HPPE: HeatMap Positional Embedding for Background Noise Suppression in Object Detection

Yuxin Liu[1], Hongyun Liu[2], Yusen Wu[3], Yangchen Zeng[4]✉

[1]Shenyang University of Technology, [2]Shenyang University of Chemical Technology, [3]Fujian University of Technology, [4]Southeast University
liuyuxin@smail.sut.edu.cn, 3315109664@qq.com,
3231319130@smail.fjut.edu.cn, ydmpzb@seu.edu.cn

Abstract. Despite the rapid proliferation of Transformer-based architectures in small object detection, mitigating background noise and improving query quality remain critical bottlenecks. To address these challenges, this study presents HeatMap Positional Embedding (HPPE), an adaptive optimization framework that intrinsically couples positional encoding with semantic detection priors through heatmap-guided masking. Furthermore, we develop a novel visualization scheme for HPPE, providing an intuitive perspective on feature embeddings to facilitate hyperparameter tuning. Building upon this mechanism, two specialized modules are proposed: the Multi-Scale Object Box-Heatmap Fusion Encoder (MOFE) and the HeatMap Induced High-Quality Queries for Decoder (HIQQ). These components are tailored to generate semantically rich queries while actively suppressing irrelevant background interference. By incorporating heatmap positional embeddings alongside standard baseline feature extractors like Linear-Snake Conv (LSConv), our approach effectively handles the massive diversity of small object categories and drastically minimizes the required number of decoder multi-head layers. Extensive evaluations demonstrate that our framework achieves absolute mAP improvements of 2.3% on the small object benchmark (NWPU VHR-10) and 1.8% on the general dataset (PASCAL VOC) compared to the baseline.

Keywords: Object Detection, Vision Transformer, Heatmap.

1 Introduction

Small object detection continues to be a cornerstone task in computer vision, underpinning critical applications across diverse domains such as drone-based aerial imaging (UAVs), remote sensing (RSI), infrared detection (IF), and high-resolution wide (HRW) imagery[1–3]. Despite its broad utility, this task is fraught with inherent difficulties, including the dense clustering of targets, the minuscule footprint of objects, and overwhelmingly high background-to-object ratios.

In recent years, Transformer-based architectures have gained substantial traction over traditional CNN detectors due to their superior capacity for global context

modeling. However, state-of-the-art DETR-like models still grapple with notable limitations when applied to small object scenarios:

1. Suboptimal Positional Utilization: Existing methods often generate queries without fully exploiting the intrinsic correlations between positional encodings and object semantics, leading to degraded performance, especially under limited training data.
2. Query Quality Degradation: The inability to selectively suppress low-scoring, background-contaminated queries compromises both bounding box localization and classification accuracy.
3. Decoder Inefficiency: The multi-head attention mechanism inherently scales with model depth. The lack of refined, information-dense embeddings forces the reliance on redundant decoder layers, severely slowing down inference speeds.

To tackle these persistent bottlenecks, we propose HeatMap Positional Embedding (HPPE), an adaptive optimization paradigm that dynamically couples positional encoding with semantic detection priors via heatmap-guided learning. By seamlessly integrating spatial, categorical, and geometric data through gradient-weighted class activation maps, HPPE dramatically accelerates network convergence and elevates inference quality. Additionally, we introduce a novel visualization scheme for HPPE, offering intuitive insights into feature embeddings for precise parameter tuning.

Building upon the HPPE mechanism, we design two specialized modules to fortify the encoder-decoder architecture:

The Multi-Scale ObjectBox-Heatmap Fusion Encoder (MOFE) integrates class and bounding-box semantics into multi-scale positional embeddings, effectively expanding receptive fields while suppressing background noise.

Conversely, the HeatMap Induced High-Quality Queries for Decoder (HIQQ) distills these mixed heatmap features into semantically rich queries, mitigating decoder redundancy. To further enhance the extraction of sparse features in intricate small targets, we optionally deploy standard baseline techniques like Linear-Snake Conv (LSConv).[4–7]

In summary, our core contributions are: (1) Formulating the HPPE framework along with an interpretable visualization method to dynamically align positional and semantic features. (2) Proposing the MOFE and HIQQ modules to generate high-quality, noise-resistant queries for the encoder and decoder. (3) Demonstrating that HPPE significantly trims the required decoder layers, achieving faster inference without sacrificing accuracy.

2 Proposed Method

2.1 HPPE:HeatMap Positional Embedding And Visualization

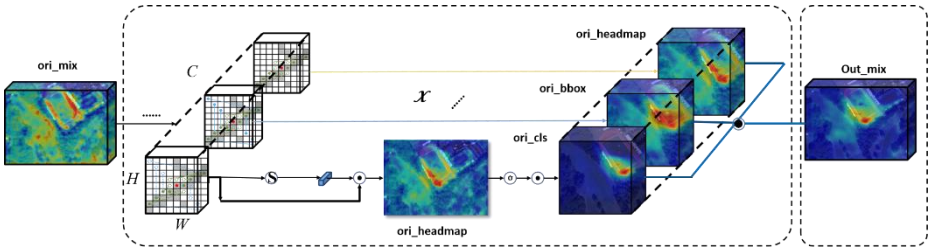

Fig. 1. The pipeline of the proposed method HeatMap Positional Embedding. It mainly consists of the HIQQ and MOFE modules. The HPPE module appliesnorm, followed by upsampling, and employs a Mask Filter to produce high-quality queries while filtering out low-quality background noise.

While DETR-based object detection methods exhibit remarkable advantages in global context modeling, they fundamentally lack the exploration of intrinsic correlations between positional embedding and detection semantics.The absence of embedded information coupling—manifested as static positional embeddings and constrained synergy between semantic features—fundamentally limits further improvements in detection performance. As illustrated in the method pipeline figure 1, this paper proposes the HPPE (HeatMap Positional Embedding algorithm and implements HeatMap Positional Embedding Visualization, which generates high-quality positional embedding integrated with detection semantics. This advancement significantly enhances query quality in encoder-decoder structure and improves small targets detection accuracy.

Conventional approaches rely on uniform 2D Sinusoidal Positional Encoding schemes a technique that critically neglects multi-scale object variations and fails to decouple spatial dependencies between semantic and positional features. Such limitations cascade into three deficiencies: loss of high-frequency positional information crucial for small targets, mutual interference between category classification and bounding box regression tasks, and inadequate high-dimensional semantic embeddings to distinguish densely distributed small objects in cluttered environments. These shortcomings collectively undermine detection robustness, particularly in scenarios requiring precise localization of miniature or overlapping instances.

To address this limitation, we propose HeatMap Positional Embedding (HPPE), a novel mechanism that generates high semantic density heatmap matrices through gradient-weighted class activation mapping, dynamically aligning positional embedding with semantic-critical regions in multi-scale detection tasks.
As illustrated in the pipeline framework Figure 1.

Algorithm 1. HeatMap Positional Embedding (HPPE)

Input:Image $\boldsymbol{I} \in \boldsymbol{R^{3\times H_0\times W_0}}$**,Ratio** $\boldsymbol{\varepsilon} \in [\boldsymbol{0}, \boldsymbol{1}]$**,Positional Encoding DimenSion** $\boldsymbol{d_{pe}}$

Output: $\boldsymbol{D} = \boldsymbol{b}_k, \boldsymbol{c}_k{}_{k=1}^{K}$

1. $\boldsymbol{H}_{\text{cls}} \leftarrow$ **Grad-CAM**$(\boldsymbol{I})$

2. $\boldsymbol{H}_{\text{box}} \leftarrow$ **Reg-Grad**$(\boldsymbol{I})$

3. $\boldsymbol{H}_{\text{mix}} \leftarrow \boldsymbol{\epsilon} \cdot \boldsymbol{H}_{\text{cls}} + (\mathbf{1} - \boldsymbol{\epsilon}) \cdot \boldsymbol{H}_{\text{box}}$

4. **Upsample** $\boldsymbol{H}_{\text{mix}}$ **to original dimensions**

5. $\mathbf{M} \leftarrow \mathbf{1}(\boldsymbol{H}_{\text{mix}} > \boldsymbol{\tau})$

6. $\mathbf{PE} \leftarrow \text{SinusoidalPE}(\boldsymbol{d}_{\text{pe}})$

7. $\mathbf{PE}' \leftarrow \boldsymbol{M} \odot \boldsymbol{PE}$

8. **For** $\boldsymbol{l} = \mathbf{1}$ **to** $\boldsymbol{L}_{\text{enc}}$

9. $\mathbf{F}_1 \leftarrow \text{MOFE}(\boldsymbol{F}_{l-1}, \boldsymbol{PE}')$

10. **end for**

11. $\boldsymbol{Q}_0 \leftarrow \text{Linear}(\boldsymbol{H}_{\text{mix}})$

12. **for** $\boldsymbol{l} = \mathbf{1}$ **to** $\boldsymbol{L}_{\text{dec}}$ **do**

13. $\boldsymbol{Q}_l \leftarrow \text{HIQQ}(\boldsymbol{Q}_{l-1}, \boldsymbol{F}_{\text{enc}})$

14. **end for**

15. $\mathbf{D} \leftarrow \text{DetectionHead}(\boldsymbol{Q}_{L_{\text{dec}}})$

16. **return D**

- *Step 1* employs a convolutional neural network to extract input image feature tensors $\mathbf{A} \in \mathbb{R}^{\mathbb{K}\times\mathbb{H}\times\mathbb{W}}$, where $A_k(i,j)$denotes the activation intensity of the k-th filter at spatial coordinate (i, j).

For a detection class c, the gradient weighting coefficients α_{ijk}^{c} are derived by computing second- and third-order partial derivatives of the classification head's output confidence score y^c with respect to the feature tensor **A**:

$$\alpha_{ijk}^{c} = \frac{\partial^2 y^c}{\partial A_k(i,j)^2} + \frac{\partial^3 y^c}{\partial A_k(i,j)^3}$$

This formulation quantifies the nonlinear sensitivity of category confidence to spatial activation variations. The global channel-wise importance weight β_k^c for channel k is computed via spatial aggregation incorporating ReLU-based gradient filtering:

$$\beta_k^c = \sum_{i=1}^{H}\sum_{j=1}^{W} \alpha_{ijk}^{c} \cdot \text{ReLU}\left(\frac{\partial y^c}{\partial A_k(i,j)}\right)$$

Here, ReLU suppresses negative gradients to retain only activations positively correlated with class c, while α_{ijk}^{c} hierarchically amplifies regions where activation magnitudes exhibit higher-order nonlinear contributions to class discrimination.

- *Step 2* Generate H_{class} based on the channel-wise global importance weights β_k^c:

$$H_{\text{class}}(i,j) = \text{ReLU}\left(\sum_{k=1}^{K} \beta_k^c \cdot A_k(i,j)\right)$$

Analogous to H_{class}, replace the confidence score $y^c$with the bounding box regression lossL_{reg}, and obtain activation weights through Huber loss gradient backpropagation to generate the detection-box heatmap H_{bbox}.Generate the hybrid heatmap H_{mixed} by dynamically adjusting the contribution ratio of categorical and geometric information via the gating mechanism λ:

$$H_{\text{mixed}} = \lambda \cdot H_{\text{class}} + (1-\lambda) \cdot H_{\text{bbox}}$$

Obtain three upsampled masks with the same dimensions as the standard feature maps.

- *Step 3* Construct a dynamic masking mechanism MASK Filter to decouple cold and hot regions in the heatmap, enabling heatmap-guided position embedding, as shown in Equation~\eqref{eqa:mask}. The mask matrix multiplies element-wise with standard positional encoding to suppress invalid positional noise in background regions while preserving multi-scale geometric features in heat regions, thereby enhancing semantic integration of class, position, and bounding box information in encoder-decoder architectures.

As shown in Figure 2, the heatmap positional embedding visualization demonstrates background suppression through the gradient threshold of the mask filter $Mask$, which creates anisotropic attention patterns in hot/cold regions. Introducing ReLU gradient filtering during mask filter generation establishes a semantic screening mechanism: For Heat Area (Good Embeddings ($H_{\text{map}} > \tau$), retain full sinusoidal positional encoding, executing precise geometric positional matching, visually depicted as heatmap focal points. For Cold Area (Poor Embeddings ($H_{\text{map}} \leq$), binarize positional embeddings via gradient heat-value filtering (ReLU suppresses invalid background noise), degrading to global semantics, visually depicted as background. Evaluation metrics demonstrate that the ReLU mask filter endows tokens with two characteristics: Position branch enhances semantically prominent global features, and Detection branch sharpens geometry-sensitive local variations.

This section introduces HeatMap Positional Embedding (HPPE), an innovative framework that dynamically aligns position embedding with object detection semantics through heatmap-guided adaptive learning. By guiding position embedding optimization through heatmap, it effectively enhances the collaborative performance of classification and localization in object detection.

2.2 HPPE to High-Quality-Query

The HPPE method processes raw feature maps through multi-branch operations to generate three specialized heatmaps: category-predominant heatmap (H_{class}), bounding box-guided heatmap (H_{bbox}), and hybrid-guided heatmap (H_{mixed}). By leveraging the

threshold property of the Mask filter, it filters Poor embeddings response regions, constructs a binarized mask matrix, and integrates high-quality queries generated from positional encoding. We propose the MOFE and HIQQ modules to enhance the correlation between positional encoding and detection semantics in the encoder and decoder, respectively.

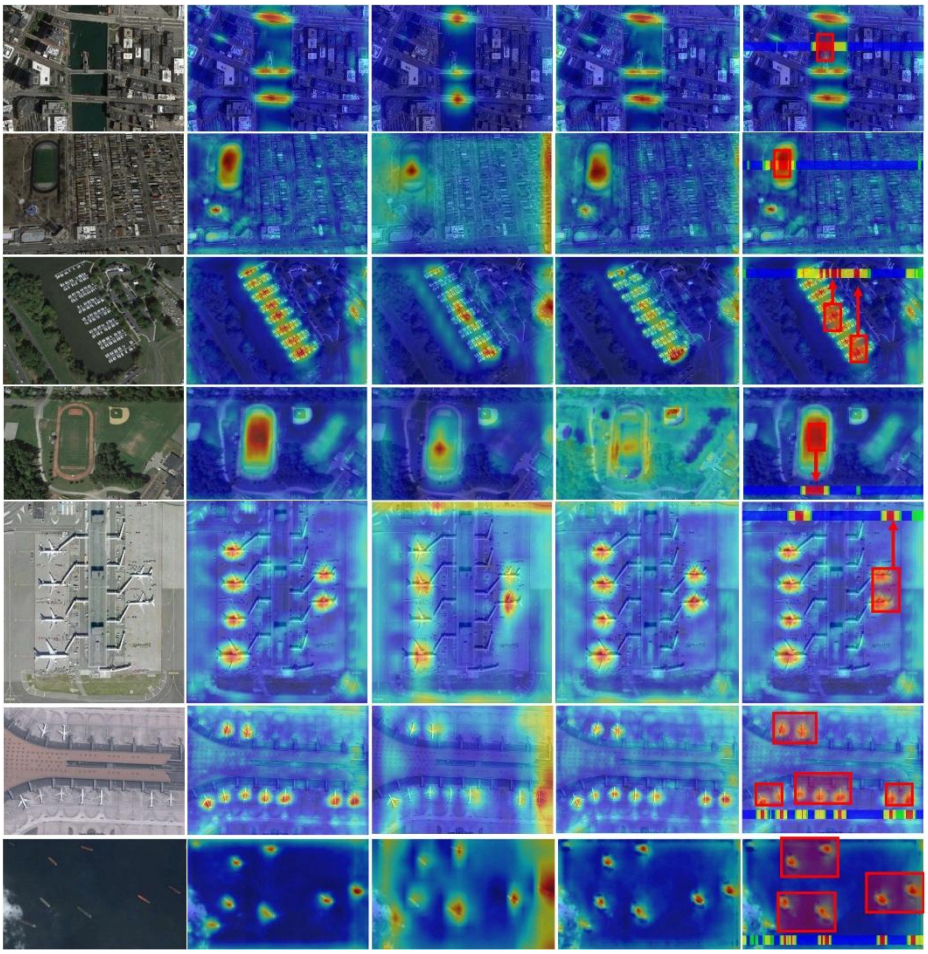

Fig. 2. Heatmap positional embedding visualized with heatbar

The heatmap embedding shows red hot middle and blue cold ends, corresponding to red good and blue poor embeddings. The first figure shows the original input image. The second figure presents the combined heatmap $\mathbf{H}_{mix}$ at the final layer with$\epsilon = 0.1$, indicating a weighted mixture of $0.9\mathbf{H}_{\text{bbox}}$ and$0.1\mathbf{H}_{class}$. The third figure displays the class activation heatmap $\mathbf{H}_{class}$ at the final layer, while the fourth figure shows the bounding-box heatmap $\mathbf{H}_{\text{bbox}}$ from the mixed layer. The last figure provides the visualized HPPE embedding with a corresponding color heatbar for reference.

MOFE:Multi-Scale ObjectBox-Heatmap Fusion Encoder

MOFE (Encoder-oriented): Its core innovation lies in establishing a conditional coupling mechanism between heatmap spectra and positional encoding, forging strong correlations between positional encoding and semantic features while enabling multi-scale feature fusion to expand receptive fields.

At the encoder stage, the MOFE module applies independent linear projections to the category heatmap H_{class} and bounding box heatmap H_{bbox}, concatenates the processed features $E_{\text{class}} \| E_{\text{bbox}}$, and generates Query/Key/Value matrices for multi-head attention: $Q_{\text{enc}} = W_Q[E_{\text{class}} \| E_{\text{bbox}}]$, $K_{\text{enc}} = W_K[E_{\text{class}} \| E_{\text{bbox}}]$,$V_{\text{enc}} = W_V[E_{\text{class}} \| E_{\text{bbox}}]$. This process facilitates implicit high-level decoupling between semantic features and position features.

The mask matrix $Mask(i,j)$ zeros out positional embeddings for poor embeddings,

$$Mask(i,j) = \begin{cases} 0, (i,j) \in \{ H_{i,j}^{grand} \} \leq \tau \} \\ 1, \qquad\qquad\qquad else \end{cases}$$

via:

$$\text{PE}(i,j,d) = \text{Mask}(i,j) \odot \left[\sin\left(\frac{i}{\tau_d}\right) + \cos\left(\frac{j}{\tau_d}\right) \right]$$

as shown in the above equation, dynamically suppressing redundant noise in background regions. Here $Mask \in \{0,1\}^{H \times W}$ is a binary spatial filter generated by gradient thresholding, $\tau_d = 10000^{2d/D}$,$(d \in \{0,1,\dots,D/2 - 1\})$, and $\odot$ denotes Hadamard product for element-wise masking.

Meanwhile, heat regions (good embeddings) preserve multi-scale sinusoidal encoding characteristics, constructing target-sensitive context modeling.

2.3 HIQQ: HeatMap Induced High-Quality Queries for Decoder

HIQQ (Decoder-oriented): The HIQQ module transforms embedding features of the mixed heatmap (H_{mixed}) into high-quality query vectors. The mixed heatmap undergoes linear transformation to generate initial queries$Q_{\text{dec}} = W_Q' E_{\text{mixed}}$, followed by a deformable attention mechanism $DeformAttn(Q_{\text{dec}}, K_{\text{enc}}, V_{\text{enc}})$.

HPPE-generated high-quality queries (serving as the exclusive QUERY) suppress background-contaminated low-quality queries through attention reweighting, thereby simultaneously improving the decoder's precision in bounding box regression and robustness in semantic category prediction.

This work proposes that HPPE-generated high-quality queries significantly enhance object detection performance.

In encoding, the MOFE module innovatively integrates class and bbox heatmaps, employs a gradient-based Mask filter to build robust positional encodings, and suppresses background noise while expanding receptive fields through feature disentanglement andmulti-scale embeddings.

In decoding, the devised HIQQ converts mixed heatmap features into high-quality initial queries, leveraging deformable attention mechanisms to consolidate semantic-

locational correlations, thereby substantially improving the detection head's regression accuracy for geometric attributes and semantic features.

3 Experimental results and analysis

Table 1. Comparison of different methods on VOC07+12 And NWPU VHR-10 datasets. The optimal results are highlighted in red, and the suboptimal results are in blue.

	Method	BackBone	Datasets				GFLOPs	Parameters
			PASCAL	VOC	NWPU			
			mAP	mAP@0.95	mAP	mAP@0.95		
CNN	SSD	ResNet18	51.10	22.98	40.34	16.01	353	26.29
	FCOS	ResNeXt	67.41	43.64	84.35	55.68	32	31.8
	Faster RCNN	VGG16	65.97	38.63	89.15	61.23	63	41.1
	CenterNet	ResNet34	57.64	30.90	81.28	48.76	130	41.7
	MobilenetV3	MBNV3	52.40	30.12	62.97	40.56	14	8.86
	YOLOV3	ResNet50	64.08	43.76	84.50	52.34	112	43.1
	YOLOV4	Darknet-53	62.70	49.90	87.80	60.89	101	44.9
	YOLOV5	CSPDarknet53	67.10	49.73	89.11	65.43	109	46
	YOLOV6	CSPDarknet53	67.16	54.28	89.78	52.10	150	59
	YOLOV7	CSPRep53	67.80	44.30	91.03	63.12	104	36
	YOLOV8	CSPDarknet18	53.01	36.66	90.43	60.80	32	16.1
	YOLOV11	CSPResp53	71.70	47.3	92.68	68.89	105	38.8
Transformer	DETR	ResNet50	62.40	43.00	81.28	35.9	101	36.74
	Deformable DETR	ResNet50	60.56	51.92	79.30	40.50	179	39.83
	PR-Deformable DETR	ResNet50	/	/	88.30	43.20	151	46.07
	RT-DETR	HGNetv1	69.41	50.59	92.60	60.62	136	163
	Our Work	HGNetv2	**71.20**	**52.51**	**94.92**	**68.10**	57	77.3

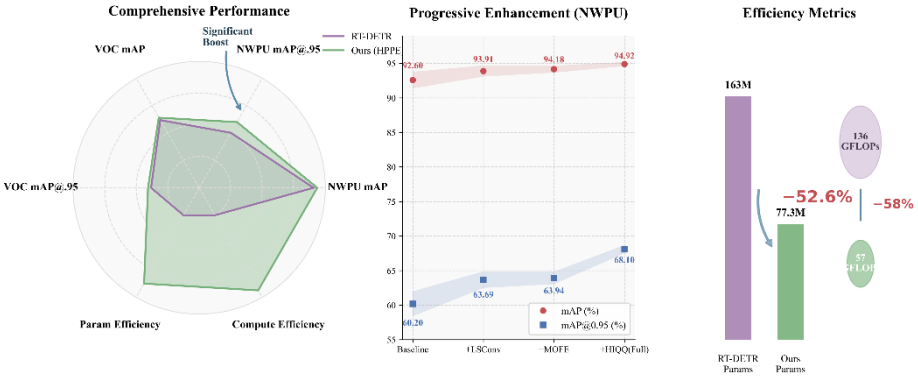

Fig. 3. A comprehensive evaluation of HPPE versus RT-DETR. (a) Radar chartshows balanced gains on NWPU and VOC. (b) Progressive enhancement illus-trates stable improvements. (c) Efficiency metrics confirm 52.6% parameter and 58% FLOP reduction.

3.1 Datasets and Evaluation Criteria

Our evaluation uses two benchmarks: NWPU VHR-10[8] for small objects, and Pascal VOC[9] for general detection. Pascal VOC merges 2007 and 2012 splits, containing 16,551 training and 4,952 testing images. NWPU VHR-10 provides 650 foreground and 150 background images. Evaluation protocols follow[10].

3.2 Comparison with State-of-the-art Methods

To validate our model's superiority, we benchmark it against leading CNN-based and Transformer-based detectors. The CNN cohort includes FCOS[11], RetinaNet[12] , Faster R-CNN[13], CenterNet[14], MobileNet-V3[15], and the YOLO series (v3 to v11)[16, 17]. The Transformer cohort includes DETR, Deformable DETR, PR-Deformable DETR*, and RT-Detr. Hyperparameters are kept consistent within each cohort.

As shown in Table 2, CNN models typically converge in 75 epochs, whereas Transformer models require 100-125 epochs. Transformer detectors are inherently data-hungry, which explains their suboptimal mAP95 on the sparse NWPU VHR-10 dataset (e.g., 35.9% for DETR). However, our HPPE framework defies this limitation. It achieves 94.92% mAP and 68.10% mAP95 on NWPU, outperforming YOLOv11 on mAP and substantially surpassing computationally heavy models like RT-Detr. Remarkably, HPPE secures these gains while requiring 58.8M fewer parameters than RT-Detr, proving its efficiency in generating stable, high-quality queries even with sparse training data.

Table 2. NWPU VHR-10 Ablation Experiments

Model	DSconv	Linear_snake	Heat Map MOFE	Positional EMB HIQQ	Precesions	Recall	mAP	mAP@95
Baseline	-	-	-	-	88.9	89.9	92.6	60.2
Baseline+Parts	√		-	-	89.0↑+0.10	90.10↑+0.20	92.71↑+0.11	62.32↑+2.12
	√	√	-	-	90.74↑+1.84	0.62↑+0.72	93.91↑+1.31	63.69↑+3.49
	√	√	√	-	90.75↑+1.85	0.51↑+0.61	94.18↑+1.58	63.94↑+3.74
	√	√	√	√	93.17↑+4.97	0.40↑+0.50	94.92↑+2.32	68.10↑+7.90

3.3 Ablation Studies

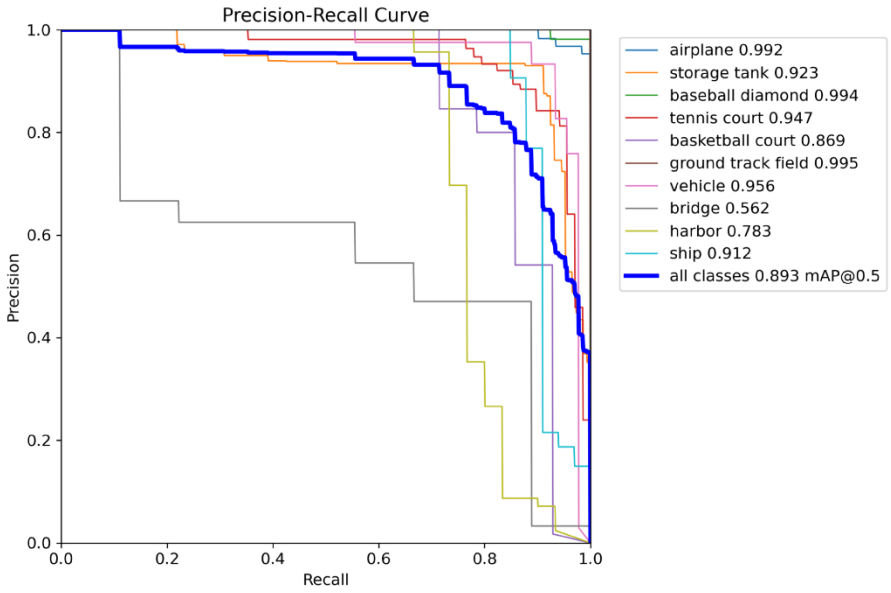

Fig. 4. Precision-Recall (PR) curve of the proposed method on the NWPU VHR-10 dataset, demonstrating consistent performance across different classes.

Results on NWPU VHR-10. The systematic experiments validate the progressive enhancement mechanism of each component on the NWPU VHR-10 dataset. The baseline model achieves 92.6% mAP and 60.2% mAP95 without any proposed modules. The incorporation of Linear-Snake Conv significantly improves fine-grained geometric perception through dual-path complementary architecture feature extraction, boosting mAP95 by 3.4% to 63.6% alongside a 1.8% precision improvement to 90.7%. Activating MOFE elevates mAP to 94.18% through heatmap-driven positional encoding optimization, representing a 1.58% gain over baseline. Finally, integrating the HIQQ module achieves comprehensive performance breakthroughs: the heatmap-induced query refinement sharpens detection precision to 93.1% (+4.97% absolute improvement) while propelling mAP to 94.92% (+2.32% over baseline) and mAP95 to 68.10% (+7.90% over baseline). This progressive improvement demonstrates that MOFE enhances background suppression through semantics-sensitive positional encoding, while HIQQ further optimizes object localization precision via decoder-side heatmap-guided high quality queries refinement.

- Results on PASCAL VOC. The progressive implementation demonstrates consistent improvements across all metrics. Starting from the baseline (69.4% mAP, 50.5% mAP95), the sequential integration of Linear-Snake (+0.91% mAP95) confirms their geometric modeling benefits. While MOFE strengthens positional-semantic coupling (+0.83% mAP95), the complete framework with HIQQ achieves

optimal performance: +2.10% precision (78.80%), +1.80% mAP (71.20%) and 52.51% mAP95 (+2.01%). Notably, the recall-precision equilibrium improves progressively, maintaining a 0.82% recall gain. Though the gains are relatively smaller than NWPU due to the dataset's lower object density, these results validate the framework's cross-domain robustness, particularly the critical role of HIQQ in distilling heatmap-derived queries.

Systematic experiments on the NWPU VHR-10 and PASCAL VOC datasets validate the progressive enhancement mechanisms of each component. HIQQ significantly improves the quality of query representations in both datasets, particularly achieving substantial performance gains in the high-density target scenarios of NWPU. MOFE enhances background suppression through optimized positional encoding driven by semantic detection. Linear-Snake Conv improves feature extraction for small objects through dual-path complementary architecture feature extraction, regardless of whether the scenario involves dense or sparse targets.

To further illustrate the detection performance, Figure 4 presents the Precision-Recall (PR) curve on the NWPU VHR-10 dataset, demonstrating high average precision across various small object categories. Additionally, Figure 5 visualizes the normalized confusion matrices for both the NWPU VHR-10 and PASCAL VOC datasets, showing the model's strong capability in distinguishing between different classes and reducing background false positives.

Model	DSconv	Linear_snake	Heat Map	Positional EMB	Precesions	Recall	mAP	mAP@95
			MOFE	HIQQ				
Baseline	-	-	-	-	76.7	76.7	69.4	50.5
Baseline+Parts	√		-	-	77.20↑+0.50	62.93↑+0.03	70.29↑+0.89	51.17↑+0.67
	√	√	-	-	78.41↑+1.71	63.10↑+0.20	70.19↑+0.79	51.41↑+0.91
	√	√	√	-	78.52↑+1.82	63.55↑+0.65	70.47↑+1.07	51.33↑+0.83
	√	√	√	√	78.80↑+2.10	63.72↑+0.82	71.20↑+1.80	52.51↑+2.01

3.4 Ablation Study on MultiHead Of Detr

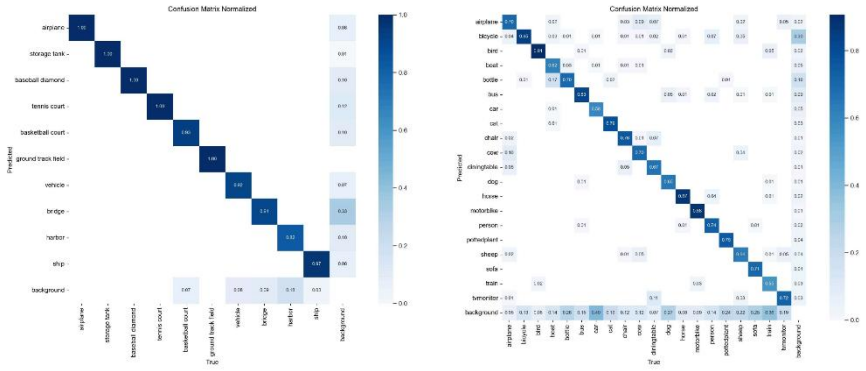

Fig. 5. Normalized confusion matrices of the proposed method on the NWPUVHR-10 dataset (left) and the PASCAL VOC dataset (right).

- Results on Encoder-Decoder architecture params.The encoder uses 3.01 MB for feature extraction, while the decoder requires 16.8 MB. This highlights the decoder's critical role in precise localization, ensuring efficiency with balanced parameters.

4 Conclusion and Outlook

In this paper, we proposed HeatMap Positional Embedding (HPPE), a novel framework for improving background noise suppression in object detection by adaptively coupling positional encoding with semantic detection priors. HPPE leverages gradient-weighted class activation maps to guide the generation of high-quality positional embeddings, enabling more effective feature integration in the encoder-decoder architecture. We introduced two key components: the Multi-Scale ObjectBox-Heatmap Fusion Encoder (MOFE) and the HeatMap Induced High-Quality Queries for Decoder (HIQQ), which jointly enhance feature representation and query quality while suppressing irrelevant background noise.

Extensive experiments show that HPPE significantly improves detection performance on both general and small-object benchmarks. On NWPU VHR-10, our method achieves 94.92% mAP and 68.10% mAP@0.95, outperforming state-of-the-art models like YOLOv11 and RT-DETR. On PASCAL VOC, we obtain 71.2% mAP and 52.51% mAP@0.95, with substantial improvements in precision and recall. Notably, HPPE reduces model parameters by 52.6% and FLOPs by 58% compared to RT-DETR, demonstrating its efficiency.

Our proposed visualization framework provides interpretable insights into how positional embeddings are refined, enabling better understanding of the model’s detection behavior. This interpretability not only aids in debugging and tuning but also provides a foundation for future research on semantically guided attention mechanisms.

Future work will explore the extension of HPPE to video detection and segmentation tasks, where temporal consistency and dynamic background modeling can further benefit from its adaptive fusion strategy. We believe HPPE offers a scalable and effective path toward lightweight, high-performance visual transformers for real-world applications.

5 Reference

1. Chen, K., Wang, Z., Wang, X., Gong, D., Yu, L., Guo, Y., Ding, G.: Towards real-time object detection in GigaPixel-level video. Neurocomputing. (2022).

2. Fan, J., Liu, H., Yang, W., See, J., Zhang, A., Lin, W.: Speed Up Object Detection on Gigapixel-Level Images With Patch Arrangement. In: CVPR (2022).

3. Han, J., Ding, J., Li, J., Xia, G.-S.: Align deep features for oriented object detection. IEEE TGRS. (2021).

4. Li, W., Chen, Y., Hu, K., Zhu, J.: Oriented reppoints for aerial object detection. In: CVPR (2022).

5. Yang, X., Yan, J., Liao, W., Yang, X., Tang, J., He, T.: Scrdet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. IEEE TPAMI. (2022).

6. 涂淑琴, 汤寅杰, 李承桀, 梁云, 曾扬晨, 刘晓龙: Behavior recognition and tracking of group-housed pigs based on improved ByteTrack algorithm. Transactions of the Chinese Society for Agricultural Machinery. 53, 264–272 (2022).

7. Zeng, Y.: HMPE: HeatMap Embedding for Efficient Transformer-Based Small Object Detection. arXiv preprint arXiv:2504.13469. (2025).

8. Cheng, G., Zhou, P., Han, J.: Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE transactions on geoscience and remote sensing. 54, 7405–7415 (2016).

9. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV. (2010).

10. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16965–16974 (2024).

11. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9627–9636 (2019).

12. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017).

13. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NeurIPS (2015).

14. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6569–6578 (2019).

15. Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., others: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019).

16. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767. (2018).

17. Khanam, R., Hussain, M.: Yolov11: An overview of the key architectural enhancements (2024). arXiv preprint arXiv:2410.17725. (2024).