDING DETR, USE THE SMALLEST SCALE BACKBONE. M AND B MEAN MALIGNANT TUMOR CATEGORY AND BENIGN. PMA-DETR PERFORMS WELL ON BOTH TYPES OF TUMORS.

Model	mAP50(M)	mAP50(B)	mAP50(all)	mAP50-95
Yolov5 [8]	69.4%	87.0%	78.2%	53.9%
Yolov7 [9]	72.0%	85.3%	77.1%	52.4%
Yolov8 [10]	72.6%	87.1%	78.6%	53.8%
SMCADETR [11]	69.9%	84.0%	74.3%	50.9%
DETR-DC5 [12]	69.7%	86.2%	75.5%	51.4%
Df-DETR [13]	67.3%	83.2%	76.0%	52.8%
DINO [14]	70.0%	86.7%	78.3%	53.2%
Our model	**74.7%**	**88.7%**	**80.2%**	**60.2%**

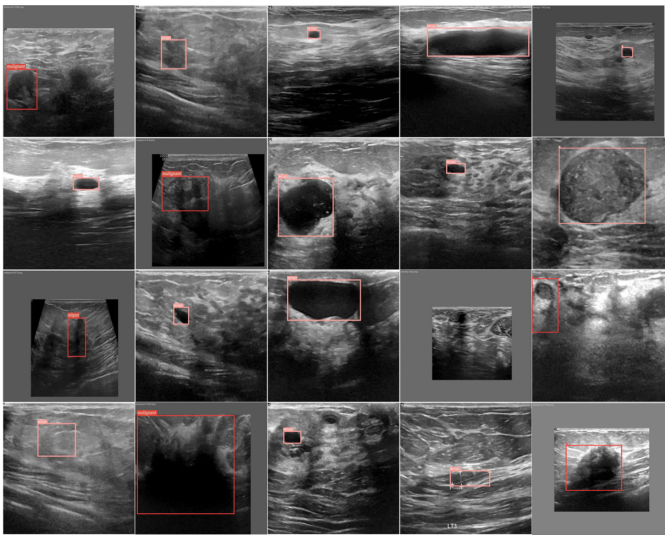

Fig. 4. Predicted results on the BUSI dataset. The pink and red boxes represent the detection results of benign and malignant tumors, respectively.

TABLE II
BRAINTUMORV2.0 DATASET COMPARISON EXPERIMENT. ONE STAGE AND DETR SERIES TARGET DETECTORS ARE SET UP. SET THE SAME AS BUSI DATASET. PMA-DETR ACHIEVES THE BEST PERFORMANCE UNDER THE SAME SCALE BACKBONE, ACHIEVES SIGNIFICANT LEADERSHIP IN BOTH MAP50 AND MAP50-90 INDICATORS.

Model	mAP50	mAP50-95
Yolov5 [8]	79.8%	53.7%
Yolov7 [9]	78.8%	52.3%
Yolov8 [10]	79.3%	52.5%
SMCADETR [11]	77.2%	50.7%
DETR-DC5 [12]	77.0%	51.3%
Df-DETR [13]	80.0%	52.7%
DINO [14]	78.6%	51.9%
Our model	**83.9%**	**62.9%**

C. Ablation Experiment

In order to verify the effect of different components, we conduct a detailed ablation experiment. We enable different modules on the basic network structure for testing. Specifically, we test the following situations:

- Enable the Parallel Multi-view Scale Convolution and Dilated Convolution Module(PMCD) separately;
- Enable the Integrated Multi-level Contextual Attention Aggregation Module(IMCA) separately;
- Enable the Multi-axis Deformable Convolution Module(MDC) separately;
- Enable PMCD, IMCA and MDC modules simultaneously in PMA-DETR.

As shown in Table III, when PMCD, IMCA and MDC modules are enabled at the same time, the experimental results are optimal. Specifically, the mAP50 of malignant tumors is increased from 68.4% to 74.7%, the mAP50 of benign is increased from 86.2% to 88.6%, the overall mAP50 is increased to 80.2%, and the mAP50-95 is also increased from 57.3% to 60.2%. These results show that new module effectively solves the challenges when processing medical data and significantly improves detection performance on complex tumors.

TABLE III
ABLATION EXPERIMENT WITH OUR MODULE ON BACKBONE. COMPARED TO THE ORIGINAL DETR NETWORK STRUCTURE, OUR NEW MODULES SIGNIFICANTLY IMPROVE MAP INDICATORS OF BOTH MALIGNANT AND BENIGN TUMORS.

PMCD	IMCA	MDC	mAP50(M)	mAP50(B)	mAP50(all)	mAP50-95
			68.4%	86.2%	77.4%	57.3%
✓			**75.0%**	83.8%	79.1%	58.0%
	✓		72.2%	83.1%	78.3%	57.8%
		✓	73.9%	85.0%	79.4%	58.4%
✓	✓	✓	74.7%	**88.7%**	**80.2%**	**60.2%**

D. Attention Experiment

In order to verify the effectiveness of the hybrid attention module proposed in this paper and other types of attention modules, we conduct detailed comparative experiments covering a variety of common attention mechanisms, including methods based on channel attention and spatial attention. As shown in Table IV, the hybrid attention module proposed in this paper combines spatial multi-scale and channel attention mechanisms and achieves the best results in experiments, with 80.2% mAP50 and 60.0% mAP50-90.

TABLE IV
COMPARATIVE EXPERIMENTS VERIFYING THE HYBRID ATTENTION MODULE PROPOSED IN THIS PAPER WITH OTHER TYPES OF ATTENTION. THE EXPERIMENTS DISTINGUISH BETWEEN COMMON OBJECT DETECTION METHODS BASED ON CHANNEL ATTENTION AND SPATIAL ATTENTION. OUR ATTENTION ACHIEVES THE BEST RESULTS BY COMBINING SPATIAL MULTI-SCALE AND CHANNEL ATTENTION MECHANISMS.

Attention Block	mAP50	mAP50-90	From	Dimension
Triplet Attention [15]	78.1%	56.2%	WACV2021	Channel
P2TAttention [16]	78.0%	56.3%	TPAMI2022	Spatial
MuLinearAttention [17]	79.4%	58.0%	ICCV2023	Spatial
Sgformer [18]	78.6%	57.1%	ICCV2023	Spatial
CoTAttention [19]	78.4%	57.7%	TPAMI2022	Spatial
CGAFusion [20]	78.9%	56.7%	TIP2024	Mix
Our model	**80.2%**	**60.2%**	/	Mix

IV. CONCLUSION

Our paper focuses on adaptively enhancing the Transformer-based object detection model RT-DETR for medical image tumor data. We introduce a Parallel Multi-view Scale Convolution and Dilated Convolution Module to adapt to low-resolution and unclear lesion boundaries by dynamically adjusting receptive fields. Additionally, an Integrated Multi-level Attention Aggregation Module is designed to enhance local and global feature representation. We also optimize the convolution in the original DETR with a Multi-axis Deformable Convolution Module to improve tumor feature capture. Our method outperforms other object detection techniques on medical image datasets. Incorporating the three proposed modules improves the model's mAP50 and mAP50-90 metrics by 2.8% and 2.9%, respectively. The resulting PMA-DETR architecture enables state-of-the-art oncology testing and will be a powerful tool for tumor lesion detection.

REFERENCES

[1] Thierry AGM Huisman, "Tumor-like lesions of the brain," *Cancer Imaging*, vol. 9, no. Special issue A, pp. S10, 2009.

[2] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[3] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[4] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2024, pp. 16965–16974.

[5] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, 2016, pp. 779–788.

[6] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser, "Dilated residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, July 2017.

[7] Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, and Guanyu Yang, "Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision(ICCV)*, 2023, pp. 6070–6079.

[8] Glenn Jocher, "Yolov5 release v7.0," https://github.com/ultralytics/yolov5/tree/v7.0, 2022.

[9] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[10] Glenn Jocher, "Yolov8," https://github.com/ultralytics/ultralytics/tree/main, 2023.

[11] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li, "Fast convergence of detr with spatially modulated co-attention," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3601–3610, 2021.

[12] Radhika Baskar, Rajat Kumar Dwibedi, Makhan Kumbhkar, Swati Sah, Harshal Patil, and Firos A, "Detr-dc5 approaches for improved spatial object detection in satellite imagery," in *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, 2023, pp. 1–5.

[13] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *ArXiv*, vol. abs/2010.04159, 2020.

[14] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun-Juan Zhu, Lionel Ming shuan Ni, and Heung yeung Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *ArXiv*, vol. abs/2203.03605, 2022.

[15] Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou, "Rotate to attend: Convolutional triplet attention module," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 3138–3147.

[16] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng, "P2t: Pyramid pooling transformer for scene understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 12760–12771, 2023.

[17] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han, "Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 17256–17267.

[18] Qitian Wu, Wentao Zhao, Chenxiao Yang, Hengrui Zhang, Fan Nie, Haitian Jiang, Yatao Bian, and Junchi Yan, "Sgformer: Simplifying and empowering transformers for large-graph representations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[19] Yehao Li, Ting Yao, Yingwei Pan, and Tao Mei, "Contextual transformer networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1489–1500, 2023.

[20] Zixuan Chen, Zewei He, and Zhe-Ming Lu, "Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention," *IEEE Transactions on Image Processing*, vol. 33, pp. 1002–1015, 2024.