Resolution-Aware Criss-Cross Attention Detector for Small Object Detection in Aerial Images

Heyu Sun
sunheyu@mail.sdufe.edu.cn
School of Computing and Artificial Intelligence,
Shandong University of Finance and Economics
Jinan, China

Taoying Liu
lty@ict.ac.cn
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China

Xingzhou Zhang
zhangxingzhou@ict.ac.cn
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China

Qiang Guo*
guoqiang@sdufe.edu.cn
School of Computing and Artificial Intelligence,
Shandong University of Finance and Economics
Jinan, China

Abstract

Detecting small objects in large-scale, high-resolution aerial images presents significant challenges. Most existing detectors focus primarily on the design of detection heads and fusion layers, often overlooking information loss in the backbone and the excessive computational resources required, which are particularly constrained in aerial image analysis. To address the aforementioned challenges, we propose the Resolution-Aware Criss-Cross Attention Detector (RACDet), which effectively leverages the contextual information embedded in an innovative backbone RACNet of aerial images. By decomposing the position information into orthogonal horizontal and vertical components, we achieve efficient modeling of spatial dependencies. For each pixel, RACNet gathers contextual information from all other pixels in the same position, establishing position relationships early, which can guide the subsequent processing in convolutional networks across different resolutions. The proposed method not only provides an adaptive representation of feature maps at multi-scale resolutions using normalized position encoding, but also enhances the detection accuracy of small objects by leveraging a regression loss function based on smooth Gaussian Wasserstein distance. We evaluate our method on two challenging aerial image datasets, including VisDrone2019 and UAVDT. Comprehensive experiments show that our approach achieves state-of-the-art performance while significantly decreasing the number of FLOPs.

CCS Concepts

• **Computing methodologies → Object detection**.

Keywords

Aerial Images; Small Object Detection; RACDet; RACNet; GWD

ACM Reference Format:
Heyu Sun, Taoying Liu, Xingzhou Zhang, and Qiang Guo. 2025. Resolution-Aware Criss-Cross Attention Detector for Small Object Detection in Aerial Images. In *Proceedings of the 48th International ACM ICMR Conference on Research and Development in Information Retrieval (ICMR '25), June 30–July 3, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3731715.3733420

1 Introduction

Leveraging advancements in Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) [4], current object detectors (e.g., Faster R-CNN [22], YOLO [21], and DETR [1]) have achieved substantial progress. However, constructing detectors for aerial images remains a significant challenge, as current methods fail short of satisfactory results in terms of accuracy and efficiency.

Aerial images have several unique characteristics compared to traditional images. Firstly, they are characterized by high resolution and complex scenes. As illustrated in Fig. 1 (a), objects in aerial images, viewed from a top-down perspective, are surrounded by similar objects in both the horizontal and vertical directions. Traditional detectors typically downsample images without considering the viewpoint, which can significantly affect the resolution of objects. This leads to the weakening or disappearance of small object features, thereby degrading the detector's performance. Secondly, Fig. 1 (b) shows that aerial datasets consist of a significant proportion of small and medium objects. An analysis of the VisDrone2019 [8] dataset reveals that over 60.1% of the annotation samples are small objects, 34.2% are medium, and 5.7% are large. Similarly, in the UAVDT [7] training annotations, small objects account for 61.9%, medium objects for 36.3%, and large objects for just 1.8%. In general, small objects refer to objects with an area less than 32×32 [18], so effectively detecting small objects in high-resolution aerial images is key to improving detector performance.

CNNs have long served as the backbone for object detection methods [2, 12, 14, 20, 25, 37]. Although these methods show considerable potential, there remains significant room for improvement. During the forward propagation process, they only have a

[*] Corresponding author

ICMR '25, Chicago, IL, USA.

ACM ISBN 979-8-4007-1877-9/2025/06
https://doi.org/10.1145/3731715.3733420

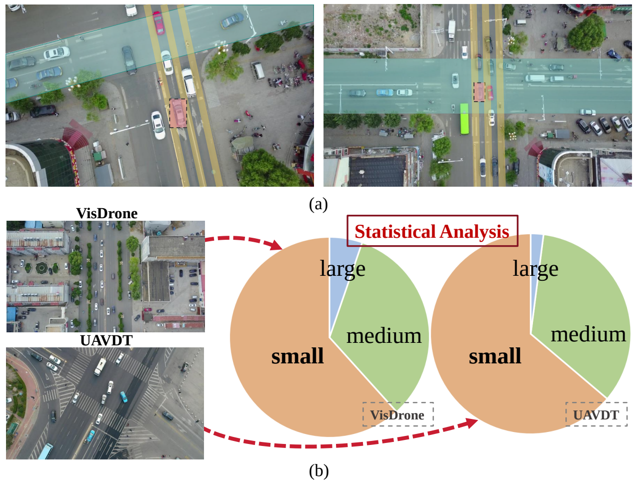

Figure 1: (a) In aerial images, objects are typically located in complex road scenes. (b) Statistical analysis reveals the distribution of small, medium, and large objects in the VisDrone2019 and UAVDT training datasets.

small receptive field, which is beneficial for detecting small objects, but the lack of fully considering the contextual information embedded within the feature maps leads to information loss in the backbone during the training phase. ViTs [5] are widely used as alternatives to CNNs for extracting contextual information by leveraging self-attention mechanisms to capture long-range dependencies. However, despite the theoretical advantages shown by transformer-based models [3, 32, 33, 35, 38, 40] in small object detection for aerial images, architectures based solely on attention still face challenges due to the unique characteristics of aerial images, including high resolution, dense object distribution, and complex scenes. The global self-attention incurs high computational costs on large images, while the weakening of local features can cause missed detections. Therefore, integrating the local perception advantages of convolutions remains necessary to balance detection accuracy and efficiency.

This raises the question: **Can high-resolution network captures global contextual information, similar to self-attention, by linking global features with smaller receptive fields to improve the accuracy and efficiency of small object detection?**

To address the information loss in high-resolution images during the backbone stage and improve the performance of small object detection while maintaining the overall lightweight structure, we propose the Resolution-Aware Criss-Cross Attention Detector (RACDet). Our detector uses an anchor-free CenterNet [42] as the baseline, directly predicting the center points of objects from the extracted feature maps without relying on anchor boxes. CenterNet predicts the heatmap of the object's center point and uses features around the center to infer attributes such as width and length. In the backbone stage, we introduce the Resolution-Aware Criss-Cross Network (RACNet), a high-resolution network integrated with global contextual information. Specifically, during feature propagation, we apply the Resolution-Aware Criss-Cross Attention enhanced with self-attention mechanisms to capture long-range dependencies and contextual information in aerial images. The resulting features are then fed into the high-resolution network [29] for detailed extraction across multiple resolution branches. In order to obtain accurate high-resolution representations while maintaining the overall lightweight structure of RACDet, we forgo designing complex feature fusion layers during the feature fusion stage after the backbone. Instead, the high-resolution branch outputs the feature map directly, while the low-resolution branch uses simple bilinear upsampling and channel adjustment to match it. This design is motivated by the parallel connection of multi-resolution convolutional streams during the backbone stage, where repeated multi-resolution fusion enables consistent maintenance of high-resolution representations throughout training.

In addition, to improve the detection accuracy of small objects, we design a smooth Gaussian Wasserstein distance (GWD) [36] as the loss function. This loss function reduce the confusion and redundancy of small objects in the background. As a result, it enhances the sensitivity to small object features. Additionally, to prevent performance degradation when detecting large objects, we combine the traditional Smooth L_1 loss. This hybrid approach balances the weights of large and small objects during optimization.

Our contributions can be summarized as follows.

- We propose the anchor-free Resolution-Aware Criss-Cross Attention Detector (RACDet) and design a backbone RACNet that leverages global contextual information to guide different resolution convolution streams for efficient and accurate small object detection in aerial images.
- We propose a regression loss function based on smooth Gaussian Wasserstein distance (GWD), which effectively enhances small object detection accuracy and improves sensitivity to small object features.
- Through extensive experiments conducted on the challenging VisDrone2019 and UAVDT datasets, we thoroughly demonstrate the effectiveness of RACDet, achieving state-of-the-art performance.

2 Related Work

2.1 General object detection

General object detection methods in computer vision are mainly classified into anchor-based and anchor-free approaches. Anchor-based methods are further divided into two-stage and one-stage methods. Two-stage methods, such as Faster R-CNN [23], generate candidate regions for classification and localization, while Mask R-CNN [10] extends Faster R-CNN by adding instance segmentation. One-stage methods, such as YOLO [21], directly regress bounding boxes and categories, improving detection speed. Anchor-free methods like FCOS [26] predict distances to the bounding box sides, while CenterNet [9] localizes objects using center points.

2.2 Small Object Detection in Aerial Images

In aerial image small object detection, many methods adopt a coarse-to-fine framework to address challenges like small object distribution and scale variation. ClusDet [34] refines search after coarse detection using a clustering-based query embedding, while DMNet

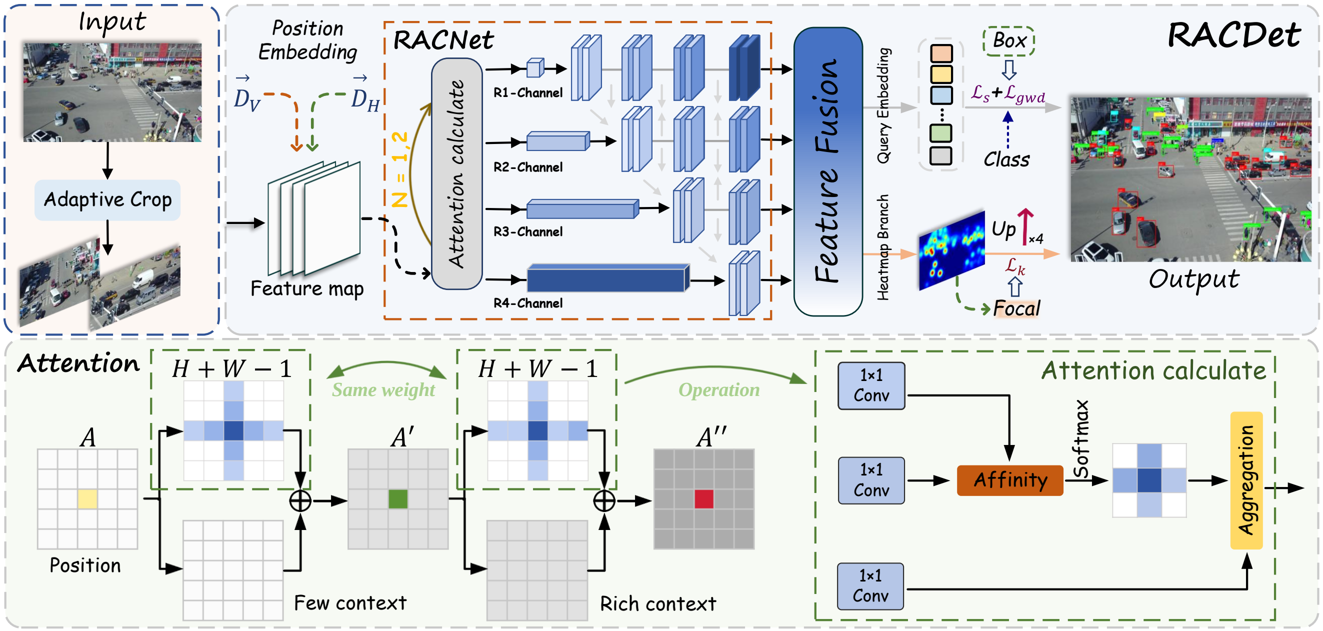

Figure 2: Overview of the proposed RACDet. The final detection results are obtained from the combination of the image crops and entire images. This includes the backbone RACNet, where *n* represents the number of iterations for the Attention module. R_i-channel (i = 1,2,3,4) refers to the adjustment of the feature map channels, with the range from 1 to 4 channels corresponding to resolutions from high to low. Finally, the query embedding and the heatmap are combined to obtain the detection results.

[16] simplifies training with a density map for cluster prediction. CRENet [30] improves clustering and fine-grained prediction, but these methods often incur high computational costs due to multiple inference stages, limiting their use on resource-constrained UAV platforms. Focus-and-Detect [15] tries to reduce complexity with a Gaussian mixture model, but many existing methods overlook information loss in the backbone network, negatively impacting real-time performance and detection accuracy.

2.3 Attention mechanisms in Small Object Detection

Attention mechanisms, particularly self-attention, have shown great potential for small object detection by capturing long-range dependencies in images [13]. For example, PSANet [39] uses point-wise spatial attention and bi-directional information propagation for scene parsing, while Channel Attention-based Detection [27] emphasizes more discriminative channels for small object detection. However, these methods struggle with high-resolution aerial images due to high computational complexity and large resource consumption. To address these challenges, we propose RACDet, which efficiently captures long-range dependencies and contextual information while guiding feature training at multiple resolutions.

3 METHODOLOGY

In this section, we provide a detailed introduction to our proposed Resolution-Aware Criss-Cross Attention Detector (RACDet), which is an improved version of CenterNet [9]. This detector is a powerful and efficient anchor-free object detection framework that utilizes high-resolution feature maps for small object prediction. In the backbone stage, we propose the Resolution-Aware Criss-Cross Network (RACNet). RACNet incorporates an enhancement of the Resolution-Aware Criss-Cross Attention through self-attention [28] mechanisms, capturing long-range dependences to extract contextual information from aerial images and accurately guiding the training of features at different resolutions.

3.1 RACNet

As shown in Fig. 2 , RACDet takes aerial images of different resolutions as input. After an initial convolution that generates the initial feature map without positional information, the feature map is passed through two consecutive Resolution-Aware Criss-Cross Attention modules. In the first module, local features gather contextual information from both the horizontal and vertical directions. The feature map generated by the first resolution-aware criss-cross attention module is then fed into the second module, where additional contextual information from the criss-cross paths is obtained. This process ultimately establishes entire image dependencies across all positions. To balance accuracy and efficiency, we introduce criss-cross positional encoding to enhance positional representation and constrain the parameters of both module to maintain a lightweight model.

3.1.1 Criss-Cross Position Encoding. In the computation of the attention matrix, we consider the influence of keys at different positions. To capture the spatial relationships in the attention mechanism, absolute positional encoding is applied to preprocess the feature maps $\mathbf{H} \in \mathbb{R}^{C\times W\times H}$. Normalize the positions of the feature map along the height H and width W.

$$\mathbf{p}_h = \sum_{i=0}^{H-1} \frac{i}{H}, \quad \mathbf{p}_w = \sum_{j=0}^{W-1} \frac{j}{W} \tag{1}$$

The position encodings $\mathbf{p}_h$ and $\mathbf{p}_w$ are normalized along the height H and width W dimensions.

$$\begin{aligned} \mathbf{pos}_h &= \mathbf{p}_h \cdot \exp\left(\frac{-\log(10000) \cdot \mathbf{k}_h}{C}\right) \\ \mathbf{pos}_w &= \mathbf{p}_w \cdot \exp\left(\frac{-\log(10000) \cdot \mathbf{k}_w}{C}\right) \end{aligned} \tag{2}$$

The $\mathbf{pos}_h$ and $\mathbf{pos}_w$ represent the position encodings for each pixel along the height and width of the feature map. $\mathbf{k}_h$ and $\mathbf{k}_w$ are the indices that scale the positional information. This scaling is crucial for capturing long-range dependencies.

$$\mathbf{pos} = \left[\, \sin(\mathbf{pos}_h) + \sin(\mathbf{pos}_w),\ \cos(\mathbf{pos}_h) + \cos(\mathbf{pos}_w) \,\right] \tag{3}$$

The final position information $\mathbf{pos}$ is represented by the sum of the sine and cosine functions of $\mathbf{pos_h}$ and $\mathbf{pos_w}$, allowing the position information to be encoded as periodic high-dimensional vectors, which can be easily incorporated into the attention operations.

3.1.2 Resolution-Aware Criss-Cross Attention. The predominant viewpoint of aerial images is top-down, with scenes often consisting of "criss-cross" environments, such as intersections and crosswalks, which makes them particularly suited for Resolution-Aware Criss-Cross Attention. The top-down perspective provides broad spatial context, while the "criss" environment leverages the attention mechanism to capture global dependencies.

Resolution-Aware Criss-Cross Attention is primarily an improvement on self-attention [28], where two consecutive row-column correlation matrix transformations replace the global correlation matrix transformation in self-attention. This modification significantly reduces the parameter count required by self-attention while maintaining high accuracy. However, single-head attention processes the input data in a single representation space, which may limit the model's capacity to capture the full complexity of the data. In contrast, multi-head attention maps the input data into multiple representation subspaces, allowing the model to capture different aspects of the information and integrate them into the final output.

In the specific implementation, early feature map $\mathbf{H} \in \mathbb{R}^{C \times W \times H}$ is input into the resolution-aware criss-cross attention module. Three 1×1 convolutions are applied to obtain the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ matrices, where $\{\mathbf{Q}, \mathbf{K}\} \in \mathbb{R}^{C' \times W \times H}$, and C' is the result of dimensionality reduction of C. Next, the **Affinity** operation is performed on $\mathbf{Q}$ and $\mathbf{K}$ to obtain the attention diagram $P \in \mathbb{R}^{(H+W-1)\times(H \times W)}$. For every position m of $\mathbf{Q}$ in the space dimension, a vector $\mathbf{Q_m} \in \mathbb{R}^{C'}$ can be obtained. Feature vectors can be extracted from the corresponding rows and columns in K to get a vector set $\mathbf{\Phi_m} \in \mathbb{R}^{(H+W-1)\times C'}$, then the **Affinity** operation is according to the following equation:

$$L_{i,m} = \mathbf{Q}_m^T \mathbf{\Phi}_{i,m} \tag{4}$$

The i-th element of $\mathbf{\Phi_m}$, denoted as $\mathbf{\Phi}_{i,m} \in \mathbb{R}^{C'}$, is used to calculate $L_{i,m} \in L$, which represents the degree of correlation between Q_m and $\Phi_{i,m}$, where $L \in \mathbb{R}^{(H+W-1)\times(H \times W)}$. After performing the **Affinity** operation on $\mathbf{Q}$ and $\mathbf{K}$, the result L is normalized via softmax, and the attention matrix $\mathbf{B}$ is obtained.

Then, $\mathbf{B}$ needs to perform the Aggregation operation.$\mathbf{V} \in \mathbb{R}^{C \times H \times W}$ is used for feature adaptation.$\mathbf{V}_m \in \mathbb{R}^{C}$ is a vector at position m in the spatial dimension of $\mathbf{V}$ and $\mathbf{\Psi}_m \in \mathbb{R}^{(H+W-1)\times C}$ is a collection of vectors in the same row and column as $\mathbf{V}_m$. The final feature map is obtained by the following equation:

$$A'_m = \sum_{i=0}^{H+W-1} B_{i,m} \Psi_{i,m} + A_m \tag{5}$$

where $\mathbf{B}_{i,m}$ is the weight assigned to the i-th vector at position m (with $i = [1, 2, \ldots, H+W-1]$), and $\mathbf{B}_{i,m}$ is a scalar. $\mathbf{A'}_m \in \mathbb{R}^C$ is the feature vector at position m, which is calculated by the resolution-aware criss-cross attention module.

3.1.3 Contextual Information for Resolutions. The resolution-aware criss-cross attention module allows each pixel in the feature map to gather information from its respective criss-cross path. However, pixels outside of this range cannot obtain information. To address this limitation, we stack two attention modules. The first module enables each pixel in the feature map to collect information from its corresponding criss-cross path. The second module then replicates this process, allowing each pixel to acquire global contextual information. The feature map $\mathbf{A'}$ is fed into another resolution-aware criss-cross attention module, resulting in a new feature map $\mathbf{A''}$. Thus, every pixel in $\mathbf{A''}$ contains the global spatial contextual information of the entire feature map. In contrast to traditional self-attention [28], which directly generates global contextual information, our method reduces both temporal and spatial complexity from $O(N^2)$ to $O(N\sqrt{N})$ by sequentially stacking two our modules.

We input the feature map $\mathbf{A''}$ into the subsequent high-resolution network, feeding global contextual information into the high-resolution convolutional network [29] for detailed feature extraction across different resolutions. This allows the model to learn the overall contextual information of the entire image across multiple subrepresentational spaces. By feeding the $\mathbf{A''}$ feature map into the branches of different resolutions, we effectively combine the local information from the low-resolution path while preserving the high-resolution features. The specific implementation is as follows:

$$A_i^{out} = \sigma(F_{attn}(A'', A_i)) \cdot \mathcal{T}_j(A_i) + A_i \tag{6}$$

In this formulation, A_i^{out} is the output feature map of the i-th resolution branch after attention enhancement. A is the input attention feature map containing global contextual information, and A_i is the original feature map from the i-th resolution branch. $\sigma(\cdot)$ is a nonlinear activation function. $F_{\text{attn}}(A'', A_i)$ is the attention interaction function same as Eq. 3.1.2 , calculates the attention weights to enhance A_i with A. $\mathcal{T}_j(\cdot)$ represents the feature transformation operation. The operator $\cdot$ stands for element-wise multiplication for adaptive enhancement. Finally, the residual connection formed by adding A_i helps alleviate the vanishing gradient problem while preserving the original feature information.

3.2 High-Resolution Representation

To achieve accurate high-resolution representations while maintaining the overall lightweight structure of RACDet, we avoid using complex feature fusion layers during the feature fusion stage after the backbone. Instead, the high-resolution branch directly outputs the feature map without any additional operations, while the low-resolution branch employs simple bilinear upsampling to match the high-resolution branch and adjusts the channels using 1D convolution. This results in the final high-resolution feature map. The motivation behind this approach lies in the fact that during the backbone stage. By leveraging low-resolution representations to enhance the high-resolution ones, we ensure that high-resolution representations are consistently maintained throughout the training process. The upsampling equation for the feature fusion layer is as follows:

$$A_{out} = \sum_{i=1}^{N-1} UpSample\left(R_i, 2^i\right) \cdot A_i^{out} \tag{7}$$

where R_i represents the different resolution branches, as shown in Fig. 2, when i=1, it corresponds to the high-resolution branch, while i=2,3,4 correspond to the low-resolution branches. The low-resolution branches upsample with a scaling factor of 2^i. This process enhances high-resolution features while maintaining a lightweight structure, resulting in the final high-resolution output A_i^{out}.

To improve the accuracy of object detection in densely packed small object regions, RACDet employs higher-resolution heatmaps for prediction. In CenterNet, each object is modeled as a single point representing the center of its bounding box, depicted by a Gaussian kernel in the heatmap. However, the sampling rate of the heatmap is reduced by a factor of 4× compared to the input image. This downsampling causes small objects to collapse into a few points, or even a single point, on the heatmap, making it difficult to accurately localize their centers. To address this issue, we focus on high-resolution representations of objects, which helps retain fine-grained detail and spatial structure, thereby improving the precision in localizing and detecting small objects.

3.3 Loss Function

CenterNet[9] combines three different losses to jointly optimize the entire network.

$$L_{det} = L_k + \lambda_{size} L_{size} + \lambda_{off} L_{off} \tag{8}$$

L_k is the modified focal loss used in CenterNet [42], L_{off} represents the center point offset loss, and L_{size} denotes L_1 loss. By default, λ_{size} and λ_{off} are set to 0.1 and 1.

However, the method of size regression using L_1 loss has limitations. L_1 loss is not sufficiently sensitive to small objects and updates slowly when gradients are small. Since objects in aerial images are often densely distributed and exhibit significant scale variations, L1 loss is more suitable for handling objects with distinct features in typical background scenarios. To address this issue, we introduce a smooth Gaussian Wasserstein distance, which effectively distinguishes small objects in densely packed regions. Specifically, for a bounding box $R = (cx, cy, w, h)$, where (Cx, Cy) denote the coordinates of the center, and w and h represent the width and height, respectively. The equation of its inscribed ellipse can be represented as

$$\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{\left(y-\mu_y\right)^2}{\sigma_y^2} = 1 \tag{9}$$

where (μ_x, μ_y) represents the center coordinates of the ellipse, where $\mu_x = cx$, $\mu_y = cy$, and the semi-axes lengths along x and y axes are given by $\sigma_x = \frac{w}{2}$, $\sigma_y = \frac{h}{2}$. The probability density function of a 2D Gaussian distribution is:

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})\right)}{2\pi |\Sigma|^{\frac{1}{2}}} \tag{10}$$

where $\mathbf{x}$, $\boldsymbol{\mu}$ and Σ denote the coordinate (x, y), the mean vector and the co-variance matrix of Gaussian distribution as:

$$(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) = 1 \tag{11}$$

According to Optimal Transport Theory, the Wasserstein distance between two distributions μ and ν can be computed as:

$$\mathbf{W}(\mu; \nu) := \inf \mathbb{E}\left(\|\mathbf{X} - \mathbf{Y}\|_2^2\right)^{1/2} \tag{12}$$

Given two 2D Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}1, \Sigma 1)$ and $\mathcal{N}(\boldsymbol{\mu}2, \Sigma 2)$, the Wasserstein distance is:

$$\mathbf{d}^2 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathbf{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right) \tag{13}$$

Note in particular we have:

$$\mathbf{Tr}\left(\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right) = \mathbf{Tr}\left(\left(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}\right)^{1/2}\right) \tag{14}$$

Since the detection task uses horizontal bounding boxes as the ground truth, we have $\Sigma 1 \Sigma 2 = \Sigma 2 \Sigma 1$. Then Eq.(13) can be rewritten as follows [36]:

$$\begin{aligned} \mathbf{d}^2 &= \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\right\|_F^2 \\ &= l_2\text{-norm}\left(\left[x_1, y_1, \frac{w_1}{2}, \frac{h_1}{2}\right]^\top, \left[x_2, y_2, \frac{w_2}{2}, \frac{h_2}{2}\right]^\top\right) \end{aligned} \tag{15}$$

where $\| \|_F$ is the Frobenius norm.

The loss defined in Eq.(15) may be overly sensitive to large errors. To address this, a nonlinear function is introduced, transforming the loss into an affinity measure $\frac{1}{\tau + f(\mathbf{W}^2)}$. The resulting GWD-based loss is expressed as [36]:

$$L_{gwd} = 1 - \frac{1}{\tau + f\left(\mathbf{d}^2\right)}, \quad \tau \geq 1 \tag{16}$$

where $f(\cdot)$ is the non-linear function to make the Wasserstein distance more smooth and expressive. τ is a modulated hyperparameter, which is empirically set as 1.

The smooth L_1 distance of the five points from the predicted point set and ground-truth bounding box is calculated using L_{smooth}, which is defined in Eq.17.

$$L_{smooth}(t) = \begin{cases} 0.5t^2 & \text{if } |t| < 1 \\ |t| - 0.5 & \text{otherwise} \end{cases} \tag{17}$$

Table 1: Comparison in terms of AP (%), Latency, and FLOPs on VisDrone. o, ca, aug respectively stand for the original validation set, cluster-aware cropped images, and augmented images. "-" indicates that the result is not reported.

Model	backbone	imgsz	test	AP^{val}	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{s}	AP^{val}_{m}	AP^{val}_{l}	Latency(ms)	FLOPs(G)
QueryDet	ResNet50	800	o	19.6	35.7	19.0	-	-	-	288	-
RetinaNet	ResNet50	800	o	20.2	36.9	19.5	-	-	-	14.7	210
Faster-RCNN	ResNet50	800	o	21.4	40.7	19.9	11.7	33.9	54.7	21.2	285
RTMDet-L	CSPNeXt-L	640	o	23.7	37.4	25.5	12.5	38.7	50.4	13.9	50.4
CenterNet	Hourglass104	800	o	27.8	47.9	27.6	21.3	42.1	49.8	95.2	1855
HRDNet	HRDNet	1333	o	28.3	49.3	28.2	-	-	-	-	421
GFLV1	ResNet50	1333	o	28.4	50.0	27.8	-	-	-	525	-
CEASC	ResNet50	1333	o	28.7	50.7	28.4	-	-	-	43.8	150
RACDet	**RACNet**	1024	o	**35.6**	**60.0**	**36.0**	**27.1**	**47.7**	**57.4**	175	104
ClusDet	ResNet50	1000	o+ca	26.7	50.6	24.7	17.6	38.9	51.4	273	-
DMNet	ResNet50	1500	o+ca	28.2	47.6	28.9	19.9	39.6	55.8	290	-
CDMNet	ResNet50	1000	o	29.2	49.5	29.8	20.8	40.7	41.6	-	-
GLASN	ResNet50	600	o+ca	30.7	55.4	30.0	-	-	-	-	-
AMRNet	ResNet50	1500	o+aug	31.7	-	-	23.0	43.4	**58.1**	-	-
YOLC	HRNet	1024	o+ca	31.8	55.0	31.7	24.7	42.3	45.0	441	151
CZDet	ResNet50	1200	o+ca	33.2	58.3	33.2	26.0	42.6	43.4	-	-
UFPMP-Det	ResNet50	1333	o+ca	36.6	62.4	36.7	-	-	-	152	205
RACDet	**RACNet**	1024	o+ca	**38.3**	**62.5**	**40.1**	**31.6**	**47.9**	52.5	402	155

Table 2: The detection performance of each class on VisDrone validation set. Ped. and Awn. are short for Pedestrian and Awning-tricycle. RS is short for random sampler.

Method	Backbone	Ped.	Person	Bicycle	Car	Van	Truck	Tricycle	Awn.	Bus	Motor
Comparison with other detectors											
RetinaNet+RS	ResNet50	13.0	7.9	1.4	45.5	19.9	11.5	6.3	4.2	17.8	11.8
FRCNN+RS	ResNet50	21.4	15.6	6.7	51.7	29.5	19.0	13.1	7.7	31.4	20.7
FRCNN+DSHNet	ResNet50	22.5	16.5	10.1	52.8	32.6	22.1	17.5	8.8	39.5	23.7
CenterNet	Hourglass104	28.7	18.5	12.7	57.9	37.6	29.6	17.4	11.4	43.2	20.9
YOLC	HRNet	37.4	24.3	21.3	64.3	43.8	34.0	26.5	17.9	53.2	33.6
RACDet	**RACNet**	**39.7**	**26.4**	**23.7**	**66.3**	**46.2**	**36.8**	**29.6**	**20.3**	**59.0**	**36.3**

During the early stages of model training, we observed that the Wasserstein distance d is typically high when handling large objects, while the L_1 loss tends to be overly sensitive to small objects. To address these issues, we propose an improved strategy that combines the GWD loss, L_1 loss, and L_{smooth} loss. The final loss function is formulated as follows:

$$L_{det} = L_k + \lambda_{gwd}L_{gwd} + \lambda_{l1}L_1 + \lambda_{l_s}L_{smooth} \tag{18}$$

where we set λ_{gwd} to 2, λ_{l1} to 0.5 and λ_{l_s} to 0.1.

4 Experiment

4.1 Datasets and Evaluation Metrics

We evaluate our approach on two publicly available aerial image datasets: VisDrone 2019 [8] and UAVDT [7]. The details of these datasets are as follows:

4.1.1 VisDrone. This dataset comprises 7,019 images captured by drones, with 6,471 images designated for training and 548 images for validation. It includes annotations across ten categories: bicycle, awning tricycle, tricycle, van, bus, truck, motor, pedestrian, person, and car. The images have an approximate resolution of 2000 x 1500 pixels, we use the validation set for evaluation.

4.1.2 UAVDT. The dataset consists of 38,327 images with an average resolution of 1,080×540 pixels. It includes three categories: car, bus, and truck. The dataset is split into 23,258 images for training and 15,069 images for testing.

4.2 Implementation Details

Using PyTorch and MMDetection, we trained models from scratch on the VisDrone [8] and UAVDT [7] datasets for 120 epochs, The model is trained for 160 epochs using the SGD optimizer with a momentum of 0.9 and weight decay of 0.0001 ,and applied data

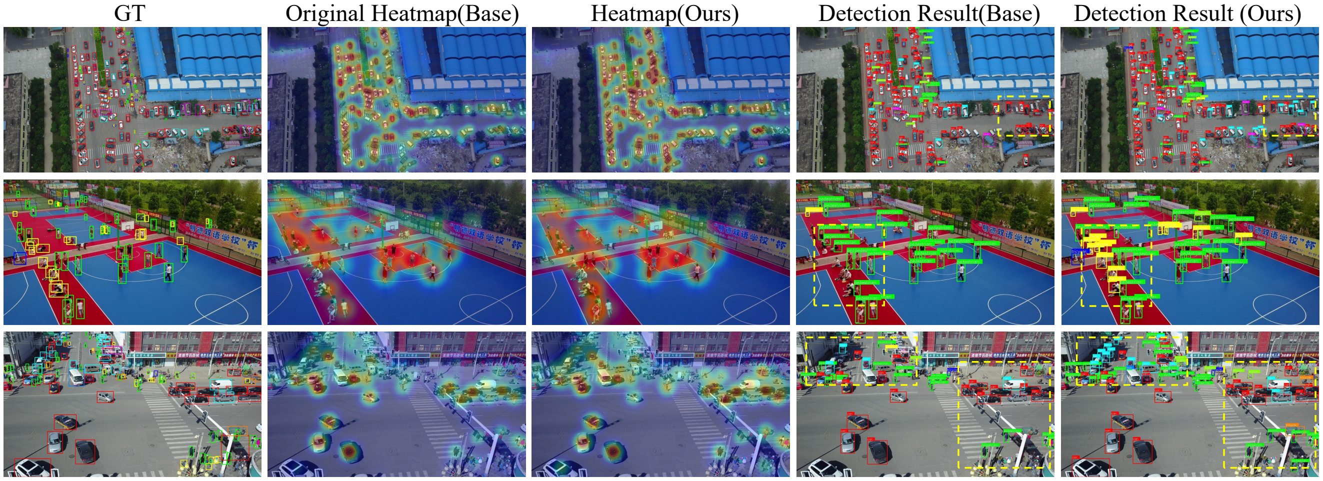

Figure 3: Visualization of the detection results and heatmaps on VisDrone. The highlighted areas represent the regions that the network is focusing on. The yellow box indicates the superiority of our method compared to the base model.

augmentation techniques such as mixup and Mosaic. It is run on the NVIDIA RTX 4090 GPUs platform with a batch size of 2. The initial learning rate is set to 0.0025 with a linear warm-up. The input resolution is configured as 1024 × 640 for both datasets, and inference is conducted using a single 4090 GPU.

4.3 Evaluation Measures

Following the evaluation protocol of the MS COCO [18] dataset, we utilize AP, AP_{50}, and AP_{75} as evaluation metrics. Here, AP refers to the average precision across all categories, while AP_{50} and AP_{75} indicate the average precision at IoU thresholds of 0.5 and 0.75, respectively. Additionally, we report the average precision for each object category to assess class-specific performance. To evaluate performance across different object scales, we adopt three metrics: AP_{small}, AP_{medium}, and AP_{large}. Finally, the efficiency of our approach is measured by calculating the processing time for a single original image per GPU.

4.4 Comparison with SOTA on Aerial Datasets

4.4.1 Results on VisDrone Dataset. The proposed detector demonstrates significant improvements in the key evaluation metric, mean Average Precision (mAP), on the VisDrone dataset compared to existing models. To further emphasize the balance between detection accuracy and efficiency on aerial images, we conduct comparisons with general object detectors. Evaluations are performed on both the original dataset and a cluster-aware cropped version, following the YOLC [19] approach.

On the original validation set, the results in Table 1 show that our proposed RACDet detector outperforms existing models, achieving a 7.8% improvement in AP over CenterNet. RACDet also surpasses the previous state-of-the-art lightweight detector CEASC [6] by 0.8%, and improves detection accuracy by 15.4% compared to the mainstream backbone RetinaNet [17]. It should be noted that on a single RTX 4090 GPU (without applying any model acceleration techniques), our model achieves an inference latency of 175 ms while significantly reducing FLOPs. After incorporating the cluster-aware Crops method, RACDet achieves the best performance to date with an mAP of 38.3%, representing a 5.1% improvement over AMRNet [31] and a 1.7% improvement over UFPMP-Det [37]. As shown in Table 2, our method attains the best performance across all object categories on the VisDrone validation set.

Table 3: Comparison in terms of AP (%) on UAVDT.

Model	AP^{val}	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{s}	AP^{val}_{m}	AP^{val}_{l}
R-FCN	7.0	17.5	3.9	4.4	14.7	12.1
FRCNN+FPN	11.0	23.4	8.4	8.1	20.2	26.5
CenterNet	13.2	26.7	11.8	7.8	26.6	13.9
ClusDet	13.7	26.5	12.5	9.1	25.1	31.2
DMNet	14.7	24.6	16.3	9.3	26.2	35.2
CDMNet	16.8	29.1	18.5	11.9	29.0	15.7
GLSAN	17.0	28.1	18.8	-	-	-
CEASC	17.1	30.9	17.8	-	-	-
AMRNet	18.2	30.4	19.8	10.3	31.3	33.5
RACDet	**19.0**	**33.0**	**20.2**	**13.3**	**32.1**	18.4

4.4.2 Results on UAVDT Dataset. The performance evaluation on the UAVDT dataset [7] leads to conclusions similar to those obtained from the VisDrone dataset. This indicates that general object detectors fail to achieve satisfactory detection results. Based on the experimental results using the UAVDT dataset (as shown in Table 3), we demonstrate that the proposed method outperforms the current state-of-the-art model AMRNet [31], achieving the highest detection accuracy with an mAP of 19.0%. Furthermore, compared to other methods, our approach significantly improves the accuracy of small and medium object detection, despite exhibiting relatively poorer performance for large objects. These findings further validate the effectiveness of our proposed detector.

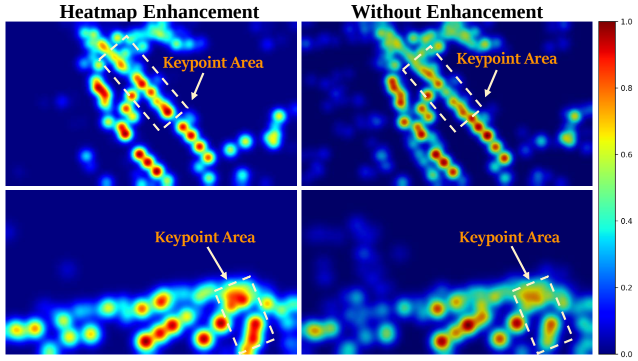

Figure 4: Comparing the enhanced heatmap (left) with the original (right), highlighting the keypoint regions within dense areas.

4.5 Ablation Study

To validate the effectiveness of the components in our detector, we conducted ablation experiments on the VisDrone dataset. In our detector, RACNet serves as a key component, demonstrating the most significant improvement in overall performance. Furthermore, the performance is further enhanced by designing a more efficient loss function.

4.5.1 Ablation of Backbones. To evaluate the effectiveness of our proposed backbone design, we conduct ablation studies by replacing the original Hourglass104 backbone in CenterNet [9] with RACNet and ResNet50 [11], a widely adopted backbone in many state-of-the-art detectors.In order to ensure a fair comparison, only the backbone is substituted, while other components such as the loss functions remain unchanged. Experimental results on the VisDrone validation set demonstrate that RACNet improves AP by 7.7%, AP_{50} by 10.6%, and AP_{75} by 8.6% compared to Hourglass104. Furthermore, RACNet consistently outperforms ResNet50 in all evaluation metrics. As shown in Table 5, these results validate that the proposed RACNet can significantly enhance the detection performance of existing architectures.

Table 4: Ablation study of different backbones.

Backbone	AP^{val}	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{s}	AP^{val}_{m}	AP^{val}_{l}
Hourglass104	28.5	49.5	29.2	21.5	43.5	50.1
ResNet50	29.3	51.5	28.2	20.2	41.8	47.5
RACNet(ours)	36.0	60.4	36.7	27.7	47.9	57.0

4.5.2 Ablation of Loss function. To further optimize the proposed RACDet, we made modifications to the regression loss. Specifically, we replaced the original loss with L_{smooth} and L_{gwd}. These losses were chosen to address the sensitivity to scale variations and the inconsistency between metrics and losses. The results of these experiments are shown in Table 5. When using the regression loss with $L_1 + L_{smooth}$, we observed an improvement in detection performance, with AP increasing to 29.5%. When $L_1 + L_{gwd}$ was used, the performance improved further, reaching 35.5% for AP. Combining both L_{smooth} and L_{gwd} with L_1 resulted in the highest performance, achieving AP of 36.0%, as shown in Table 5.

This demonstrates the impact of the modified loss functions on performance. When we finally combined L_{smooth}, we found that the detection performance showed only a slight improvement of 0.5%, which we attribute to small objects already obtaining global contextual information during the forward pass. During backpropagation, we typically only need to consider the loss between objects of different resolutions. Additionally, we tested IoU-based losses such as GIoU [24] and DIoU [41], and we also modified the GIoU loss to adapt it to our RACDet. These loss functions were chosen to address the sensitivity between different objects. However, after replacing the original regression loss, compared to the best performance, we observed a decrease of 1.8% in AP, 2.3% in AP_{50}, and 0.9% in AP_{75}. As shown in Fig. 3, the effectiveness of the smooth GWD loss in reducing background interference is validated.

Table 5: Ablation study of the loss function.

Method	AP^{val}	AP^{val}_{50}	AP^{val}_{75}	AP^{val}_{s}	AP^{val}_{m}	AP^{val}_{l}
$L_1 + L_{smooth}$	33.0	58.3	33.5	26.3	43.0	50.0
$L_1 + L_{gwd}$	35.5	58.5	36.2	27.1	47.5	55.6
$L_1 + L_{GIoU}$	34.2	58.1	35.6	26.0	46.7	54.9
$L_1 + L_{smooth} + L_{gwd}$	36.0	60.4	36.7	27.7	47.9	57.0

4.5.3 Design of the Heatmap Branch. In CenterNet [9], each object is modeled as a point representing the center of its bounding box, depicted by a Gaussian kernel in the heatmap. However, the sampling rate of the heatmap is reduced by a factor of 4× compared to the input image. To address this issue, RACDet focuses on high-resolution representations of objects and employs higher-resolution heatmaps for prediction. We enhance the heatmap branch by applying two transpositions, ensuring that the heatmap resolution matches that of the input image. We refer to this process as heatmap enhancement. As shown in Fig. 4 , the enhanced heatmap makes the keypoint regions more prominent compared to the original heatmap, leading to improved detection accuracy, particularly in densely packed areas.

5 Conclusions

This paper presents RACDet, an anchor-free object detection detector for high-resolution aerial images. We propose RACNet, which uses an improved self-attention mechanism to capture long-range dependencies and contextual information for small objects, while maintaining a lightweight structure. Experiments on two datasets show that our method outperforms state-of-the-art approaches. We hope that this work will inspire future research in aerial image detectors.

Acknowledgments

This work was supported by the Key Research and Development Program of Shandong Province under Grant 2024TSGC0109.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. Springer, 213–229.

[2] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. 2017. Learning efficient object detection models with knowledge distillation. *Advances in Neural Information Processing Systems* 30 (2017).

[3] Guang Chen and Yi Shang. 2022. Transformer for tree counting in aerial images. *Remote Sensing* 14, 3 (2022), 476.

[4] Wuyang Chen, Ziyu Jiang, Zhangyang Wang, Kexin Cui, and Xiaoning Qian. 2019. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8924–8933.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Representation Learning*.

[6] Bowei Du, Yecheng Huang, Jiaxin Chen, and Di Huang. 2023. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 13435–13444.

[7] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. 2018. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision*. 370–386.

[8] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. 2019. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. 2019. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6569–6578.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[13] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 603–612.

[14] Shuaixiong Hui, Qiang Guo, Xiaoyu Geng, and Caiming Zhang. 2023. Multi-guidance CNNs for salient object detection. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3 (2023), 1–19.

[15] Onur Can Koyun, Reyhan Kevser Keser, Ibrahim Batuhan Akkaya, and Behçet Uğur Töreyin. 2022. Focus-and-Detect: A small object detection framework for aerial images. *Signal Processing: Image Communication* 104 (2022), 116675.

[16] Weisheng Li, Xiayan Zhang, Yidong Peng, and Meilin Dong. 2020. DMNet: A network architecture using dilated convolution and multiscale mechanisms for spatiotemporal fusion of remote sensing images. *IEEE Sensors Journal* 20, 20 (2020), 12190–12202.

[17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Dollar Piotr. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*. 2980–2988.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[19] Chenguang Liu, Guangshuai Gao, Ziyue Huang, Zhenghui Hu, Qingjie Liu, and Yunhong Wang. 2024. YOLC: You Only Look Clusters for Tiny Object Detection in Aerial Images. *IEEE Transactions on Intelligent Transportation Systems* (2024).

[20] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision*. 116–131.

[21] J Redmon. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2016), 1137–1149.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[24] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 658–666.

[25] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4510–4520.

[26] Zhi Tian, Xiangxiang Chu, Xiaoming Wang, Xiaolin Wei, and Chunhua Shen. 2022. Fully convolutional one-stage 3d object detection on lidar range images. *Advances in Neural Information Processing Systems* 35 (2022), 34899–34911.

[27] Wei Tong, Weitao Chen, Wei Han, Xianju Li, and Lizhe Wang. 2020. Channel-attention-based DenseNet network for remote sensing image scene classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), 4121–4132.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[29] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 10 (2020), 3349–3364.

[30] Yi Wang, Youlong Yang, and Xi Zhao. 2020. Object detection using clustering algorithm adaptive searching regions in aerial images. In *Proceedings of the European Conference on Computer Vision*. Springer, 651–664.

[31] Z Wei, C Duan, X Song, Y Tian, and H Wang. 2020. AMRNet: Chips augmentation in aerial images object detection. *arXiv preprint arXiv:2009.07168* (2020).

[32] Xiangkai Xu, Zhejun Feng, Changqing Cao, Mengyuan Li, Jin Wu, Zengyan Wu, Yajie Shang, and Shubing Ye. 2021. An improved swin transformer-based model for remote sensing object detection and instance segmentation. *Remote Sensing* 13, 23 (2021), 4779.

[33] Zhenhua Xu, Yuxuan Liu, Lu Gan, Yuxiang Sun, Xinyu Wu, Ming Liu, and Lujia Wang. 2022. Rngdet: Road network graph detection by transformer in aerial images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 4707612.

[34] Fan Yang, Heng Fan, Peng Chu, Erik Blasch, and Haibin Ling. 2019. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8311–8320.

[35] Hao Yang, Dinghao Zhang, Anyong Hu, Che Liu, Tie Jun Cui, and Jungang Miao. 2022. Transformer-based anchor-free detection of concealed objects in passive millimeter wave images. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 5012216.

[36] Xue Yang, Junchi Yan, Qi Ming, Wentao Wang, Xiaopeng Zhang, and Qi Tian. 2021. Rethinking rotated object detection with gaussian wasserstein distance loss. In *Proceedings of the International Conference on Machine Learning*. PMLR, 11830–11841.

[37] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6848–6856.

[38] Yan Zhang, Xi Liu, Shiyun Wa, Shuyu Chen, and Qin Ma. 2022. GANsformer: A detection network for aerial images with high performance combining convolutional network and transformer. *Remote Sensing* 14, 4 (2022), 923.

[39] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. 2018. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision*. 267–283.

[40] Yongbin Zheng, Peng Sun, Zongtan Zhou, Wanying Xu, and Qiang Ren. 2021. ADT-Det: Adaptive dynamic refined single-stage transformer detector for arbitrary-oriented object detection in satellite optical imagery. *Remote Sensing* 13, 13 (2021), 2623.

[41] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12993–13000.

[42] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. 2019. Objects as points. *arXiv preprint arXiv:1904.07850* (2019).