This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Feature Information Driven Position Gaussian Distribution Estimation for Tiny Object Detection

Jinghao Bian[1] Mingtao Feng[1,3*] Weisheng Dong[1] Fangfang Wu[1] Jianqiao Luo[2]
Yaonan Wang[2] Guangming Shi[1]
[1]Xidian University, [2]Hunan University
[3]Jiangxi Communication Terminal Industrial Technology Research Institute

Abstract

Tiny object detection remains challenging in spite of the success of generic detectors. The dramatic performance degradation of generic detectors on tiny objects is mainly due to the the weak representations of extremely limited pixels. To address this issue, we propose a plug-and-play architecture to enhance the extinguished regions. We for the first time exploit the regions to be enhanced from the perspective of pixel-wise amount of information. Specifically, we model the entire image pixels feature information by minimizing Information Entropy loss, generating an information map to attentively highlight weak activated regions in an unsupervised way. To effectively assist the above phase with more attention to tiny objects, we next introduce the Position Gaussian Distribution Map, explicitly modeled using a Gaussian Mixture distribution, where each Gaussian component's parameters depend on the position and size of object instance labels, serving as supervision for further feature enhancement. Taking the information map as prior knowledge guidance, we construct a multi-scale position gaussian distribution map prediction module, simultaneously modulating the information map and distribution map to focus on tiny objects during training. Extensive experiments on three public tiny object datasets demonstrate the superiority of our method over current state-of-the-art competitors.

1. Introduction

Deep learning-based architectures have significantly advanced general object detection [27, 32, 34]. However, their performance declines sharply when detecting tiny objects [8]. As defined in the AI-TOD benchmark [44], tiny objects are categorized by pixel count: very tiny (2-8 pixels), tiny (8–16 pixels), and small (16-32 pixels). Given the prevalence of tiny objects in real-world contexts (e.g., traffic monitoring [24], sea rescue [56], and wildlife population assessment [19]), it is crucial to bridge the gap between tiny and general object detection.

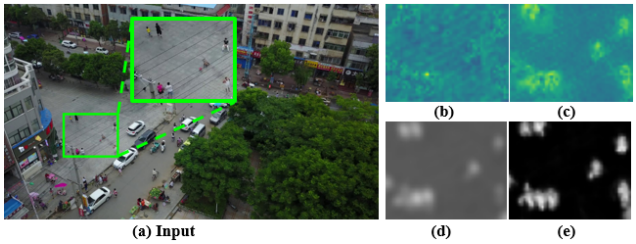

Figure 1. (a) Input image; (b) feature map P_2 from DetectoRS [31]; (c) enhanced feature; (d) and (e) are proposed information map σ and predicted position distribution map M_{pd2}.

Several pioneering efforts have been introduced to address this challenge. One prominent approach is scale-aware feature fusion, which constructs hierarchical feature representations [12, 22, 25, 26, 31, 38, 49]. This method leverages network features at varying depths to detect objects at corresponding scales, and fusion between deep and shallow features bridges the spatial-semantic gap between lower and higher pyramid levels, enhancing small object detection. Another approach emphasizes target regions using attention mechanisms [23, 29, 50, 52, 55]. Inspired by the human visual system [3, 9], attention-based methods filter irrelevant information and highlight tiny objects by modeling feature map importance weights.

Although existing methods partially mitigate this issue, they remain ineffective when tiny object pixel counts are extremely limited. We contend that previous approaches overlook a fundamental challenge: weak tiny object representations of extremely limited pixels. Unlike normal-sized objects, tiny object representations with limited pixels are consistently diminished through successive neural network downsampling, ultimately resulting in the suppressed activation of features. The information loss caused by downsampling is extremely fatal to tiny objects compared to general objects, leading to extremely weak and indiscriminative representations. As shown in Fig. 1(b), representations for tiny objects in generic detectors are faint and nearly in-

*Corresponding author

distinguishable from the background. Consequently, even when tiny object features are enriched through scale-aware feature fusion, the intrinsic weak representations of tiny objects still struggle with accurate detection. Moreover, attention-based methods, which rely on heuristic attention map generation, often fail due to the sparse pixels of tiny objects, allowing the background to dominate in local patches. So the attention map derived from such weak regions becomes unreliable, leading to suboptimal detection results.

Based on the preceding discussion, we argue that enhancing indiscriminative regions suffering from information loss is crucial for detecting tiny objects. To achieve this, we propose a novel plug-and-play framework for feature enhancement. Inspired by information theory [37], we first estimate entire image pixel-wise amount of information unsupervisedly by minimizing Information Entropy loss (which is a reflection of encoding cost of all pixels). The generated information map emphasizes salient regions, with salient targets having more amount of information than smooth background, displaying attentive regions for enhancement. To assist the information map with more attention to tiny objects, we introduce the Position Gaussian Distribution Map, which is modeled via Gaussian Mixture distribution, with each Gaussian component parameters adjusting to the position and size of the instance label. The distribution map can capture salient targets, with tiny objects having higher pixel intensities than general objects, increasing more attention to tiny objects for enhancement. We incorporate information map prior with multi-scale features to predict the distribution map in a supervised manner, simultaneously modulating the information map and distribution map to focus on tiny objects during training. As shown in Fig. 1, our information map and predicted distribution map successfully identify and enhance the key information loss regions, making the tiny object representation salient. Extensive experiments on three datasets validate the superiority of our method. Our main contributions are:

- We are the first to enhance the weak representations of tiny objects from the perspective of pixel-wise amount of information. The attentive information map is captured by minimizing Information Entropy loss in an unsupervised manner.
- With the information map as a prior guide, we leverage multi-scale features to predict Position Gaussian Distribution Map, further highlighting tiny objects.
- The proposed method can be flexibly integrated into any FPN-like detectors. Experimental results demonstrate the consistent gain of our plug-and-play module and its superiority over existing state-of-the-art methods.

2. Related works

2.1. General Object Detection

Anchor-based. Anchor-based detectors predefine anchors covering different scales and aspect ratios and train the network through a label assignment strategy to adjust and refine the anchors. Specifically, they can be subdivided into one-stage methods (such as YOLOV3 [33], SSD [27], RetinaNet [35]) and two-stage methods (such as Faster R-CNN [34], Cascade R-CNN [5], Cascade RPN [42]).

Anchor-free. Anchor-free detectors use the center point or key point to predict the target position directly, which effectively alleviates the problem of hyper-parameter sensitivity of predefining anchors. FCOS [40] and FoveaBox [18] predict the bounding box in a central points fashion. Another type of paradigm is to locate objects by key point. The representative works are CornerNet [20], Grid R-CNN [28], ExtremeNet [57] and RepPoints [51].

2.2. Tiny Object Detection

Sample-oriented strategies. Kisantal *et al.* [17] oversamples images with small objects and augments them by copy-pasting small objects. NWD-RKA [47] replaces the standard label assignment strategy IoU with Normalized Wasserstein Distance and ranking-based assigning strategy. RFLA [48] measures the similarity between the Gaussian receptive filed and ground truth to alleviate the problem of insufficient positive samples for tiny objects.

Scale-aware methods. FPN [25] constructs feature-level pyramid for multi-scale learning, promoting promising improvement of tiny objects. BiFPN [38] employs a weighted bi-directional feature pyramid architecture to facilitate multi-scale fusion. DetectoRS [31] further develops it by incorporating extra feedback connections and switchable atrous convolution. Gong *et al.* [12] introduces fusion factor to control the information flow between two adjacent layers of FPN, easing the burden of learning for tiny objects.

Attention-based frameworks. Attention mechanisms are used to filter out unimportant regions and increase attention to tiny objects. KB-RANN [52] utilizes attention mechanism to refine key features in an iterative way. SCRDet [50] introduces channel attention and pixel attention network to jointly explore tiny objects. AFF-SSD [29] designs a dual-path attention module to screen feature information, improving the detection performance of one-stage methods.

Mimic learning algorithms. The core of mimic learning is to use high-quality features of large instances to boost low-quality representations of tiny objects. Perceptual GAN [21] and MT-GAN [1] draw inspiration from generative adversarial networks [13] and generate super-resolved representations for tiny objects. Another line of efforts [16, 46, 53] aim to narrow the distance between large and small instances in feature space with the help of similarity measurement.

Previous methods often overlook the performance limitations caused by weak representations suffering from information loss. The recent work, SR-TOD [6], uses a difference map created by subtracting the restored image from the original to identify regions affected by information loss.

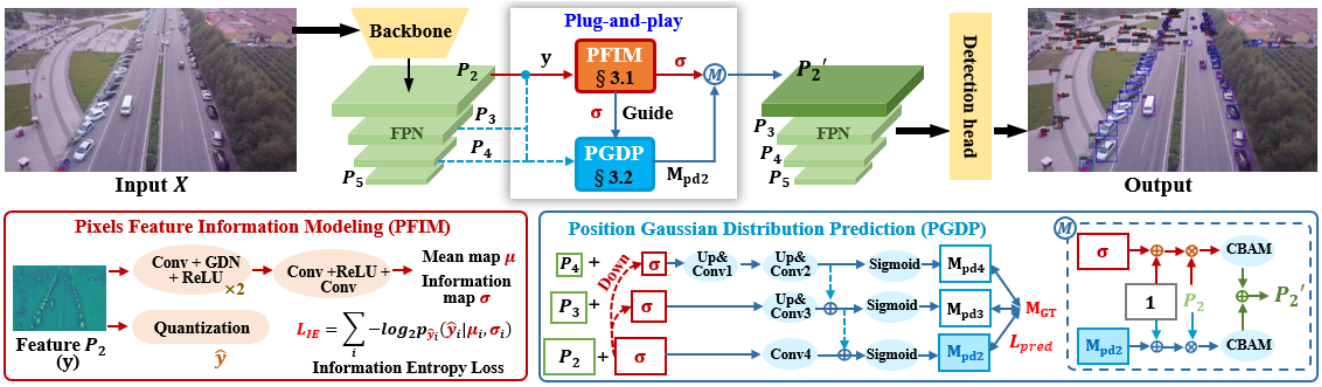

Figure 2. Method overview. GDN represents [2] Generalized Divisive Normalization which is suitable for density modeling, CBAM represents Convolutional Block Attention Module [45].

However, it depends heavily on the quality of the restored image, and its downsampling process further reduces difference map information. In contrast, supported by information entropy theory and distribution map modeling, our method directly identifies regions of information loss at the feature map level, demonstrating superior performance.

3. Methodology

Overview. As illustrated in Fig. 2, given input X, a feature pyramid is constructed from P_2 to P_5. Enhancement is applied to P_2, which primarily handles tiny objects. First, P_2 is quantized to estimate its information map σ, with salient objects having more amount of information. Then, P_2 to P_4 are used to predict the Position Gaussian Distribution Map M_{pd_2}, guided by prior knowledge σ, simultaneously assisting σ with more attention to tiny objects. After σ and M_{pd_2} enhance the information loss regions of P_2, they will respectively flow through attention module and finally merge to obtain $P_2^{'}$. As a substitute for P_2, $P_2^{'}$ will be sent to the detection head for detection task.

3.1. Pixels Feature Information Modeling

The downsampling process in neural networks results in the loss of target information as network depth increases, which is particularly detrimental for tiny objects due to their limited pixel occupancy. Enhancing the information loss regions is therefore essential for improving tiny object detection. To address this, we emphasize salient regions requiring enhancement from a novel perspective of the amount of information. Specifically, for a signal x with probability $p(x)$, amount of information is defined as follows [41]:

$$I(x) = -\log_2 p(x) \tag{1}$$

Assuming the signal feature representation is highly ordered and compact, salient and informative fragments within the feature typically have lower occurrence probabilities. According to Shannon entropy theory [37, 39], the amount of a signal feature information is closely linked to its distribution. For a stream of discretized elements $\hat{y}$, the cross entropy between the actual marginal distribution $m(\hat{y})$ and its approximation $p_{\hat{y}}(\hat{y})$ reflects the lower bound of the rate cost (commonly measured in bits in communication systems) for encoding $\hat{y}$:

$$R = \mathbb{E}_{\hat{y}\sim m}[-\log_2 p_{\hat{y}}(\hat{y})] = H(m) + D_{KL}(m||p_{\hat{y}}) \tag{2}$$

where $H(\cdot)$ denotes the entropy function, and $D_{KL}(\cdot)$ represents the KL-divergence function. When the modeling $p_{\hat{y}}(\hat{y})$ perfectly matches the marginal $m(\hat{y})$, the encoding cost R is minimized, as the mean amount of information is reduced to its lowest. In communication systems, to achieve minimum encoding cost for storage and transmission, more encoding resources are allocated to the salient and informative fragments with more amount of information, while fragments with higher occurrence probabilities incur lower encoding costs due to their less information.

Given an input image $X \in \mathbb{R}^{H\times W\times 3}$ with the bottom-level feature $P_2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ (denoted as y) from the FPN for tiny object detection, pixels containing more amount of information in y often correspond to salient regions, as the occurrence probability of salient regions is typically lower than that of smooth background. Thus, we leverage the amount of pixels information to identify and enhance these salient regions for subsequent detection. By minimizing the overall encoding cost based on a well-estimated distribution, the network adaptively allocates more cost to salient regions, while smooth background regions incur less cost. This overall encoding cost optimization process effectively mines spatial structures and reduces spatial redundancy.

We first estimate the distribution of y. To encode the feature y with a finite number of encoding bits, the continuous feature variables are discretized using quantization Q, resulting in the discrete feature map $\hat{y}$. This quantization reduces pixel-wise redundancy to some extent. To make the quantization function differentiable, additive uniform noise $\mathcal{U}(-\frac{1}{2},\frac{1}{2})$ is applied during training [2]:

$$\hat{y} = Q(y) = y + \mathcal{U}(-\frac{1}{2},\frac{1}{2}) \tag{3}$$

The distribution of $\hat{y}$ is estimated with a fully factorized density model, where each pixel element $\hat{y}_i$ is independently modeled as a Gaussian distribution with mean μ_i and standard deviation σ_i:

$$p_{\hat{y}}(\hat{y}|\mu,\sigma) = \prod_i (\mathcal{N}(\mu_i, \sigma_i^2) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2}))(\hat{y}_i) \quad (4)$$

where the mean map $\mu \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ and scale map $\sigma \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ are predicted from the bottom-level feature map y using a CNN parameters stimation module, as shown in Fig. 2. Due to the ability of CNN in local receptive field within Gaussian distribution prediction, our pixel-level modeling is equipped with local structure information. We convolve the Gaussian density with a unit uniform distribution, which has been shown to be effective in matching the estimated distribution with the unknown actual marginal distribution [30].

After determining the encoding distribution, the likelihood of each element $\hat{y}_i$ can be obtained via the binned area under the probability density function:

$$\begin{aligned} p_{\hat{y}_i}(\hat{y}_i|\mu_i,\sigma_i) &= (\mathcal{N}(\mu_i,\sigma_i^2) * \mathcal{U}(-\frac{1}{2},\frac{1}{2}))(\hat{y}_i) \\ &= \int_{\hat{y}_i-\frac{1}{2}}^{\hat{y}_i+\frac{1}{2}} \mathcal{N}(y|\mu_i,\sigma_i^2)\, dy \\ &= F(\frac{\hat{y}_i+\frac{1}{2}-\mu_i}{\sigma_i}) - F(\frac{\hat{y}_i-\frac{1}{2}-\mu_i}{\sigma_i}) \end{aligned} \quad (5)$$

where F denotes the cumulative distribution function of a standard normal Gaussian distribution.

Finally, the encoding cost of the element $\hat{y}_i$ is calculated by the negative log-likelihood:

$$R_{\hat{y}_i} = -\log_2 p_{\hat{y}_i}(\hat{y}_i|\mu_i,\sigma_i) \quad (6)$$

The Information Entropy loss $\mathcal{L}_{IE}$ is defined as the sum of the encoding cost of all elements, which is a reflection of mean amount of feature information:

$$\mathcal{L}_{IE} = \sum_i R_{\hat{y}_i} \quad (7)$$

We minimize the Information Entropy loss $\mathcal{L}_{IE}$ to adaptively capture regions with more amount of information, such as tiny objects. As shown in Fig. 3(a), the predicted mean map μ is typically close to $\hat{y}$, minimizing the Information Entropy loss. For smooth backgrounds, the output likelihood is higher than that for salient object regions, indicating that the latter contain more amount of information and incur higher encoding costs. Salient regions output larger σ, requiring more bits to encode. In contrast, background regions have smaller σ, resulting in fewer encoding bits and less randomness. Thus, the predicted scale map σ is positively correlated with the amount of information map (calculated by Eq. (6)) and highlights potential salient regions

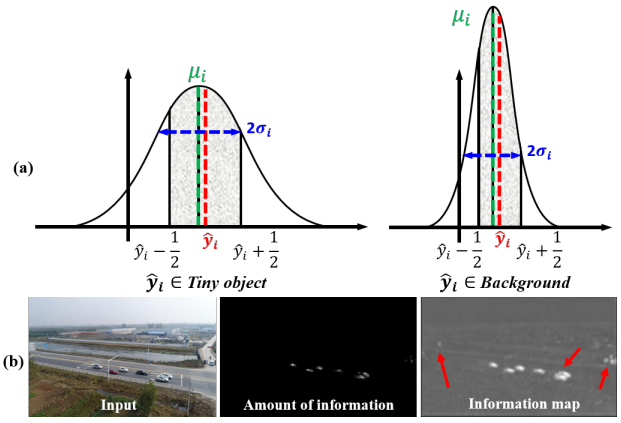

Figure 3. (a) Gaussian distributions of tiny object and background. (b) Amount of information map & information map σ comparison.

for enhancement. While the scale map σ is more visually salient, since the amount of information map is the direct optimization target (see Fig. 3(b)). Therefore, we treat the scale map σ as a better representation of the amount of information (denoted as information map σ by default in this paper) and refine the feature map P_2 by σ to enhance salient regions, as formulated below:

$$\sigma = Mean(\sigma), \quad y_1 = y \otimes (1+\sigma) \quad (8)$$

where $Mean$ denotes the averaging operation along the channel dimension, and $\otimes$ represents pixel-wise multiplication. Note that $1+\sigma$ is used to preserve useful contextual information in the feature map P_2, preventing it from being influenced by values near zero in σ.

3.2. Position Gaussian Distribution Prediction

Position Gaussian Distribution Map. To effectively make the information map pinpoint tiny objects, we introduce Position Gaussian Distribution Map at the region level. Ground truth distribution map for tiny object detection follows two principles: 1) foreground-background distinction (foreground has higher values), and 2) differentiation between tiny and general objects (tiny objects have relatively higher values).

We employ a Gaussian Mixture distribution to model the feature map and emphasize object locations, as shown in Fig. 4. Given input X with N object instances, we set the size of the generated ground truth map to match P_2 dimensions ($\frac{H}{4} \times \frac{W}{4} \times 1$) to reduce computational cost. Assuming that an object's main body is centered in its annotated bounding box [43, 48], each Gaussian component in the feature map distribution is defined as a two-dimensional Gaussian distribution with $(\boldsymbol{\mu_i^{box}}, \boldsymbol{\Sigma_i^{box}})$, where the bounding box center serves as the mean vector, and the squared side length forms the covariance matrix:

$$\boldsymbol{\mu_i^{box}} = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \boldsymbol{\Sigma_i^{box}} = \begin{bmatrix} w_i^2 & 0 \\ 0 & h_i^2 \end{bmatrix} \quad (9)$$

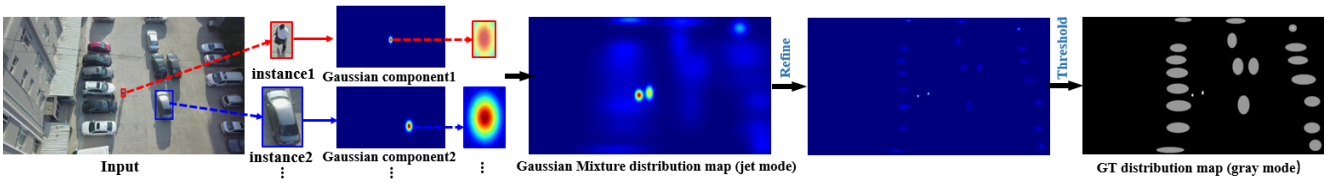

Figure 4. Generation pipeline of Position Gaussian Distribution Map. GT is displayed in gray mode for clarity.

where (x_i, y_i, w_i, h_i) represents the location of the i-th bounding box in P_2. As shown in the Gaussian Mixture distribution map in Fig. 4, regions with dense, large objects exhibit chaotic salient characteristics, while tiny objects display significantly higher peak values than general objects, introducing noise and increasing the learning burden. To create a distinct distribution map for easier training, we refine the covariance matrix with scaling factors α_i based on the area of each annotated box. Following the object definitions in [44], scaling factors are set to 4, 6, 8, and 10 for very tiny (2–8 pixels), tiny (8–16 pixels), small (16–32 pixels), and general objects, respectively, as formulated below:

$$\boldsymbol{\mu_i^{box}} = \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \boldsymbol{\Sigma_i^{box}} = \begin{bmatrix} (\frac{w_i}{\alpha_i})^2 & 0 \\ 0 & (\frac{h_i}{\alpha_i})^2 \end{bmatrix} \tag{10}$$

The final Gaussian Mixture distribution of the feature map P_2 is derived by combining each Gaussian component:

$$f(p) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{N}(p|\boldsymbol{\mu_i^{box}}, \boldsymbol{\Sigma_i^{box}}) \tag{11}$$

where p denotes the pixel position in P_2. To further accentuate the foreground-background contrast, we multiply $f(p)$ by N and post-process the distribution map using a threshold th (positions exceeding th are increased by 0.5) to obtain the final ground truth position Gaussian distribution map M_{GT}. This process is simplified as follows:

$$th = \frac{1}{|S|} \sum_{p \in S} N \cdot f(p)$$
$$M_{GT} = (Sign(N \cdot f(p) - th) + 1) \times 0.25 + N \cdot f(p) \tag{12}$$

where S is the set of all pixel positions in the feature map and $Sign$ is the mathematical Sign function.

As shown in Fig. 4, the designed position Gaussian distribution map effectively captures image saliency, with dense tiny objects assigned relatively higher values. This map thus serves as a guide to enhance the representations of tiny objects that are otherwise weakly activated.

Prediction. We predict the position Gaussian distribution map from multi-scale features under supervision, guided by the information map σ derived from the pixels feature information modeling module. The prediction process guided by σ can simultaneously modulate the information map and distribution map to spotlight tiny targets, both providing attention regions for enhancing weak representations. Given the distinct semantic and spatial details across FPN levels, we use P_2, P_3 and P_4 to predict the map. The overall prediction network structure is shown in Fig. 2, with further details in the supplementary material. Specifically, incorporating prior amount of information from σ, we add information map σ to each feature level as inputs for the prediction network, where σ is downsampled accordingly. The inputs are given as $inputs = [P_4 + \frac{1}{4}\sigma, P_3 + \frac{1}{2}\sigma, P_2 + \sigma]$, with $\frac{1}{4}$ and $\frac{1}{2}$ representing downsampling. A series of convolutional and transposed convolutional layers [54] (except in the P_2 branch) perform feature extraction and learnable upsampling, respectively. Notably, we merge features via skip connections from deeper side-outputs (pre-Sigmoid) to shallower layers to aid in locating salient regions and refining complex prediction maps [14]. Each branch then predicts the position Gaussian distribution map, yielding M_{pd_2}, M_{pd_3} and M_{pd_4}. The process is simplified as: $[M_{pd_4}, M_{pd_3}, M_{pd_2}] = \phi(inputs)$, where ϕ represents the parameterized multi-scale prediction network.

We apply deep supervision for the three side-output predictions which are the same size as ground truth position Gaussian distribution map:

$$\mathcal{L}_{pred} = \sum_{i=2}^{4} MSE_{weighted}(M_{pd_i}, M_{GT}) \tag{13}$$

where $MSE_{weighted}$ denotes the weighted Mean Squared Error loss. To address foreground-background imbalance, we set the weight for object regions exceeding the threshold th to 10 (and 0.1 for the background) to emphasize positive pixels.

In general, incorporating the information map σ allows feature information to effectively modulate distribution map predictions, while optimizing $\mathcal{L}_{pred}$ aids in generating a better information map σ that identifies information-rich tiny object regions.

Similar to information map, M_{pd_2} is used to enhance tiny object representations:

$$y_2 = y \otimes (1 + M_{pd_2}) \tag{14}$$

The enhanced features y_1 and y_2 are fed into the Convolutional Block Attention Module (CBAM) [45] for further attention to important regions, enabling global exploration of the features. The two attentive features are fused by element-wise addition to form the enhanced feature map $P_2^{'}$. $P_2^{'}$, which replaces P_2 in the new feature pyramid will be sent to the detection head for downstream tasks.

Methods	Venue	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s
Faster R-CNN [34]	TPAMI'2017	23.9	42.2	23.8	0.1	6.5	21.1
Cascade R-CNN [5]	CVPR'2018	25.2	42.6	25.9	0.1	7.0	22.5
DetectoRS [31]	CVPR'2021	26.3	43.9	26.9	0.1	7.5	23.3
NWD-RKA [47]	ISPRS'2022	27.2	47.7	26.8	4.1	12.3	23.7
RFLA [48]	ECCV'2022	27.2	48.0	26.6	4.5	13.0	23.6
CEASC [10]	CVPR'2023	25.1	42.6	25.7	2.0	8.6	20.7
PKS R-CNN [4]	CVPR'2024	24.3	42.4	24.4	0.1	7.3	22.5
Salience DETR [15]	CVPR'2024	28.4	48.7	28.5	5.4	13.7	25.0
SR-TOD [6]	ECCV'2024	27.3	46.9	27.5	2.3	11.5	24.7
Faster R-CNN [34] w/ ours	-	26.8$_{+2.9}$	47.8$_{+5.6}$	26.6$_{+2.8}$	2.6$_{+2.5}$	12.3$_{+5.8}$	25.2$_{+4.1}$
Cascade R-CNN [5] w/ ours	-	28.1$_{+2.9}$	48.1$_{+5.5}$	28.3$_{+2.4}$	2.9$_{+2.8}$	12.3$_{+5.3}$	26.4$_{+3.9}$
DetectoRS [31] w/ ours	-	28.3$_{+2.0}$	48.5$_{+4.6}$	28.8$_{+1.9}$	3.5$_{+3.4}$	12.6$_{+5.1}$	26.1$_{+2.8}$
RFLA [48] w/ ours	-	**29.0**$_{+1.8}$	**50.7**$_{+2.7}$	**29.0**$_{+2.4}$	**7.4**$_{+2.9}$	**14.9**$_{+1.9}$	**26.5**$_{+2.9}$

Table 1. Experimental results on VisDrone2019. The bottom-right corner of the result represents the gain towards its baseline.

3.3. Loss Function

Without loss of generality, we unify the original detection network loss as $\mathcal{L}_{det}$, comprising both regression and classification losses. The total loss function is expressed as:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda_1 \mathcal{L}_{IE} + \lambda_2 \mathcal{L}_{pred} \tag{15}$$

where λ_1 and λ_2 are balancing hyper-parameters.

4. Experiments

4.1. Evaluation setup

Datasets. Experiments are conducted on three tiny object detection datasets: VisDrone2019 [11], AI-TOD [44], and AI-TODv2 [47]. VisDrone2019 consists of 10 categories and 10,209 high-resolution (2,000 × 1,500) drone images captured in diverse urban scenes from various angles, featuring numerous small object instances. AI-TOD contains 28,036 aerial images and 700,621 annotated instances across 8 categories, with a mean instance size of 12.8 pixels. AI-TODv2, a refined version, includes 752,745 instances and a reduced mean instance size of 12.7 pixels.

Implementation Details. Experiments are implemented in PyTorch using MMDetection [7], with all models trained on a single NVIDIA RTX 4090 GPU. We follow the experimental settings of SR-TOD [6]. By default, ResNet50-FPN [25] pretrained on ImageNet [36] is used for feature extraction. Models are optimized using the SGD optimizer with 0.9 momentum, 0.0001 weight decay, and 2 batch size. Training consists of 12 epochs, with learning rate of 0.005 that decays at the 8th and 11th epochs. λ_1 and λ_2 are set to 0.01 and 1.0, respectively. Ablation experiments and analysis are conducted on VisDrone2019 using DetectoRS [31].

Evaluation Metrics. We follow the AI-TOD benchmark [44] evaluation metrics, which comprise the Average Precision (AP), $AP_{0.5}$, $AP_{0.75}$, AP_{vt} (very tiny), AP_t (tiny) and AP_s (small).

4.2. Comparison with State-of-the-art Methods

Comparisons on VisDrone2019. We compare our method with state-of-the-art approaches on VisDrone2019, as shown in Tab. 1, leading to several key observations: 1) Integrating our method into baseline detectors as plug-and-play components consistently improves detection performance, with the most notable gain of 5.8 points in AP_t for Faster R-CNN. On the challenging AP_{vt} metric, the gain reaches 3.4 points, highlighting our method's advantage for tiny objects. 2) When combined with RFLA, our method outperforms all existing approaches on all metrics, achieving the highest AP_{vt} of 7.4, 2 points higher than the next best competitor. 3) Similar to our approach, SR-TOD uses a difference map for feature enhancement. However, our method achieves a greater overall improvement, as the difference map only captures partial information loss regions, proving the superiority of our method in identifying information loss regions. 4) RFLA addresses tiny object detection by optimizing label assignment, and our results show that improvements in both label assignment and feature representation significantly boost performance, suggesting the potential of combining these strategies in future research.

Methods	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s
Fas.R-CNN [34]	11.7	27.4	8.2	0.0	8.6	23.7
Cas.R-CNN [5]	14.0	31.2	10.7	0.1	10.3	26.2
DetectoRS [31]	14.6	31.8	11.5	0.0	11.0	27.4
NWD-RKA [47]	20.5	48.7	13.8	8.1	20.6	25.6
RFLA [48]	21.7	50.5	15.3	8.3	21.8	26.3
CEASC [10]	14.4	33.2	10.2	3.2	15.2	18.5
PKS R-CNN [4]	9.3	22.1	6.3	0.0	6.6	18.9
Sal. DETR [15]	19.7	48.4	12.7	7.4	19.7	25.9
SR-TOD [6]	21.9	50.6	15.6	**9.6**	22.4	26.7
[34] w/ ours	20.6$_{+8.9}$	49.7$_{+22.3}$	13.5$_{+5.3}$	6.6$_{+6.6}$	22.1$_{+13.5}$	25.7$_{+2.0}$
[5] w/ ours	22.6$_{+8.6}$	52.1$_{+20.9}$	16.4$_{+5.7}$	8.4$_{+8.3}$	23.5$_{+13.2}$	27.6$_{+1.4}$
[31] w/ ours	**24.3**$_{+9.7}$	**54.4**$_{+22.6}$	**18.3**$_{+6.8}$	8.5$_{+8.5}$	**24.9**$_{+13.9}$	**29.8**$_{+2.4}$
[48] w/ ours	22.6$_{+0.9}$	52.7$_{+2.2}$	15.9$_{+0.6}$	8.2$_{-0.1}$	22.7$_{+0.9}$	27.2$_{+0.9}$

Table 2. Experimental results on AI-TOD.

Comparisons on AI-TOD and AI-TODv2. We also evaluate our method on AI-TOD and AI-TODv2 to assess its generalization. Results on AI-TOD, shown in Tab. 2, indicate that nearly all models outperform the baseline, with a maximum gain of 22.6 points. Our method, integrated with DetectoRS, achieves the best performance on all metrics except AP_{vt}, where it ranks second. Similarly, as shown in Tab. 3, results on AI-TODv2 mirror those on AI-TOD,

Methods	AP	$AP_{0.5}$	$AP_{0.75}$	AP_{vt}	AP_t	AP_s
Cas.R-CNN [5]	14.9	33.4	11.0	0.1	11.0	26.7
DetectoRS [31]	16.1	35.5	12.5	0.1	12.6	28.3
NWD-RKA [47]	22.2	52.5	15.1	7.8	21.8	28.0
RFLA [48]	22.8	53.8	15.4	7.9	22.2	28.8
CEASC [10]	17.0	39.5	11.5	3.9	16.1	22.0
PKS R-CNN [4]	11.1	25.1	8.5	0.0	7.3	22.1
Sal. DETR [15]	20.9	52.6	12.6	8.2	20.4	27.0
SR-TOD [6]	22.9	54.0	15.7	7.7	22.9	28.4
[5] w/ ours	23.7$_{+8.8}$	55.3$_{+21.9}$	16.8$_{+5.8}$	7.4$_{+7.3}$	23.6$_{+12.6}$	29.1$_{+2.4}$
[31] w/ ours	**25.5**$_{+9.4}$	**58.2**$_{+22.7}$	**18.4**$_{+5.9}$	**8.6**$_{+8.5}$	**25.7**$_{+13.1}$	**30.6**$_{+2.3}$
[48] w/ ours	23.9$_{+1.1}$	55.8$_{+2.0}$	16.6$_{+1.2}$	7.3$_{-0.6}$	23.5$_{+1.3}$	29.0$_{+0.2}$

Table 3. Experimental results on AI-TODv2.

PFIM	PGDP	AP	$AP_{0.5}$	AP_{vt}	AP_t	AP_s
		26.3	43.9	0.1	7.5	23.3
✓		28.2	48.4	3.3	12.2	26.0
	✓	27.6	47.3	3.4	11.2	24.6
✓	✓	**28.3**	**48.5**	**3.5**	**12.6**	**26.1**

Table 4. Effect of our proposed modules.

with our best method significantly outperforming competitors across all metrics. These experiments further validate the effectiveness of our approach for tiny object detection.

4.3. Ablation Study

Proposed modules. We validate the effectiveness of our two proposed components (Pixels Feature Information Modeling and Position Gaussian Distribution Prediction) by progressively applying them to enhance information loss regions. As shown in Tab. 4, each component improves performance relative to the baseline, with further gains observed when both are combined. This demonstrates that the components complement each other in capturing information loss regions to enhance feature representation.

Different distribution modeling methods. We investigate various strategies for modeling the distribution map. Specifically, we refine the covariance matrix using fixed scaling factor α=1, in contrast to the instance-size-dependent scaling factors used in our method. Additionally, we construct a binary mask based on ground truth bounding box, where background values are 0 and the bounding box is 1. We also utilize self-attention to generate a weight distribution map. As shown in Tab. 5, all approaches underperform compared to our method. The fixed scaling factor results in an inflated peak for tiny targets, increasing the network's learning burden. The binary mask approach is overly simplistic and introduces irrelevant background noise, leading to suboptimal performance. Weight distribution map generated from self-attention at limited pixels is susceptible to interference from irrelevant background, resulting in performance degradation for tiny objects. In contrast, our smooth distribution map effectively highlights salient features without noise, yielding superior results.

Different prior guidance. As shown in Tab. 6, we use information map σ to guide the prediction of the distribution map through various methods: multiplying input features by σ, by $(1+\sigma)$, and concatenating input features with σ. Although multiplying by σ yields the best performance on both AP and AP_s, other metrics show significant declines, primarily due to the loss of useful context in regions where σ is near 0. The concatenation method introduces noise, as pyramid features (which capture spatial and semantic information) and σ (which represents the amount of information) have different meanings. Multiplying by $(1+\sigma)$ performs similarly to ours, and we adopt element-wise addition for its simplicity.

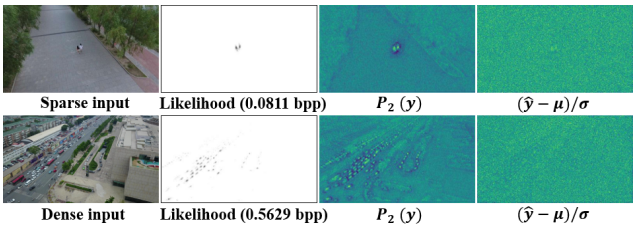

Figure 5. Visualizations of sparse and dense scenes, and columns show the likelihood map (corresponding bits per pixel), feature map P_2 and normalized feature map $\frac{\hat{y}-\mu}{\sigma}$.

Different fusion strategies. We explore various fusion strategies for two enhanced features, y_1 and y_2, including element-wise multiplication and concatenation. As shown in Tab. 7, element-wise addition proves to be the most effective fusion method, as it complements the features and prevents information loss. In contrast, both element-wise multiplication and concatenation lead to performance degradation. Element-wise multiplication causes nonlinear changes by magnifying or shrinking feature values, resulting in critical information loss. Additionally, since the two enhanced features are highly similar, concatenation introduces redundancy and increases network complexity.

4.4. Further Analysis

Analysis on Pixels Feature Information Modeling. We analyze the effectiveness of Pixels Feature Information Modeling module in capturing salient objects with high amount of information. First, we visualize the estimated information map σ in Fig. 6(a), which clearly captures the target's spatial structure. The spatial structure of the objects becomes more prominent due to their higher amount of information, confirming the effectiveness of using information map σ to enhance the feature. To illustrate how Pixels Feature Information Modeling module captures spatial structure through different Gaussian distributions, we visualize the feature map P_2 (denoted as y) and its normalized version (calculated by $\frac{\hat{y}-\mu}{\sigma}$) in Fig. 5. Since each element of the normalized feature map follows a standard Gaussian distribution, the structure appears blurry and disordered, indicating that feature redundancy is removed and the spatial structure is effectively captured by the modeled distribution.

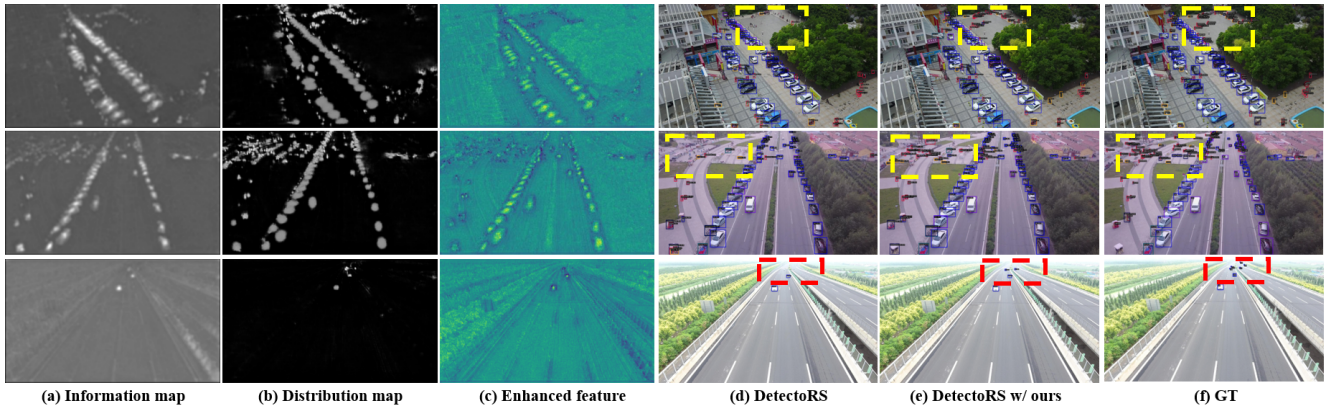

Figure 6. Visualization of qualitative results. (a)-(c) are estimated information map, predicted distribution map and final enhanced feature, (d)-(f) are the detection results of DetectoRS [31], our method and GT, dotted boxes are drawn for better comparison.

Settings	AP	$AP_{0.5}$	AP_{vt}	AP_t	AP_s
fixed α=1	27.9	47.9	3.4	11.8	25.6
binary mask	27.9	47.6	3.4	11.8	25.1
self-attention	27.7	47.5	3.0	11.7	25.1
Ours	**28.3**	**48.5**	**3.5**	**12.6**	**26.1**

Table 5. Different distribution modeling methods.

Settings	AP	$AP_{0.5}$	AP_{vt}	AP_t	AP_s
$\otimes\,\sigma$	**28.3**	48.2	2.7	12.2	**26.1**
$\otimes\,(1+\sigma)$	28.1	48.4	**3.7**	12.5	25.9
concat	27.9	47.8	2.7	12.2	25.4
Ours	**28.3**	**48.5**	3.5	**12.6**	**26.1**

Table 6. Different amount of information guidances. Element-wise product $\otimes$.

Settings	AP	$AP_{0.5}$	AP_{vt}	AP_t	AP_s
multiplication	27.5	47.0	2.5	11.4	24.4
concat	27.8	47.4	2.7	11.3	24.8
Ours	**28.3**	**48.5**	**3.5**	**12.6**	**26.1**

Table 7. Effect of different fusion strategies.

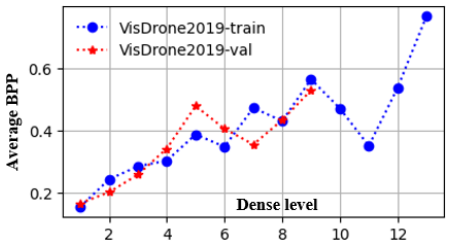

Figure 7. Average BPP (bits per pixel) curves corresponding to different dense levels on VisDrone2019 training and validation set.

We adopt the concept of bits per pixel (bpp) from communication systems [37] to better understand the role of Information Entropy loss in the module. Bpp is defined as the average encoding cost per pixel, i.e., the Information Entropy loss $\mathcal{L}_{IE}$ divided by the total number of pixels. As shown in Fig. 5, the value at tiny objects in the likelihood map is much lower than the background, indicating higher encoding cost (see Eq. (6)) and more information. Consequently, dense scenes exhibit higher bpp than sparse ones, with values of 0.5629 and 0.0811, respectively. We define dense level based on the number of instances in the image, increasing one level for every 40 instances. The average bpp across each density interval is plotted in Fig. 7, showing an overall increasing trend, consistent with our analysis. This supports the effectiveness of Information Entropy loss in capturing salient objects with high amount of information.

Analysis on Position Gaussian Distribution Map. We visualize the predicted distribution map in Fig. 6(b), where the objects and background are clearly separated, with the value of tiny objects being notably larger than that of general objects. Both the information map and distribution map highlight the regions to be enhanced, guiding feature improvement. As a result, the enhanced feature $P_2^{'}$ in Fig. 6(c) becomes more distinct. Furthermore, as shown in Fig. 6(d)-(f), our method outperforms the baseline generic detector by detecting more challenging tiny objects within the dotted box, demonstrating the effectiveness of our feature enhancement for tiny objects.

5. Conclusion

This paper introduces a novel approach to enhance weak regions in tiny object detection based on pixel-level feature information. We minimize Information Entropy loss in an unsupervised manner to generate the attentive information map, where higher values correspond to salient regions with more amount of information. Then, to make the information map accentuate tiny objects, we predict the Position Gaussian Distribution Map guided by information map supervisedly. Using these two attentive maps, we enhance previously indiscriminative tiny object features, and the experimental results on three datasets validate our method's effectiveness.

6. Acknowledgments

This work was supported in part by Science and Technology Project of Jiangxi province (20232ACC01007), in part by Ji'an Science and Technology Project (20233TGV06020), in part by the National Natural Science Foundation of China under Grant (62373293, 62463020, 62403189), in part by Jiangxi Provincial Natural Science Foundation (20242BAB20050), and in part by Ji'an Science and Technology Plan Natural Science Foundation (20244018591).

References

[1] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 206–221, 2018. 2

[2] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. Density modeling of images using a generalized normalization transformation. *arXiv preprint arXiv:1511.06281*, 2015. 3

[3] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2012. 1

[4] Xinhao Cai, Qiuxia Lai, Yuwei Wang, Wenguan Wang, Zeren Sun, and Yazhou Yao. Poly kernel inception network for remote sensing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27706–27716, 2024. 6, 7

[5] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2, 6, 7

[6] Bing Cao, Haiyu Yao, Pengfei Zhu, and Qinghua Hu. Visible and clear: Finding tiny objects in difference map. *arXiv preprint arXiv:2405.11276*, 2024. 2, 6, 7

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[8] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[10] Bowei Du, Yecheng Huang, Jiaxin Chen, and Di Huang. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13435–13444, 2023. 6, 7

[11] Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue Zhang, et al. Visdrone-det2019: The vision meets drone object detection in image challenge results. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 6

[12] Yuqi Gong, Xuehui Yu, Yao Ding, Xiaoke Peng, Jian Zhao, and Zhenjun Han. Effective fusion factor in fpn for tiny object detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1160–1168, 2021. 1, 2

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[14] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3203–3212, 2017. 5

[15] Xiuquan Hou, Meiqin Liu, Senlin Zhang, Ping Wei, and Badong Chen. Salience detr: Enhancing detection transformer with hierarchical salience filtering refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17574–17583, 2024. 6, 7

[16] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Robust small-scale pedestrian detection with cued recall via memory learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3050–3059, 2021. 2

[17] Mate Kisantal. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019. 2

[18] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 2

[19] Satish Kumar, Bowen Zhang, Chandrakanth Gudavalli, Connor Levenson, Lacey Hughey, Jared A Stabach, Irene Amoke, Gordon Ojwang, Joseph Mukeka, Stephen Mwiu, et al. Wildlifemapper: Aerial image analysis for multi-species detection and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12594–12604, 2024. 1

[20] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[21] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1222–1230, 2017. 2

[22] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6054–6063, 2019. 1

[23] Yangyang Li, Qin Huang, Xuan Pei, Yanqiao Chen, Licheng Jiao, and Ronghua Shang. Cross-layer attention network for small object detection in remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:2148–2161, 2020. 1

[24] Cheng-Jian Lin and Jyun-Yu Jhang. Intelligent traffic-monitoring system based on yolo and convolutional fuzzy neural networks. *IEEE Access*, 10:14120–14133, 2022. 1

[25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1, 2, 6

[26] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 1

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 1, 2

[28] Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7363–7372, 2019. 2

[29] Xiaocong Lu, Jian Ji, Zhiqi Xing, and Qiguang Miao. Attention and feature fusion ssd for remote sensing object detection. *IEEE Transactions on Instrumentation and Measurement*, 70:1–9, 2021. 1, 2

[30] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. *Advances in neural information processing systems*, 31, 2018. 4

[31] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10213–10224, 2021. 1, 2, 6, 7, 8

[32] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 1

[33] Joseph Redmon. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 2

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 1, 2, 6

[35] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 2

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 6

[37] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2, 3, 8

[38] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 1, 2

[39] MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006. 3

[40] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *Proc. Int. Conf. Computer Vision (ICCV)*, 2019. 2

[41] Hu Kuo Ting. On the amount of information. *Theory of Probability & Its Applications*, 7(4):439–447, 1962. 3

[42] Thang Vu, Hyunjun Jang, Trung X Pham, and Chang Yoo. Cascade rpn: Delving into high-quality region proposal network with adaptive convolution. *Advances in neural information processing systems*, 32, 2019. 2

[43] Jinwang Wang, Wen Yang, Heng-Chao Li, Haijian Zhang, and Gui-Song Xia. Learning center probability map for detecting objects in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4307–4323, 2020. 4

[44] Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-Song Xia. Tiny object detection in aerial images. In *2020 25th international conference on pattern recognition (ICPR)*, pages 3791–3798. IEEE, 2021. 1, 5, 6

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3, 5

[46] Jialian Wu, Chunluan Zhou, Qian Zhang, Ming Yang, and Junsong Yuan. Self-mimic learning for small-scale pedestrian detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2012–2020, 2020. 2

[47] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190: 79–93, 2022. 2, 6, 7

[48] Chang Xu, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. Rfla: Gaussian receptive field based label assignment for tiny object detection. In *European conference on computer vision*, pages 526–543. Springer, 2022. 2, 4, 6, 7

[49] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 13668–13677, 2022. 1

[50] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8232–8241, 2019. 1, 2

[51] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9657–9666, 2019. 2

[52] Kai Yi, Zhiqiang Jian, Shitao Chen, and Nanning Zheng. Feature selective small object detection via knowledge-based recurrent attentive neural network. *arXiv preprint arXiv:1803.05263*, 2018. 1, 2

[53] Xiang Yuan, Gong Cheng, Kebing Yan, Qinghua Zeng, and Junwei Han. Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6317–6327, 2023. 2

[54] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 international conference on computer vision*, pages 2018–2025. IEEE, 2011. 5

[55] Fan Zhang, Licheng Jiao, Lingling Li, Fang Liu, and Xu Liu. Multiresolution attention extractor for small object detection. *arXiv preprint arXiv:2006.05941*, 2020. 1

[56] Hangyue Zhao, Hongpu Zhang, and Yanyun Zhao. Yolov7-sea: Object detection of maritime uav images based on improved yolov7. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 233–238, 2023. 1

[57] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 850–859, 2019. 2