Dense Pedestrian Detection Algorithm Based on an Improved Real-Time DEtection TRansformer

Xue Li, Yujun Zhang*

Abstract—**Pedestrian detection in crowded scenes remains challenging due to occlusion, scale variation, and computational constraints. This paper proposes an enhanced RT-DETR-based framework with three key improvements: 1) Wavelet Pooling replaces conventional up/downsampling operations to preserve small-object details and enhance multi-scale feature extraction; 2) Cascaded Group Attention (CGA) substitutes traditional multi-head self-attention, reducing computational complexity while improving global context modeling; 3) Small-Object Enhanced Pyramid (SOEP) optimize multi-scale feature fusion robustness. Evaluations on the CrowdHuman dataset demonstrate performance gains of 1.6% and 2.7% in Precision and mAP50-95, respectively, with exceptional results in occluded and small-object scenarios. The solution not only delivers efficient real-time detection but also exhibits strong generalizability for applications including instance segmentation, video surveillance, and autonomous driving.**

Index Terms—**pedestrian detection, dense scenes, real-time detection, deep learning.**

I. Introduction

With the rapid advancement of computer vision and deep learning technologies, pedestrian detection has made substantial progress [1] and has found widespread applications in areas such as autonomous driving, video surveillance, and person re-identification [2]. Despite these advancements, pedestrian detection in dense crowd scenarios still faces significant challenges, including occlusion, target overlap, scale variation, and computational limitations [3].

Pedestrian detection technology originates from general object detection algorithms. Contemporary object detection methods primarily rely on deep learning and can be broadly categorized into three major paradigms: two-stage R-CNN variants, single-stage YOLO models [4], and DETR-based approaches. Among these, single-stage CNN-based detectors are particularly suitable for real-time applications due to their efficiency. However, their performance degrades in dense crowd scenarios. This is primarily due to two factors: first, CNN-based models typically rely on Non-Maximum Suppression (NMS) as a post-processing step, which increases computational overhead and slows detection speed when handling numerous overlapping targets. Second, in densely populated scenes such as crowded streets, occlusion among pedestrians, vehicles, and buildings, as well as overlapping multi-scale objects and complex backgrounds, hinders the ability of convolutional kernels to extract detailed object features and spatial relationships, ultimately reducing detection accuracy [5].

In contrast, Transformers have shown superior capabilities over traditional CNNs in modeling long-range dependencies within sequential data. The multi-head self-attention mechanism enables the effective capture of global context and complex inter-object relationships. Vision Transformer (ViT), the first pure Transformer-based image classification model, has demonstrated performance surpassing CNNs when trained on large-scale datasets. This has led to the emergence of hybrid models that integrate the strengths of both CNNs and Transformers.

DETR (DEtection TRansformer) is a representative example of such hybrid architectures, combining CNNs for feature extraction with Transformers for global reasoning. Its end-to-end design eliminates the need for anchor boxes and NMS, achieving more accurate and stable object detection. However, the high computational cost associated with large Transformer modules remains a major bottleneck for real-time deployment.

To address this issue, RT-DETR [6] was proposed as a more efficient alternative. It outperforms state-of-the-art YOLO models [7] in both accuracy and inference speed. RT-DETR adopts ResNet as the backbone for feature extraction and introduces an efficient hybrid encoder that separates intra-scale attention from cross-scale fusion to capture global dependencies effectively. Additionally, it incorporates a PAFPN-like structure for top-down and bottom-up feature fusion. The resulting dense predictions are processed through a query selection mechanism and passed to the decoder, which predicts object categories and locations, retaining the Top-K results as final outputs.

Despite its improvements, RT-DETR faces limitations in dense crowd detection tasks. Specifically, the downsampling operations in the ResNet backbone cause the loss of fine-grained information relevant to small objects. As noted in prior studies, small object features are typically concentrated in shallow feature maps, yet are easily lost through multi-stage downsampling. Furthermore, existing research highlights both the strong correlation between feature maps and the inefficiency of standard convolutions due to redundant computations. To mitigate these issues, we replace conventional downsampling modules with Wavelet Pooling, which integrates wavelet transform and pooling operations, preserving structural detail and enhancing the representation of small objects.

Moreover, the original Attention-In-Feature Interaction (AIFI) module primarily focuses on global interactions within individual scales. However, in high-resolution feature maps, relying solely on global attention results in insufficient local feature extraction, limiting the ability to capture

Manuscript received April 6, 2025; revised July 4, 2025.

The work was supported by the Natural Science Foundation of China (No.62272093) and by Liaoning Key Laboratory of the Internet of Things Applicaiton Technology on Intelligent Construction.

Xue Li is a graduate student of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (e-mail: islixue@foxmail.com).

Yujun Zhang is a Professor of School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. (Corresponding author, e-mail: 1997zyj@163.com).

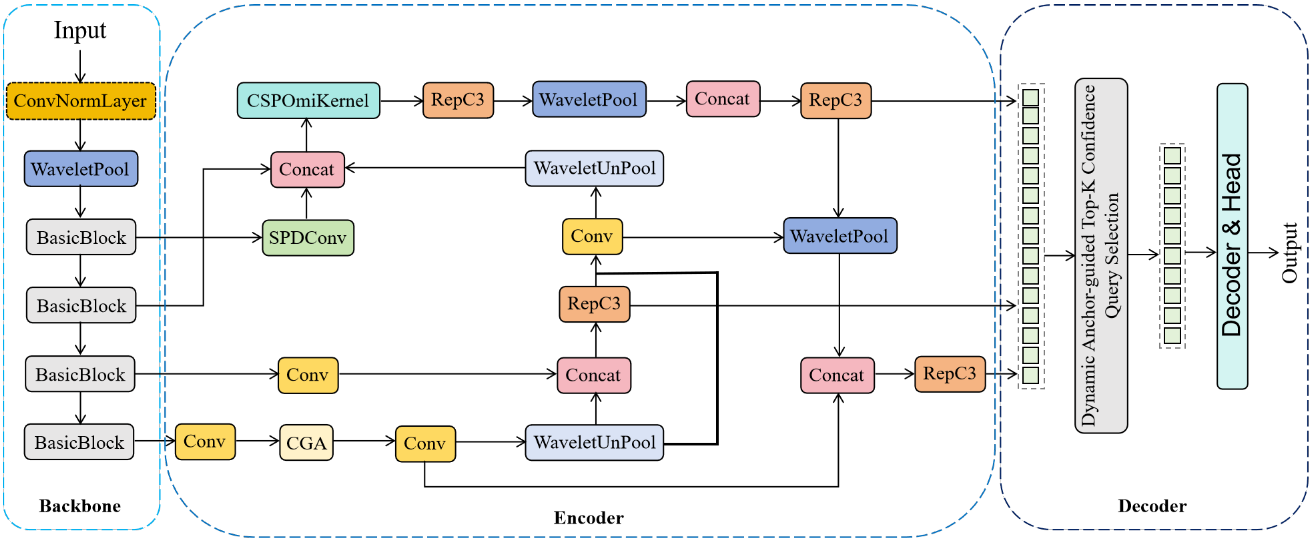

Fig. 1: Overall architecture of the model

fine-grained details such as edges and textures. Additionally, AIFI lacks effective support for multi-scale feature fusion—particularly the integration of low-level detail with high-level semantics—which leads to information loss during feature propagation. The Cross-scale Feature Fusion (CCFF) module in RT-DETR, while efficient and lightweight, mainly focuses on medium and large objects. It lacks the capacity to enhance critical regions and often processes irrelevant background regions inefficiently, which negatively impacts both detection accuracy and real-time performance—especially in crowded scenes with severe occlusion.

To further enhance the model's performance in dense pedestrian detection tasks, this paper introduces several key improvements. First, Wavelet Pooling [8] replaces traditional downsampling operations to preserve small-object features. Second, the standard multi-head self-attention mechanism is upgraded to CGA (Cross-scale Global Attention) [9], which strengthens multi-scale feature interactions and improves adaptability to densely populated scenes. Finally, the feature fusion strategy is refined by integrating the SOEP (Semantic-Oriented Edge Perception) module, which enhances the fusion of low-level spatial details and high-level semantic features. Together, these modifications improve the model's robustness, reduce computational cost, and enhance real-time detection capability in complex environments with dense, multi-scale objects [10].

In summary, the improvements and contributions of this article are as follows:

1) Integrated wavelet transform with pooling operations to optimize up/downsampling, improving multi-scale feature extraction while minimizing feature map resolution loss.
2) Replaced multi-head attention with a grouped cascaded mechanism, reducing computational complexity and enhancing long-range dependency modeling. This preserves global information representation, boosting detection accuracy.
3) Combined SPDConv(Space-to-Depth Convolution) and CSPOmniKernel module(a convolutional module based on CSP(Cross Stage Partial) structure and OmniKernel) to preprocess and fuse multi-scale features, enabling global-to-local feature learning. This elevates small-object detection precision while maintaining low computational overhead and post-processing costs.

II. Related Work

Pedestrian detection initially relied on handcrafted features such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT), combined with traditional classifiers like Support Vector Machines (SVMs) [11] and Haar cascades [12]. While these early approaches provided fundamental detection capabilities, they exhibited poor robustness in complex environments due to limited feature expressiveness and sensitivity to occlusion, scale variation, and background clutter.

The emergence of Convolutional Neural Networks (CNNs) significantly advanced the field. Two-stage detectors, such as the R-CNN series, achieved remarkable accuracy improvements but suffered from limited real-time performance. In contrast, single-stage frameworks like YOLO and SSD offered efficient end-to-end detection pipelines suitable for real-time applications. Subsequent iterations of YOLO [13] (e.g., YOLOv7 [14] and YOLOv8) introduced architectural refinements and enhanced training strategies. However, these models still encounter limitations in detecting small objects and handling densely crowded scenes [15].

More recently, Transformer-based models have opened new avenues for object detection. DETR, as a pioneering work in this domain, employs self-attention mechanisms to model global dependencies, thereby eliminating the need for components like anchor boxes and Non-Maximum Suppression (NMS). Various DETR variants have been proposed to address its limitations. Conditional DETR improves the object query mechanism, Deformable DETR introduces multi-scale deformable attention to reduce computational costs [16], and Sparse DETR adopts sparse attention for greater efficiency. Despite achieving high detection accuracy, DETR and its variants generally suffer from substantial computational overhead, limiting their deployment in real-time scenarios.

To bridge this gap, RT-DETR was proposed as a lightweight and real-time DETR-based model. It enhances

TABLE I: Advantages of Wavelet Pooling

Metric	Wavelet Pooling	Traditional Pooling
Info Retention	Preserves global structure + low-frequency texture	Keeps local maxima/mean values
Noise Robustness	Actively suppresses high-frequency noise	Vulnerable to local noise interference
Scale Adaptability	Multi-level decomposition enables multi-scale perception	Single-scale perception only
Reconstruction Capability	Partial detail recovery via IDWT (Inverse DWT)	Irreversible information loss

computational efficiency through optimized feature extraction and fusion modules while maintaining high accuracy. This advancement positions DETR-based architectures ahead of traditional YOLO-series algorithms in real-time object detection tasks, offering new momentum for progress in pedestrian detection. RT-DETR builds upon the Transformer architecture to explore more efficient strategies for feature encoding and target prediction, thereby improving both detection precision and inference speed [17]. In contrast to the YOLO series, which relies on anchor-based mechanisms and feature pyramid networks, DETR-based models utilize self-attention to enable end-to-end object detection without requiring NMS, achieving a more effective trade-off between accuracy and speed.

Currently, RT-DETR demonstrates the capability to perform object detection on high-resolution inputs at high frame rates, while maintaining precise classification and localization performance [18]. Its superior real-time characteristics make it particularly well-suited for time-sensitive applications such as autonomous driving and intelligent video surveillance. With continued architectural optimization, RT-DETR achieves a promising balance between accuracy and efficiency, marking a significant milestone in the evolution of object detection technology.

III. Methodology

The emergence of RT-DETR fills the gap in the application of DETR series in real-time monitoring. Compared with YOLO series, it achieves a more ideal balance between accuracy and speed. Traditional max pooling and downsampling convolutions lose high-frequency details such as edges and textures during dimensionality reduction. On the other hand, RT-DETR relies on the global modeling ability of the Transformer, but the basic convolution operations in the feature extraction stage cannot fully capture multi-scale information. Therefore, this paper proposes the introduction of Wavelet Pooling wavelet transform and its inverse operations to ensure bounding box regression and feature alignment for object detection tasks.

Secondly, in order to enhance multi-scale feature interaction and solve the memory consumption and computational complexity of global attention, we divide the input features into multiple groups, independently calculate attention within each group, and use cascading operations to cascade multiple attention modules, gradually enhancing the feature interaction effect.

Thirdly, replacing traditional convolutions with SPDConv [19] to convert spatial information into channel information enhances the model's ability to capture local details. Finally, by introducing the CSPOmniKernel module, efficient integration of multi task shared features was achieved, significantly optimizing the fusion effect of multi-scale features and enhancing the model's receptive field and feature representation ability. The network architecture of the model is shown in the Fig. 1.

A. Wavelet Pooling

In the RT-DETR model, upsampling and downsampling are key steps in feature extraction and fusion. Traditional pooling methods are prone to losing detailed information during downsampling, resulting in decreased performance of the model in detecting small targets and edge blurred targets. Upsampling operations such as deconvolution or bilinear interpolation are also difficult to fully recover high-frequency information, which affects the accuracy of target localization. Therefore, this article introduces WaveletPool for better information retention and reconstruction. In the upsampling stage, WaveletUnPool fuses low-frequency and high-frequency information through inverse wavelet transform to reconstruct finer feature maps, significantly improving the accuracy of object detection, especially in small object detection. In the downsampling stage, WaveletPool utilizes a multi-component decomposition mechanism to preserve the main information and store high-frequency components to assist in dimensionality reduction, while extracting multi-scale information and reducing redundant calculations, making the feature pyramid network more efficient in multi-scale object detection.

In dense pedestrian detection tasks, targets often have variable scales, severe occlusion, and blurry edges, which traditional pooling methods are difficult to handle. By replacing max pooling with WaveletPool, the model can better preserve multi-scale information in the early feature extraction stage, significantly enhancing its ability to capture pedestrian edges and texture features, and reducing information loss. By combining the inverse wavelet transform of WaveletUnPool, the model can more accurately reconstruct feature maps, improving the localization accuracy and recall rate of dense pedestrian detection.

1) WaveletPool: Traditional pooling operations (such as max pooling) are limited by local receptive fields and may lose some useful information (such as edges and textures), resulting in limited feature representation capabilities. In order to address this issue and further improve the feature extraction capability of the model by better preserving more useful information while reducing resolution, this paper introduces the Wavelet Pooling method based on Discrete Wavelet Transform (DWT) to achieve multi-scale feature compression, maximizing the preservation of key information

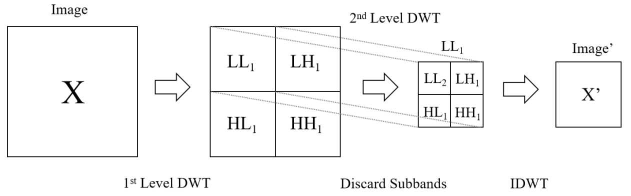

Fig. 2: WaveletPool module principle

while reducing resolution. The core process is shown in Fig. 2.

The first decomposition The input feature map $X \in \mathbb{R}^{H\times W\times C}$ is decomposed row-wise and column-wise to generate the low-frequency component LL_1 and high-frequency components LH_1, HL_1, and HH_1.

$$\mathbf{X} \rightarrow \mathbf{DWT}\{LL_1, LH_1, HL_1, HH_1\} \tag{1}$$

Where low-frequency component is the global structure and main texture extracted by the low-pass filter h, which represents the low-frequency approximation information of the image. The high-frequency component is the detail features obtained by capturing the vertical, horizontal, and diagonal directions of the image through the high pass filter g. It is sensitive to features and has high information redundancy. Specifically, it can be defined as:

$$\begin{aligned} LL_1 &= (X * h)_{\downarrow 2} * h^T_{\downarrow 2} \\ LH_1 &= (X * h)_{\downarrow 2} * g^T_{\downarrow 2} \\ HL_1 &= (X * g)_{\downarrow 2} * h^T_{\downarrow 2} \\ HH_1 &= (X * g)_{\downarrow 2} * g^T_{\downarrow 2} \end{aligned} \tag{2}$$

where $\downarrow 2$ represents downsampling with a stride of 2, and $*$ represents convolution operation.

High-frequency subbands were discarded and only LL_1 was retained as the downsampling output. The feature map X size was reduced to $\frac{H}{2} \times \frac{W}{2} \times C$.

The second decomposition Perform DWT again on the reserved low-frequency sub-band LL_1 to extract coarser grained features (such as semantic information at the object category level), while ensuring the sensitivity of the network to targets of different scales (LL_1, LL_2) with varying sizes.

Advantage comparison Compared to max pooling or average pooling, Wavelet Pooling achieves information intensive downsampling by preserving global low-frequency components. The advantages of Wavelet Pooling are shown in Table I.

2) *WaveletUnPool:* To support subsequent upsampling or feature fusion, WaveletUnPool can choose to store high-frequency components and reconstruct high-resolution features with low-frequency components through Inverse Discrete Wavelet Transform (IDWT), which can significantly reduce the problem of information loss in traditional anti pooling operations.

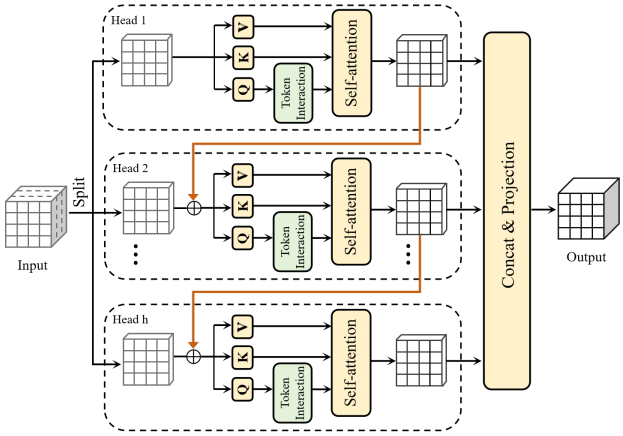

Fig. 3: CGA module

For the low-frequency sub-band $LL \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ and the high-frequency sub-band $\{LH, HL, HH\} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$, the reconstruction process of IDWT can be formally defined by Eq.3.

$$\begin{aligned} \mathbf{X_{recon}} = & \overbrace{(LL \uparrow 2) * h^T * h}^{\text{Low-frequency reconstruction}} + \overbrace{(LH \uparrow 2) * h^T * g}^{\text{Horizontal detail recovery}} \\ & + \underbrace{(HL \uparrow 2) * g^T * h}_{\text{Vertical detail recovery}} + \underbrace{(HH \uparrow 2) * g^T * g}_{\text{Diagonal detail recovery}} \end{aligned} \tag{3}$$

where notation $\uparrow 2$ denotes upsampling by a factor of two through interpolation (e.g., zero-padding or bilinear interpolation) to double the spatial dimensions of the feature maps. Through this framework, WaveletUnPool effectively restores high-frequency details, thereby enhancing both the precision and completeness of feature reconstruction.

B. *Cascaded Group Attention*

RT-DETR suffers from two key limitations due to its reliance on conventional multi-head self-attention mechanisms: First, when processing high-resolution images, its computational complexity becomes prohibitively high - particularly as the number of spatial positions N increases, the computational load grows quadratically, significantly degrading model efficiency. Second, the Adaptive Instance Feature Interaction (AIFI) module in its encoder requires explicit positional encoding to capture spatial information. This design not only increases model complexity but also restricts flexibility since the positional encoding is resolution-dependent.

To address the above issues, this paper adopts a new attention mechanism - cascaded group attention (CGA). This mechanism effectively reduces computational complexity while retaining the ability to express global information through innovative design of attention computation. The core idea is to divide the input features into multiple groups, calculate attention independently for each group, and then inject the output of the precursor group into the subsequent head through inter group cascade design to achieve feature fusion, and finally integrate the output. The process is shown in Fig. 3.

Feature Segmentation and Intra-group Attention Computation The input features $X_i \in R^{C\times N}$ are partitioned along the channel dimension into h groups (corresponding to h attention heads), where each group of features $X_{ij} \in \mathbb{R}^{(C/h)\times N}$ is independently assigned to a distinct attention

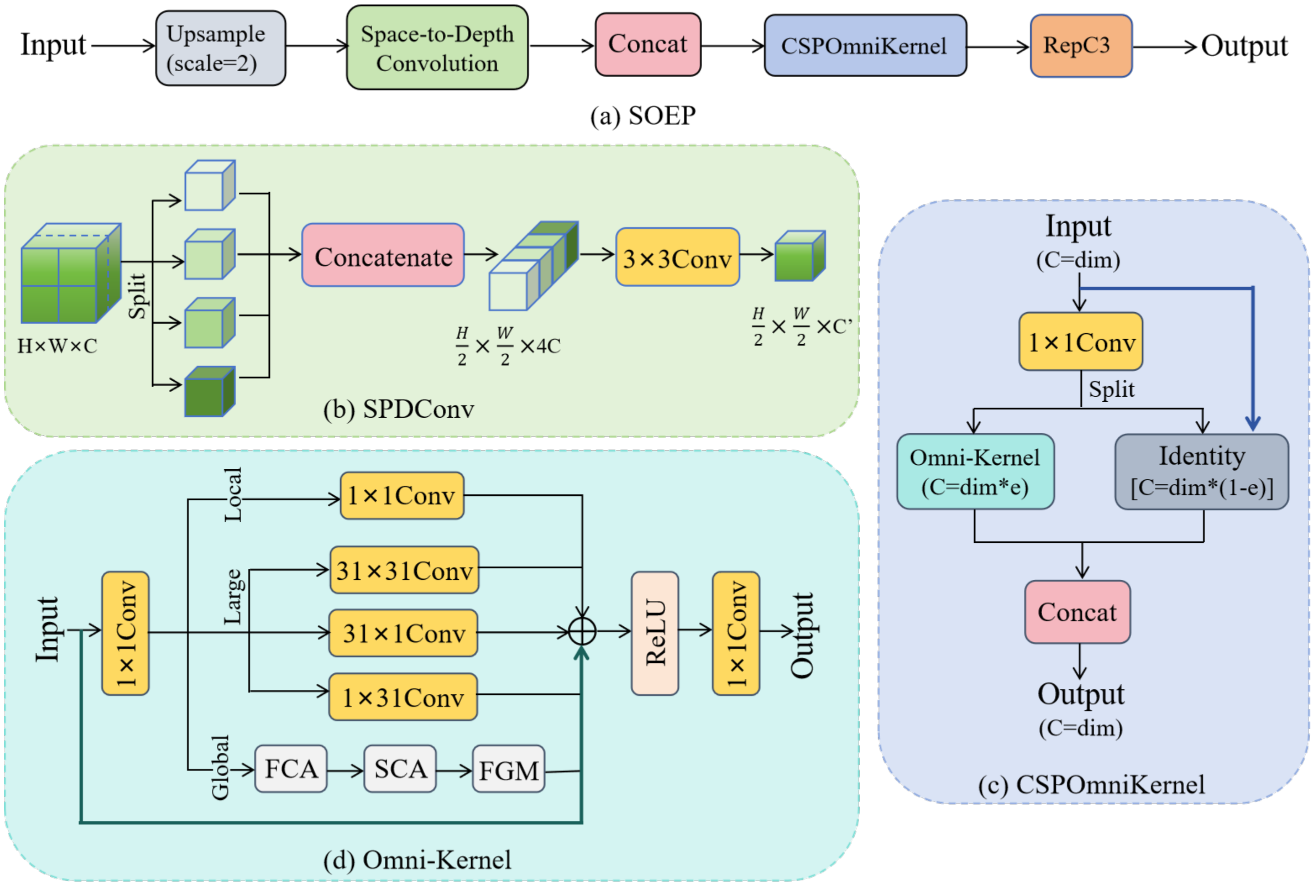

Fig. 4: SOEP architecture's main branch and sub-modules

head. Each head then computes intra-group contextual relationships via the self-attention mechanism:

$$\tilde{\mathbf{X}}_{\mathbf{ij}} = Attn(X_{ij}W_{ij}^{Q}, X_{ij}W_{ij}^{K}, X_{ij}W_{ij}^{V}) \tag{4}$$

where W_{ij}^{Q}, W_{ij}^{K}, $W_{ij}^{V} \in \mathbb{R}^{(C/h)\times d_k}$ correspond to the j-th attention head in layer i. This group-wise decomposition reduces the input channel dimension by a factor of h, this design significantly decreases the FLOPs for QKV projections and total parameters scale as $O\left((C/h) \times d_k\right)$ per head instead of $O\left(C \times d_k\right)$.

Inter-head Cascading and Feature Propagation Inject the output of the precursor head into the input of the subsequent head through cascading operations, gradually enhancing feature expression:

$$\mathbf{X}'_{\mathbf{ij}} = \begin{cases} X_{ij}, & j = 1 \\ X_{ij} + \tilde{X}_{i(j-1)}, & 2 \leqslant j \leqslant h \end{cases} \tag{5}$$

Here, X'_{ij}serves as the new input for the jth head, achieving cross head feature fusion. At the same time, a token interaction layer is introduced after Q projection to jointly model local and global relationships, further enhancing spatial perception capabilities.

Output Integration and Dimensionality Restoration The output of all heads is concatenated and linearly projected to restore dimensions:

$$\tilde{\mathbf{X}}_{\mathbf{i+1}} = \text{Concat}\left[\tilde{\mathbf{X}}_{\mathbf{i1}}, \tilde{\mathbf{X}}_{\mathbf{i2}}, \ldots, \tilde{\mathbf{X}}_{\mathbf{ih}}\right] W_i^P, \tag{6}$$

Among them, $W_i^P \in \mathbb{R}^{C\times C}$ is the dimension mapping matrix.

Through cascaded operations, the outputs from different groups are progressively propagated and ultimately aggregated into a global attention representation. This approach significantly improves computational efficiency while effectively capturing long-range dependencies, thereby enhancing object detection accuracy.

In the improved RT-DETR model proposed in this paper, CGA is applied to the AIFI (Adaptive Instance Feature Interaction) module to replace the traditional multi-head self-attention mechanism in RT-DETR. This improvement not only reduces computational complexity but also eliminates the reliance on explicit positional encoding, as CGA can implicitly capture spatial relationships through cascaded local and global information interaction. Additionally, inter-head cascading effectively increases network depth without requiring extra parameters, thereby enhancing model capacity. Experiments show that when processing high-resolution images, CGA maintains high detection accuracy while significantly reducing computational resource consumption, better supporting real-time object detection.

C. Small Object Enhance Pyramid

The issue of missed detection for small objects is particularly prominent in dense crowd detection tasks. Traditional Feature Pyramid Networks (FPN) typically rely on P3-P5 layers to process medium and large objects, exhibiting limited multi-scale representation capability and often insufficient detection accuracy for small objects. To address this problem, the Small Object Enhancement Pyramid (SOEP) proposes an innovative spatial-channel joint optimization strategy. Its core concept is to enhance small object features through the deep spatial reorganization technology SPDConv (Space-to-Depth Convolution) to achieve zero-information-loss downsampling, while incorporating the cross-domain feature fusion mechanism CSP (Cross-Stage Partial) to realize multi-granularity feature enhancement through frequency-spatial dual-domain attention. Additionally, it integrates the dynamic receptive field network OmniKernel, which employs anisotropic convolutional kernels to construct omnidirectional perception fields, thereby improving small object detection accuracy while maintaining the original FPN's computational efficiency.

The backbone architecture is shown in Fig. 4(a). First, the input feature map undergoes an SPDConv transformation to achieve lossless mapping from spatial dimensions to channel dimensions. Subsequently, it passes through the CSPOmniKernel module, where one portion of features remains unchanged while another portion undergoes Omni-Kernel processing, employing multi-scale convolutions to construct omnidirectional receptive fields and enhance small object feature representation. Finally, RepC3 is employed for feature refinement to ensure effective information transmission.

1) SPDConv: SPDConv is a convolutional operation that achieves downsampling by converting spatial information into channel information while enhancing channel-wise interactions. It can preserve more detailed information while reducing the resolution of feature maps.

For an input X of size $r \times r$, non-overlapping grid partitioning is performed through tensor slicing operations. The spatial information of each sub-region is then flattened and stacked along the channel dimension, increasing the channel count to $C \times r^2$ while reducing spatial dimensions by a factor of r. The merged feature map subsequently undergoes a 3×3 convolution to reduce channel dimensions to the target size C' while maintaining spatial resolution at half the original size. This process prevents high-frequency signal loss inherent in traditional pooling or strided convolution, preserving local spatial structures to enhance feature representation while completely retaining small object details such as edges and textures. The structure is illustrated in Fig. 4(b), where $r = 2$.

$$\mathbf{F}_{\text{SPD}} = \text{Conv}_{3\times3}\left(SPDConv\left(X\right)\right) \tag{7}$$

2) CSP: CSP is a network architecture that enhances gradient flow and feature representation capability through partial feature division and cross-stage transmission. The structure is shown in Fig. 4(c).

The input feature map F_{SPD} (denoted as X' for notational convenience in this section) $\in \mathbb{R}^{H\times W\times C}$ is first processed by a 1×1 convolution to adjust channel dimensions, yielding intermediate features $X' \in \mathbb{R}^{H\times W\times C}$. Subsequently, X' is split along the channel dimension into two parts at ratio e:Main branch X_{ok}(channel proportion e, default 0.25) undergoes Omni-Kernel operations to extract high-order features, producing X'_{ok}. Side branch X_{Identity}(remaining $1 - e$ channels) is directly propagated to avoid information loss. Finally, X'_{ok} and X_{Identity} are concatenated along the channel dimension and processed by another 1×1 convolution to adjust the channel count.

$$\begin{aligned} \mathbf{X}_{\text{ok}}, \mathbf{X}_{\text{Identity}} &= \text{Split}\left(\text{Conv}_{1\times1}(X), [eC, (1-e)C]\right) \\ \mathbf{F}_{\text{CSP}} &= \text{Conv}_{1\times1}\left(\text{Concat}\left(\text{Omni-Kernel}(X_{\text{ok}}), X_{\text{Identity}}\right)\right) \end{aligned} \tag{8}$$

This design reduces computational complexity from $O(C^2)$ to $O(eC^2)$, while the identity connection in the side branch preserves the original gradient path, mitigating the vanishing gradient problem in deep networks and facilitating gradient interaction between the main and side branches.

3) Omni-Kernel: Omni-Kernel is a feature enhancement module that integrates multi-scale depthwise convolutions with frequency-spatial attention kernels, designed to improve the model's multi-scale perception capability through heterogeneous convolutional kernels and dynamic attention mechanisms. Its core concept lies in decoupled modeling via global, large-branch, and local branches to capture full-scale information ranging from global context to local details, thereby strengthening the model's perceptual ability across different scales and locations. The structure is shown in Fig. 4(d).

The input featuresX_{ok} are first preprocessed by a 1×1 convolution followed by GELU activation, yielding X', which is then processed in parallel through Global, Large, and Local branches.

Global branch The Frequency Channel Attention (FCA) module first applies a 2D Fourier transform to X' to map the features into the frequency domain, obtaining X_{fft}, which enhances the model's sensitivity to high-frequency information such as textures and edges. It then generates frequency-domain attention weights α_{freq} through adaptive average pooling and a 1×1 convolution, which are element-wise multiplied with X_{fft} to produce X_{fca}.

$$\begin{aligned} \mathbf{X}' &= \text{GELU}(\text{Conv}_{1\times1}(X_{\text{ok}})) \\ \mathbf{X}_{\text{fft}} &= \mathcal{F}(X') \\ \alpha_{\text{freq}} &= \text{Sigmoid}\left(\text{Conv}_{1\times1}\left(\text{AdaptiveAvgPool}(X')\right)\right) \\ \mathbf{X}_{\text{fca}} &= \mathcal{F}^{-1}(\alpha_{\text{freq}} \cdot X_{\text{fft}}) \end{aligned} \tag{9}$$

Subsequently, Spatial-Channel Attention (SCA) is applied to X_{fca} for spatial compression and convolutional operations, which suppresses noise while enhancing small object saliency. The module generates spatial attention weights α_{spat}, performs element-wise multiplication with X_{fca}, and finally refines the features through the Feature Gating Module (FGM) to produce X_{sca}.

$$\begin{aligned} \alpha_{\text{spat}} &= \text{Sigmoid}\left(\text{Conv}_{1\times1}\left(\text{AdaptiveAvgPool}(X_{\text{fca}})\right)\right) \\ \mathbf{X}_{\text{sca}} &= \text{FGM}(\alpha_{\text{spat}} \cdot X_{\text{fca}}) \end{aligned} \tag{10}$$

Large and Local branches The model employs heterogeneous depthwise separable convolutions (DWConv) with varying kernel sizes — 31×1 (vertical elongated kernel), 1×31 (horizontal elongated kernel), 31×31 (global large kernel), and 1×1 (local detail) — to capture multi-scale features across different orientations and scales. These features are then combined with the original input and Global branch through residual connections.

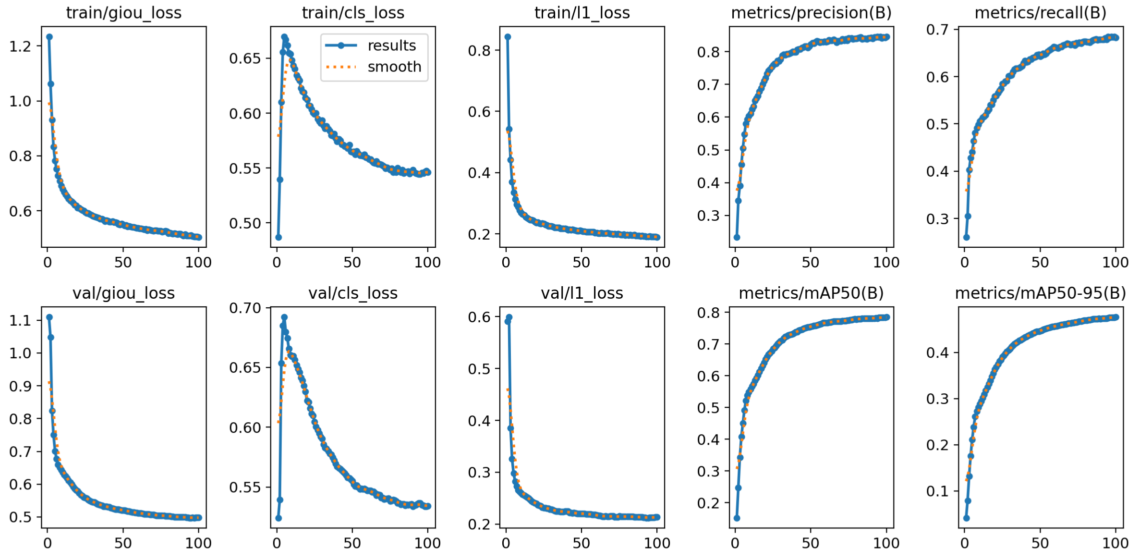

Fig. 5: Experimental result

$$
\begin{aligned}
\mathbf{X}_{\text{out}} &= \text{DWConv}_{31\times1}(X) + \text{DWConv}_{1\times31}(X) \\
&\quad + \text{DWConv}_{31\times31}(X) + \text{DWConv}_{1\times1}(X) \\
&\quad + X_{\text{sca}} + X
\end{aligned} \tag{11}
$$

Finally, the ReLU activation function and a 1×1 convolution are used to adjust the channel dimension, and the final feature is outputted as X'_{ok}

$$
\mathbf{X}'_{\text{ok}} = \text{Conv}_{1\times1}(\text{ReLU}(X_{out})) \tag{12}
$$

The Omni-Kernel method achieves collaborative optimization of multi-scale feature extraction and dynamic attention modulation while maintaining lightweight, providing an efficient solution for small object detection.

Combining these innovative designs, SOEP not only improves the detection accuracy of small targets, but also maintains low computational overhead, making it more adaptable and efficient in real-time object detection tasks.

IV. Experiments

A. Experimental environment

The configuration of the experimental environment is shown in Table 2.

B. Dataset

Due to discoveries during actual debugging and consensus among other researchers, the U-version RT-DETR model has difficulty converging. If there is too little training data, there may even be situations where indicator data is not displayed. At the same time, due to limitations in the experimental environment, this article chose a dataset with an equal number of images.

The experimental dataset in this article is a variant of the Crowd Human dataset, and the images are mainly sourced from urban environments, including streets, squares, and other places. The dataset contains 10000 images, of which 8500 are used for the training set, 500 for the validation set, and 1000 for the testing set. The total number of pedestrian annotations in the dataset exceeds 30000, and the number of pedestrian annotations in each image ranges from 1 pedestrian to hundreds of pedestrians. The main feature of the dataset is that it contains a large number of dense crowd scenes, where pedestrians are often occluded, overlapped, and obscured, posing high challenges and suitable for pedestrian detection and various visual tasks such as pedestrian tracking and pose estimation.

TABLE II: Experimental Environment Configuration

Category	**Configuration**
GPU	NVIDIA GeForce RTX 3090 (24GB)
CPU	Intel Xeon Platinum 8362, 15 vCPU, 2.80GHz
Operating System	Ubuntu 22.04 LTS
Python	3.12
Framework	PyTorch 2.3.0 + CUDA 12.1
Model Base	Ultralytics YOLOv8.0.201
Optimizer	SGD
Image Input Size	640×640
Epochs	100
Batch Size	4
Initial Learning Rate	0.01
Momentum	0.937
Weight Decay	0.0005
Loss Components	GIoU, Classification, L1
Loss Weights	$\lambda_1 = 0.05$, $\lambda_2 = 0.5$, $\lambda_3 = 0.1$

C. Evaluation metrics

To evaluate the performance of the improved RT-DETR model for dense crowd detection, the following metrics were used in this paper:

1) Precision(P): represents the proportion of samples predicted as positive by the model that are actually positive.

$$
\text{P} = \frac{\text{True Positives(TP)}}{\text{True Positives(TP) + False Positives(FP)}} \tag{13}
$$

TABLE III: Ablation experimental results

model	CGA	SOEP	Wavelet Pooling	P(%)	R(%)	mAP50(%)	mAP50-9(%)
Baseline				83.1	67.3	77.6	45.6
	✓			83.8	67.5	77.9	46.4
			✓	83.6	67.9	78.1	47.4
	✓	✓		80.5	64.3	74.8	46.9
ours	✓	✓	✓	84.7	69.4	79.3	48.3

TABLE IV: Comparative experimental results

model	P(%)	R(%)	mAP50(%)	mAP50-9(%)
Faster R-CNN	80.2	70.1	78.0	49.9
Mask R-CNN	79.8	69.5	77.0	47.0
YOLOv8n	81.5	65.3	76.4	48.1
YOLOv10n	80.5	64.3	74.8	46.9
YOLOv11n	81.0	64.9	75.4	47.4
DETR	61.1	60.5	63.1	45.9
Deformable-DETR	61.7	64.5	65.2	52.0
ours	84.7	69.4	79.3	48.3

TP represents the number of samples correctly predicted as positive by the model, while FP represents the number of samples incorrectly predicted as positive by the model. In addition, FP represents the number of samples correctly predicted by the model as negative examples, while FN represents the number of samples incorrectly predicted by the model as negative classes.

2) Recall(R): represents the proportion of samples that are actually positive and correctly predicted as positive by the model.

$$\mathrm{R} = \frac{\text{True Positives(TP)}}{\text{True Positives(TP) + False Negatives(FN)}} \tag{14}$$

3) mAP50: refers to the mean average precision when the IoU (Intersection over Union) threshold is 0.5.

$$\begin{gathered}\mathrm{AP} = \int_0^1 P(R)\,dR \\ \mathrm{mAP} = \frac{1}{N}\sum_{i-1}^{N} AP_i \\ \mathrm{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}\end{gathered} \tag{15}$$

P(R) is the precision of the recall rate of R. N is the total number of categories, and AP_i is the average precision of the i-th category. IoU is the ratio of the overlap area Area of Overlap between the predicted box and the real box to the union area Area of Union, used to measure the positioning accuracy of the detection box.

4) mAP50-90: refers to the average accuracy of the model within the range of IoU threshold from 0.5 to 0.95 (step size 0.05).

5) Experimental results and comparative analysis: During the training process, the model underwent 100 epochs of training and ultimately achieved an accuracy of 84.7% and a recall of 69.4% on the validation set. mAP50 and mAP50-95 were 79.3% and 48.3%, respectively, indicating that the model has good performance in object detection tasks, especially when the IoU threshold is 0.5. Fig. 5 presents the experimental results.

This article compares the method with the following classic object detection algorithms: R-CNN series, YOLO series, and DETR series. All comparison methods were run on the same training and testing sets, and with the same hyperparameter settings. Table IV shows the performance comparison between the proposed method and the comparative method on the crowd human validation set.

In summary, the model used in this experiment performs well in object detection tasks, especially achieving high levels of accuracy and mAP50. Although the performance decreases at higher IoU thresholds (mAP50-95), the overall model still has strong detection ability and high efficiency.

Analyzing the reason why the recall rate and mAP50-95 of Faster R-CNN are higher than our model, it is possible that Faster R-CNN uses a deeper and more complex backbone network and its two-stage detection mechanism to process detection tasks more finely, while our model has advantages in speed and real-time performance.

The mAP50-95 of Deformable DETR is particularly high, mainly because Deformable Convolution provides an adaptive receptive field and the Deformable Attention mechanism retains the advantage of sparse sampling.

In addition, the inference speed of the model is 12.5 ms per image, indicating that the model has high efficiency in practical applications and can meet the needs of real-time detection. It is expected to provide effective assistance for applications such as autonomous driving in the future.

D. Ablation experiment

In order to further demonstrate the performance of the model, we have recorded the ablation experiment results in detail, and the Table III shows the progressive growth of the model performance after adding different modules.

This paper uses RT-DETR (ReaNet-18 backbone) as the baseline model for performance reference. As shown in the table, the introduction of the CGA module leads to a slight improvement in Precision, Recall, mAP50, and mAP50-95, particularly with a 0.8 increase in mAP50-95, indicating that the CGA module enhances the model's overall detection capability by strengthening contextual information.

After incorporating the SOEP module, the improvements in Recall and mAP50-95 are more significant (Recall +1.5, mAP50-95 +1.6), demonstrating that the SOEP module, through multi-scale feature fusion, substantially enhances the model's detection performance for small objects and complex scenes. Precision and mAP50 also show slight improvements, suggesting that the SOEP module increases recall without significantly raising false detections.

With the addition of the WaveletPool module, all metrics exhibit further improvements. The notable gains in Recall and mAP50-95 indicate that the WaveletPool module strengthens the model's detection ability through more efficient feature extraction. The increase in Precision (+0.5) further confirms that the WaveletPool module improves recall while reducing false detections.

V. Conclusion

This study presents an enhanced RT-DETR-based model that significantly improves object detection performance in complex environments, especially for small and occluded targets. By integrating WaveletPool, CGA, and SOEP modules, the model strengthens feature extraction, contextual modeling, and multi-scale fusion. Experiments on the CrowdHuman dataset demonstrate that our improved model achieves increases of 1.6%, 2.1%, and 2.7% in Precision, Recall, and mAP50-95, respectively, showing superior robustness in small object and edge-blurred scenarios.These results highlight the model's robustness in detecting small and edge-blurred objects. Beyond performance, the proposed architecture offers strong adaptability to broader vision tasks, including instance segmentation, pose estimation, and medical imaging. Nevertheless, challenges such as increased training time, higher resource consumption, and decreased accuracy under extreme occlusion persist. Future work will aim to improve efficiency, explore more advanced fusion techniques (e.g., deformable attention), and enhance detection granularity through refined region proposal mechanisms or targeted data augmentation strategies.

References

[1] U. Gawande, K. Hajari, and Y. Golhar, "Real-time deep learning approach for pedestrian detection and suspicious activity recognition," *Procedia Comput. Sci.*, vol. 218, pp. 2438–2447, 2023.

[2] B. M. Alghamdi, "Human detection in crowded scenes using hybrid resnet," in *2023 International Conference on Sustainable Communication Networks and Application (ICSCNA)*. IEEE, 2023, pp. 516–520.

[3] J. Gupta and S. Sharma, "Towards real time hardware based human and object detection: A review," in *2022 4th International Conference on Artificial Intelligence and Speech Technology (AIST)*. IEEE, 2022, pp. 1–6.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[5] H. He, Z. Li, G. Tian, H. Chen, L. Xie, S. Lu, and H. Su, "Towards accurate dense pedestrian detection via occlusion-prediction aware label assignment and hierarchical-nms," *Pattern Recognition Letters*, vol. 174, pp. 78–84, 2023.

[6] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "DETRs beat YOLOs on real-time object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16 965–16 974.

[7] J. Pan and Y. Zhang, "Small object detection in aerial drone imagery based on yolov8," *IAENG Int. J. Comput. Sci*, vol. 51, no. 9, pp. 1346–1354, 2024.

[8] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *Int. Conf. Learn. Represent.*, 2018.

[9] X. Liu, H. Peng, N. Zheng, Y. Yang, H. Hu, and Y. Yuan, "EfficientViT: Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14 420–14 430.

[10] H. Cheng, "Intelligent pedestrian crossing system at urban intersection based on machine vision intelligent recognition system," in *2024 International Conference on Integrated Circuits and Communication Systems (ICICACS)*. IEEE, 2024, pp. 1–5.

[11] M. Bilal and M. S. Hanif, "Benchmark revision for hog-svm pedestrian detector through reinvigorated training and evaluation methodologies," *IEEE transactions on intelligent transportation systems*, vol. 21, no. 3, pp. 1277–1287, 2019.

[12] S. Zhang, C. Bauckhage, and A. B. Cremers, "Informed haar-like features improve pedestrian detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 947–954.

[13] X. Li and Y. Zhang, "A lightweight method for road damage detection based on improved yolov8n." *Engineering Letters*, vol. 33, no. 1, pp. 114–123, 2025.

[14] W. Teng, H. Zhang, and Y. Zhang, "X-ray security inspection prohibited items detection model based on improved YOLOv7-tiny," *IAENG Int. J. Appl. Math.*, vol. 54, no. 7, pp. 1279–1287, 2024.

[15] B. Ganga, B. Lata, and K. Venugopal, "Object detection and crowd analysis using deep learning techniques: Comprehensive review and future directions," *Neurocomputing*, p. 127932, 2024.

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

[17] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, vol. 28, no. 1, pp. 135–151, 2024.

[18] H. Peng and S. Chen, "Fedsnet: the real-time network for pedestrian detection based on rt-detr," *Journal of Real-Time Image Processing*, vol. 21, no. 4, p. 142, 2024.

[19] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2022, pp. 443–459.