EagleConv: Bio-inspired Dual-Foveated Convolution for Robust Small Object Detection

Jianwei Liu[1,2] **Lifei Hao**[1,2,*] **Baoqi Huang**[1,2] **Bing Jia**[1,2] **Xuandong Zhao**[1,2]

[1]College of Computer Science, Inner Mongolia University, Hohhot 010021, China
[2]Research Center for Spatiotemporal Intelligence, Inner Mongolia University
[*]Corresponding Author. Email: cshlf@imu.edu.cn

Abstract

Small object detection underpins wide-area vision tasks such as UAV and remote-sensing imagery. However, standard convolutions often exhibit low-pass smoothing behavior, which suppresses the sparse edge cues of tiny targets and may cause them to be overwhelmed by background clutter, leading to semantic information loss at the early stages of feature extraction. Inspired by the dual-fovea physiology of the eagle eye, we propose EagleConv, a bio-inspired operator that superimposes three Gaussian components into a center-excitation, surround-inhibition, peripheral-context response profile to selectively amplify small-object signals while attenuating noise. Architecturally, EagleConv adopts a sparse dual-pathway design that integrates partial-channel depthwise dual-foveated filtering, pointwise channel mixing, and adaptive residual gating, followed by anti-aliasing downsampling for robust dimensionality reduction. Experiments on three public benchmarks show that augmenting YOLOv11s with EagleConv delivers consistent and significant gains over multiple convolution-augmentation baselines, achieving state-of-the-art performance and confirming its transferability and generality as a plug-and-play module for wide-area small-object detection.

1 INTRODUCTION

The rapid growth of UAV and high-resolution Earth observation imagery makes efficient extraction of actionable targets (e.g., traffic-violating vehicles and search-and-rescue objects over water) crucial for balancing computational cost and operational efficiency Cheng et al. [2023]. This requires small-object detectors with high recall and accurate localization to reduce manual inspection. However, tiny targets (often $< 32 \times 32$ pixels) are easily overwhelmed by clutter, creating a severe semantic gap. From a signal-processing point of view, standard CNN kernels (typically Gaussian-initialized) behave as spatial low-pass filters Zhang [2019] (Fig. 1 (a)), which smooth high-frequency content and suppress noise but also blur tiny-target cues, making them difficult to distinguish from background clutter Qin et al. [2021].

Recent studies adopt Vision Transformers (ViTs) Dosovitskiy et al. [2021], whose global self-attention captures long-range context to relate tiny targets to scene cues (e.g., roads/coastlines), partially compensating for their weak appearance. However, the computational burden of ViTs limits real-time deployment on resource-constrained onboard edge devices. In contrast, lightweight operators such as PConvChen et al. [2023] and GhostConv Han et al. [2020] reduce FLOPs by exploiting channel redundancy, but often ignore the spectral characteristics of features; indiscriminate sparsification can suppress high-frequency edge cues that are crucial for tiny objects. Moreover, these approaches still process retained features with standard convolution kernels Li et al. [2017], Hao et al. [2024], whose intrinsic low-pass smoothing can cause fragile small-object signals to be overwhelmed by background noise and vanish in deeper layers Hu and Ramanan [2017]. This motivates an efficient feature-extraction paradigm that is edge-friendly yet fundamentally alleviates the low-pass limitation to preserve weak but informative signals.

The eagle visual system provides a useful prior for small-object detection. Its dual-fovea retina combines a deep fovea for long-range recognition with a shallow fovea for wide-field perception (Fig. 1(b)) Shi et al. [2025], Wu et al. [2024], Zhang et al. [2023]. Inspired by this mechanism, we propose EagleConv, a sparse bio-inspired convolution that separates features into deep- and shallow-fovea branches. The Dual-Foveated Kernel enhances edges and suppresses smooth backgrounds, while the shallow branch preserves global context. An adaptive gate then fuses both branches according to the complexity of the scene. As shown in Fig. 1(c), EagleConv forms a learnable band-pass response that im-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3922–3932.

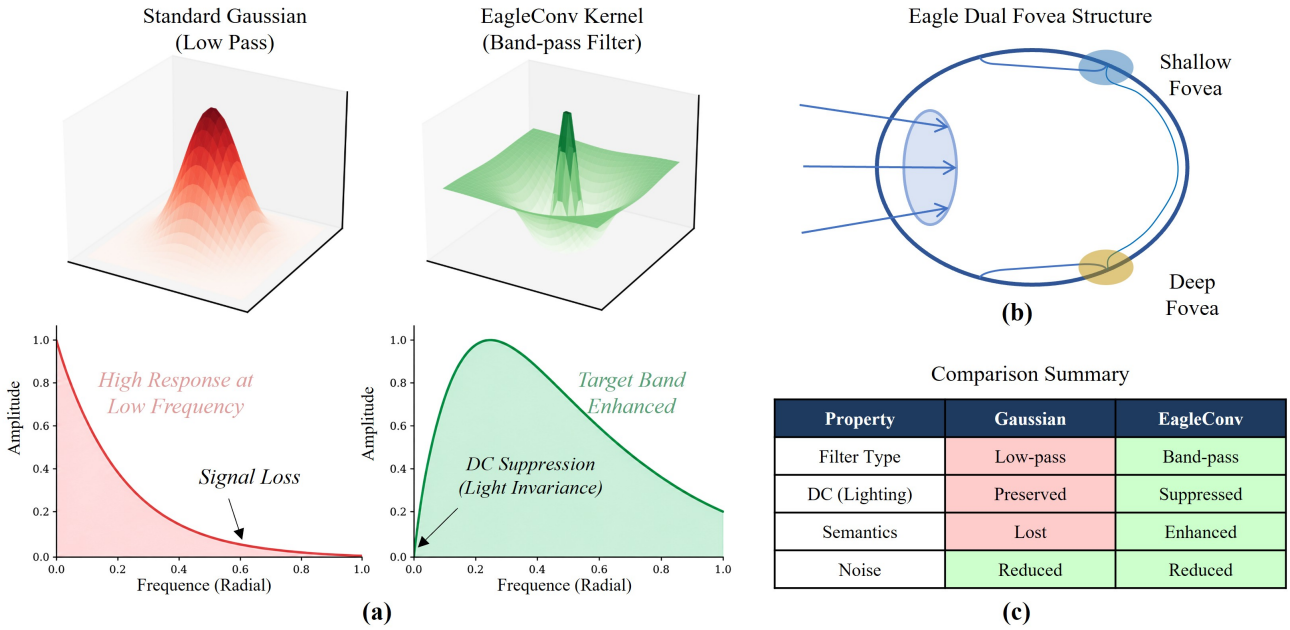


Figure 1: Bio-inspired mechanism of EagleConv and its filtering behavior versus Gaussian convolution. (a) Kernel 3D visualization and frequency responses suggest EagleConv can alleviate early semantic loss in conventional convolutions. (b) The eagle eye's dual foveae motivate multi-scale feature capture. (c) Unlike the DC-preserving, low-pass Gaussian kernel, EagleConv acts as a band-pass filter to enhance semantics and suppress illumination variations.

proves early-stage SNR and reduces the loss of tiny-object details.

Unlike existing methods, EagleConv focuses on the frequency response of convolution operators. Standard convolution tends to preserve low-frequency background and semantic information while weakening the edges and textures required for small-object recognition. EagleConv addresses this limitation with a scalable center-surround kernel that forms a band-pass response and selectively preserves target-related frequency components. In summary, our main contributions include:

- We analyze the low-pass limitation of conventional convolutions for small objects and propose a bionic, learnable band-pass kernel principle with mathematical motivation for improving SNR and mitigating semantic loss.
- We introduce a sparse dual-pathway design that orthogonally decouples high-frequency salient cues from low-frequency background flow and fuses them via adaptive gated aggregation, improving discriminability with a reduced parameter budget.

2 RELATED WORK

To facilitate a focused discussion of our core problem, we categorize the related work into three groups and elaborate as follows.

Small object detection. Detecting tiny objects, such as vehicles and pedestrians Huang et al. [2021], is challenging because targets often occupy only a few pixels, for example, fewer than 32×32 pixels, and are embedded in cluttered backgrounds such as road textures and vegetation. This results in low feature contrast and high miss rates Cheng et al. [2023]. Previous studies used feature pyramids and attention mechanisms, including sparse attention Zhu et al. [2023], DETR Carion et al. [2020], ESOD Liu et al. [2025], and DPH-YOLOv8 Wang et al. [2024]. However, these spatial-domain enhancements may not support real-time large-area deployment and often overlook spectral aliasing caused by downsampling, which can remove tiny-object cues at early stages. This motivates a high-throughput feature extractor that preserves high-frequency details for small object detection.

Biologically inspired vision and frequency-domain modeling. Biological vision addresses the trade-off between field of view and resolution through dual foveae, enabling wide-field search and high-resolution fixation. Inspired by this mechanism, BiFormer Zhu et al. [2023] introduces bi-level routing attention, while frequency-aware designs such as WTConv Finder et al. [2024] and FADC Chen et al. [2024] model spectral effects through wavelets or adaptive dilations. However, most bio-inspired methods remain heuristic and lack a signal-processing explanation of their frequency selectivity. FFT and wavelet-based approaches also introduce non-trivial transformation overhead, which

limits lightweight deployment Chi et al. [2020].

Unlike multi-scale fusion and super-resolution-assisted detectors, as well as bio-inspired methods such as Eagle-YOLO Liao et al. [2023], Eagle-eye attention Liu et al. [2022], and DeepFoveaNet Guzman-Pando and Chacon-Murguia [2021], EagleConv focuses on the convolution operator itself. Its dual-foveated three-Gaussian kernel learns a band-pass response that suppresses low-frequency background interference while preserving edges and textures of small objects. EagleConv therefore serves as a plug-and-play frequency-selective module that complements existing detection frameworks.

Lightweight convolutions and efficient architectures. Lightweight convolutions for real-time edge deployment range from DWConv Howard et al. [2017] to InceptionNeXt Yu et al. [2024] and LSConv Wang et al. [2025], which reduce parameters through operation decomposition, and PConv Chen et al. [2023], which exploits channel redundancy to reduce memory-access latency. However, PConv is computation-oriented because the retained 3×3 convolution still acts as a low-pass filter and may smooth out the weak high-frequency cues of tiny objects. In contrast, EagleConv adopts a semantics-oriented design that decouples high- and low-frequency pathways, improving efficiency while preserving semantic fidelity.

3 METHOD

To preserve tiny-target cues under limited onboard resources, we propose EagleConv, a sparse bio-inspired convolution that combines dual-fovea-style processing with frequency-selective filtering. As shown in Fig. 2, EagleConv separates features into a focused branch for fine-grained target representation and an identity branch for global context preservation. An adaptive gate dynamically fuses both pathways, enhancing tiny-object features while maintaining low computational cost.

3.1 BIOMIMETIC DUAL-FOVEATED KERNEL

This section presents the spatial-domain design of EagleConv. We analyze how the low-frequency bias of standard convolutions weakens the edge and texture cues of tiny objects, then introduce a scale-adaptive three-Gaussian kernel with zero-mean and (l_1)-normalization constraints to control Direct Current (DC) bias, amplitude, and training stability.

EagleConv abstracts the center-enhancement and surround-suppression behavior of the eagle visual system into a learnable band-pass convolution, rather than modeling the biological process precisely. Its contribution is to connect classical frequency-domain analysis and bio-inspired perception with small-object feature preservation through an interpretable convolution operator.

Spatial-Domain Representation. In small-object detection, tiny targets appear as weak, spatially localized signals embedded in low-frequency backgrounds, such as terrain and water, or high-frequency clutter, such as foliage. Standard convolutions often exhibit a bias toward low-frequency structures during optimization. Although this bias can suppress noise, it may also attenuate the high-frequency cues, including edges, textures, and local structures, that are essential for recognizing tiny objects. Consequently, discriminative evidence can be weakened during early feature extraction, increasing the semantic gap between low-level visual cues and high-level representations.

The dual-foveated kernel combines three Gaussian components motivated by biological vision and classical signal processing Jia et al. [2018], Huang et al. [2018], as illustrated in Fig. 3. We define the discrete kernel $\mathbf{K} \in \mathbb{R}^{N \times N}$ by sampling a continuous radial profile $\mathcal{K}(r)$, where N is odd and $r_{\max} = (N-1)/2$. For $(x, y) \in \{-r_{\max}, \ldots, r_{\max}\}^2$, the radius is $r = \sqrt{x^2 + y^2}$. The kernel is constructed by combining three antagonistic Gaussian terms:

$$\mathcal{K}(x, y) = \alpha \cdot G_c(r) - \beta \cdot G_i(r) + \gamma \cdot G_p(r), \quad (1)$$

where each component serves a distinct role in spatial and frequency-domain processing.

Positive-phase excitation term. The positive-phase excitation term can be represented as

$$G_c(r) = \exp\left(-\frac{r^2}{2\sigma_c^2}\right). \quad (2)$$

This component is motivated by the localized excitatory response near the retinal fovea and uses a small variance σ_c. Its narrow spatial support preserves fine-scale structures and high-frequency responses. However, when used alone, it also retains high-frequency noise and therefore cannot reliably distinguish target-related details from noise.

Negative-phase inhibition term. The negative-phase inhibition term can be represented as

$$G_i(r) = \exp\left(-\frac{r^2}{2\sigma_i^2}\right), \quad \sigma_i > \sigma_c. \quad (3)$$

This component subtracts a broader Gaussian response with $\sigma_i > \sigma_c$. Large and smooth regions, such as roads and water surfaces, are dominated by low-frequency information. The resulting center-surround antagonism attenuates these background responses and increases the relative signal-to-noise ratio of tiny targets.

Peripheral re-excitation term. The peripheral re-excitation term can be represented as

$$G_p(r) = \exp\left(-\frac{(r-\mu_p)^2}{2\sigma_p^2}\right). \quad (4)$$

This component forms an annular response centered at μ_p, providing a complementary response that preserves selected

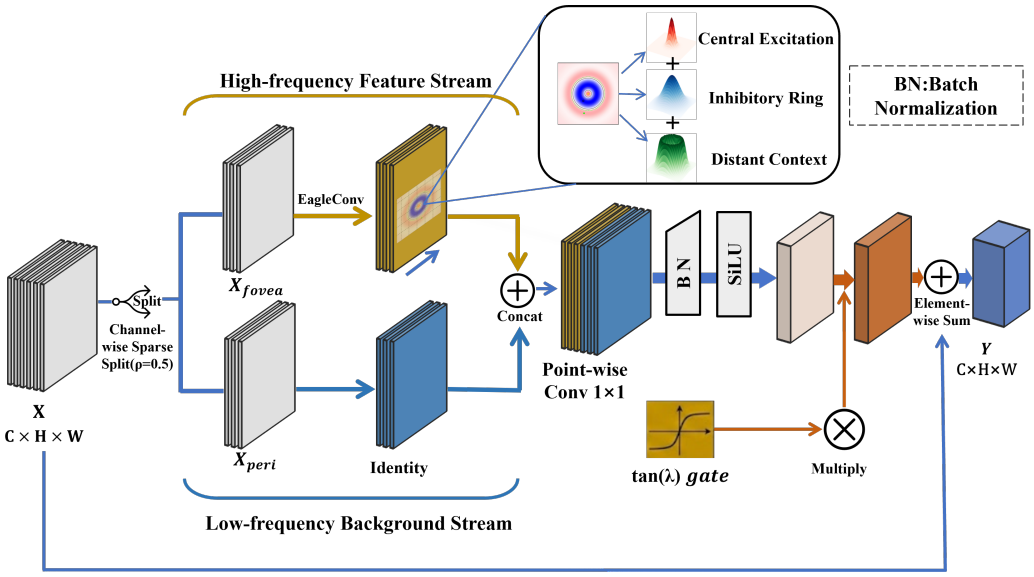
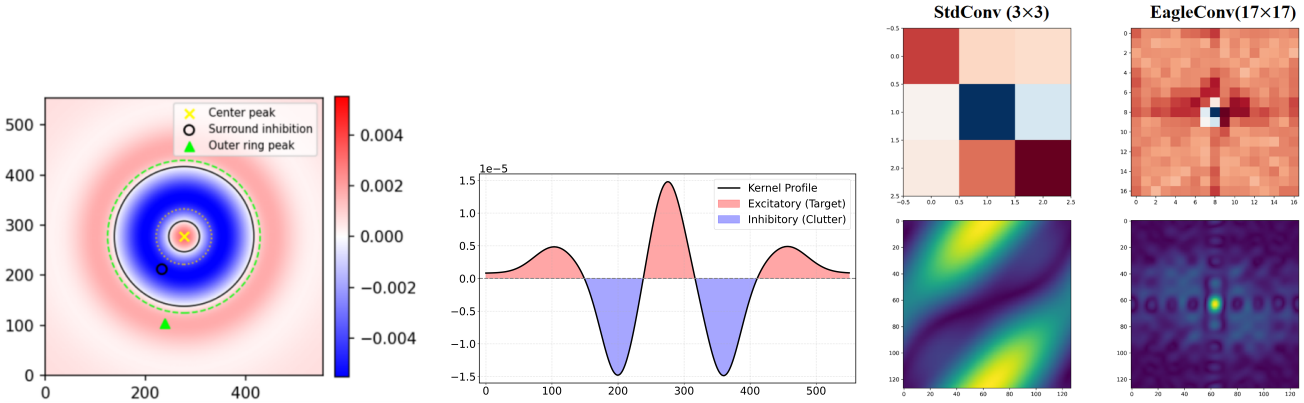


Figure 2: The architecture of EagleConv module is designed based on sparse dual pathways and adaptive gating.



(a) For clarity, we visualize a kernel of size 555×555: cross denotes the central positive peak, ○ the center-side suppressive weights, and △ the weak peripheral positive peaks (foveated depressions).

(b) The cross-sectional profile most clearly illustrates dual foveation: the red convex lobe captures target responses, while the blue concave lobe cancels background activations.

(c) The trained 3×3 convolution shows a low-pass frequency response, while EagleConv exhibits a band-pass pattern, suppressing low-frequency background components and preserving mid- and high-frequency details.

Figure 3: Visualization of the dual-foveated kernel design in EagleConv.

contextual information while reducing nearby clutter. This component balances fine local details with a broader semantic context, alleviating the tendency of conventional edge operators to preserve texture while discarding contextual information.

3.2 FREQUENCY-DOMAIN CHARACTERIZATION

As shown in Fig. 3(c), the frequency-domain comparison further indicates that EagleConv does not improve small-object detection merely by enlarging the kernel size or increasing the number of parameters. Instead, its advantage is closely related to the frequency-selective filtering behavior learned during training. By suppressing redundant background components while retaining discriminative structural details, EagleConv better preserves edges, textures, and local patterns that are essential for small-object detection.

To gain a deeper understanding of the signal-processing behavior of the dual-foveated kernel, this section provides a systematic analysis from the frequency-domain perspective. By the Fourier transform property, the 2D Fourier transform of a spatial Gaussian function $g(r) = \exp(-r^2/(2\sigma^2))$ remains Gaussian:

$$\mathcal{F}\{g(r)\}(\omega) = 2\pi\sigma^2 \exp(-2\pi^2\sigma^2\omega^2), \tag{5}$$

where $\omega = \sqrt{\omega_x^2 + \omega_y^2}$ denotes the spatial frequency of the radius. Due to the linearity of the Fourier transform, the frequency response (transfer function) of the dual-foveated kernel is given by

$$H(\omega) = \alpha H_c(\omega) - \beta H_i(\omega) + \gamma H_p(\omega). \tag{6}$$

For the central excitation and lateral inhibition components, their transfer functions are, respectively,

$$H_c(\omega) = 2\pi\sigma_c^2 \exp(-2\pi^2\sigma_c^2\omega^2), \tag{7}$$

$$H_i(\omega) = 2\pi\sigma_i^2 \exp(-2\pi^2\sigma_i^2\omega^2). \tag{8}$$

Since $\sigma_i > \sigma_c$, $H_i(\omega)$ exhibits a larger magnitude in the low-frequency regime but decays faster, whereas $H_c(\omega)$ maintains a stronger response at higher frequencies. Their difference, $\alpha H_c(\omega) - \beta H_i(\omega)$, therefore, yields a characteristic band-pass filtering behavior.

For the ring-shaped peripheral component, its Fourier transform involves a shifted Gaussian and can be derived via the Hankel transform:

$$H_p(\omega) = 2\pi\sigma_p^2 \exp(-2\pi^2\sigma_p^2\omega^2) \cdot J_0(2\pi\mu_p\omega), \tag{9}$$

where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind. This term introduces an oscillatory modulation in the frequency domain, enhancing the selectivity of the response over specific mid-frequency bands.

From a signal-processing perspective, the dual-foveated kernel of EagleConv acts as a tunable band-pass filter. Its center-surround structure suppresses DC and low-frequency backgrounds, enhances edge and texture cues of tiny objects, and attenuates high-frequency noise through Gaussian roll-off. Adjusting the spatial parameters adapts the frequency response to different object scales.

3.3 SELECTIVE DUAL-FOVEA ENHANCEMENT FOR REAL-TIME DEPLOYMENT

Applying dual-foveated filtering to all channels may over-impose the bionic prior and enforce a suboptimal fixed frequency response for some channels. Inspired by channel grouping in GhostNet Han et al. [2020] and ShuffleNet Zhang et al. [2018], we use partial-channel sparsification to balance bionic bias and representational flexibility.

Given an input feature tensor $\mathbf{X} \in R^{C\times H\times W}$, we divide it along the channel dimension into two complementary subsets:

$$\mathbf{X}_{\text{fov}} = \mathbf{X}_{1:C_p}, \quad \mathbf{X}_{\text{rest}} = \mathbf{X}_{C_p+1:C}, \tag{10}$$

where the number of channels participating in dual-fovea processing is $C_p = \lfloor \rho C \rfloor, \rho \in (0, 1)$, with $\rho = 0.5$ by default.

We implement dual-fovea filtering using depthwise convolution. Specifically, for each channel in $\mathbf{X}_{\text{fov}}$, we apply the shared dual-foveated kernel $\hat{\mathbf{K}}$:

$$\tilde{\mathbf{X}}_{\text{fov}}^{(c)} = \mathbf{X}_{\text{fov}}^{(c)} * \hat{\mathbf{K}}, \quad c = 1, \ldots, C_p, \tag{11}$$

where $*$ denotes 2D convolution. This design has computational complexity $O(C_p N^2 HW)$, offering a significant efficiency gain over standard convolution with complexity $O(C^2 N^2 HW)$.

After differential processing, we concatenate the two feature groups and employ a 1×1 pointwise convolution to enable cross-channel interaction and feature reorganization:

$$\mathbf{Y}_{\text{cat}} = [\tilde{\mathbf{X}}_{\text{fov}}, \mathbf{X}_{\text{rest}}] \in R^{C\times H\times W}, \tag{12}$$

$$\mathbf{Y}_{\text{mix}} = \phi\left(\text{BN}\left(\mathbf{W}_{1\times 1} * \mathbf{Y}_{\text{cat}}\right)\right), \tag{13}$$

where $\phi(\cdot)$ denotes the activation function, $\text{BN}(\cdot)$ denotes batch normalization, and $\mathbf{W}_{1\times 1}$ denotes the learnable weights of the pointwise convolution.

With this design, the fovea-processed channels emphasize frequency-selective edge/texture cues, while the untouched channels preserve the original full-band information. Subsequent 1×1 mixing learns channel-wise fusion weights, enabling adaptive integration of the two streams and allowing the network to learn the optimal combination strategy from the data.

3.4 DYNAMICALLY BALANCING BIOLOGICALLY INSPIRED PRIORS AND DATA-DRIVEN REPRESENTATIONS

To adaptively balance the dual-fovea inductive bias with data-driven representations, we introduce a learnable scalar gate $g \in R$, initialized as $g_0 = 0$, and combine the enhanced features with the input through a residual formulation:

$$\mathbf{Y} = \mathbf{X} + \tanh(g) \cdot \mathbf{Y}_{\text{mix}}. \tag{14}$$

The nonlinearity $\tanh(\cdot)$ constrains the gate value to $[-1, 1]$, providing three key advantages. First, bidirectional modulation allows a positive gate value ($\tanh(g) > 0$) to make the dual-fovea prior contribute positively to the final representation, while a negative gate value ($\tanh(g) < 0$) enables suppression or inversion of the prior when it is inconsistent with the target task. This modulation improves robustness when the dual-fovea assumption is only partially applicable. Second, the bounded and saturating characteristics of $\tanh(\cdot)$ prevent excessive amplification of the residual branch, improving optimization stability and reducing numerical drift. Third, when $g = 0$, $\tanh(0) = 0$, and the module degenerates to an identity mapping, that is, $\mathbf{Y} = \mathbf{X}$. Consequently, the dual-fovea branch does not perturb pretrained backbones or randomly initialized networks at the beginning of training,

facilitating unobstructed gradient propagation and providing a stable warm-up phase.

From an optimization viewpoint, the gradient with respect to g can be written as

$$\frac{\partial \mathcal{L}}{\partial g} = \left\langle \frac{\partial \mathcal{L}}{\partial \mathbf{Y}}, \mathbf{Y}_{\text{mix}} \right\rangle \cdot \operatorname{sech}^2(g), \tag{15}$$

where $\langle \cdot, \cdot \rangle$ denotes the sum of element-wise products, and $\operatorname{sech}^2(g) = 1 - \tanh^2(g)$. In particular $\operatorname{sech}^2(g)$ achieves its maximum value 1 at $g = 0$ and decreases monotonically as $|g|$ grows. Therefore, in early training ($g \approx 0$), the gate is more sensitive to loss and can rapidly adjust in the appropriate direction; later, as $|g|$ increases, the decreasing gradient naturally stabilizes the learned contribution of the dual-fovea prior and mitigates late-stage oscillations.

3.5 FIDELITY-PRESERVING DOWNSAMPLING AND FEATURE ALIGNMENT

In downsampling stages with stride (s>1), strided convolution may introduce aliasing by folding high-frequency components above the Nyquist limit, especially in edge-rich feature maps. To reduce this information loss, we use average pooling as an anti-aliasing prefilter:

$$\mathbf{Y}_{\text{down}} = \text{AvgPool}_s(\mathbf{Y}), \tag{16}$$

where AvgPool_s denotes average grouping with both both kernel size and the stride set to s. In the frequency domain, an $s \times s$ box (mean) filter has a separable transfer function given by

$$H_{\text{avg}}(\omega_x, \omega_y) = \text{sinc}(s\omega_x)\text{sinc}(s\omega_y). \tag{17}$$

Its response peaks at DC and decays with frequency, with the first zero matching the post-downsampling Nyquist limit. Thus, average pooling serves as a lightweight low-pass prefilter that suppresses aliased components. Combined with the band-pass behavior of the dual-foveated kernel, it forms an efficient cascaded anti-aliasing filter.

After downsampling, we apply a standard convolution block to project features along the channel dimension:

$$\mathbf{Z} = \text{StdConv}_k(\mathbf{Y}_{\text{down}}) \in R^{C_{\text{out}} \times H' \times W'}, \tag{18}$$

where StdConv_k consists of a $k \times k$ convolution followed by batch normalization and SiLU activation. Here, $H' = \lfloor H/s \rfloor$ and $W' = \lfloor W/s \rfloor$ denote the spatial sizes after downsampling. By default, we set $k = 1$ (i.e., a pointwise convolution) to minimize additional parameters while preserving flexible channel remapping.

In general, the forward computation of EagleConv is summarized below.

$$\mathbf{Z} = \text{StdConv}_k\left(\text{AvgPool}_s\left(\mathbf{X} + \tanh(g) \cdot \mathbf{Y}_{\text{mix}}\right)\right), \tag{19}$$

with

$$\mathbf{Y}_{\text{mix}} = \phi\left(\text{BN}\left(\text{PW}\left(\left[\text{DW}_{\hat{\mathbf{K}}}(\mathbf{X}_{\text{fov}}); \mathbf{X}_{\text{rest}}\right]\right)\right)\right). \tag{20}$$

Here, $\text{DW}_{\hat{\mathbf{K}}}$ denotes depthwise convolution with the dual-foveated kernel $\hat{\mathbf{K}}$, PW denotes pointwise (1×1) convolution, and $[\cdot;\cdot]$ indicates channel-wise concatenation.

4 EXPERIMENTS

To evaluate EagleConv for small-object detection, we conducted extensive experiments on three benchmarks and provide ablations and visualizations to quantify each component's contribution.

4.1 SETUP

Hardware. All experiments were conducted on an NVIDIA A100 GPU with 80 GB memory. Our method is implemented in PyTorch and accelerated using CUDA 11.8.

Training settings. For a fair comparison, all images are resized as 960×960. We set the optimizer momentum to 0.937, use a batch size of 16, and train for 300 epochs with an early-stopping patience of 100.

Base detector. We use YOLOv11 Khanam and Hussain [2024] as the baseline. YOLOv11 offers five scales (n/s/m/l/x) with progressively increased depth and width while keeping other designs fixed, where larger models generally improve accuracy. For efficiency and consistent settings, all experiments are conducted on YOLOv11s with depth and width multipliers set to 0.50.

Datasets. We evaluated on three representative small-object detection benchmarks, following the experimental protocols in Liu et al. [2025], Yang et al. [2022], Du et al. [2023].

- **VisDrone2019-DET** Du et al. [2019] contains 10,209 UAV images (10 classes); following the official split, we use 6,471 images for training and 548 for validation (resolutions 960×540∼ 2000×1500), with learning rate 0.01.
- **USOD** Zhang et al. [2024] includes 3,000 visible-spectrum images with 43,378 vehicles (train/test 7:3); small instances dominate (96.3% $< 16 \times 16$, 99.9% $< 32 \times 32$), and we use 2,100/900 images for train/val with learning rate 0.001.
- **AI-TOD** Wang et al. [2021] provides 28,036 aerial images with 700,621 instances in 8 categories, featuring much smaller objects (average size ≈ 12.8 pixels).

Evaluation Metrics. We report Precision (P), Recall (R), and mean Average Precision at IoU thresholds 0.50 and 0.50:0.95 (mAP_{50} and $\text{mAP}_{50:95}$) as the primary accuracy

Table 1: Results of integrating each module into YOLOv11s on VisDrone2019-DET and USOD with an input size of 960×960. All values are percentages. The best and second-best results are highlighted in **bold** and underlined, respectively, within each dataset.

	VisDrone2019-DET				USOD			
Module	P	R	mAP_{50}	$mAP_{50:95}$	P	R	mAP_{50}	$mAP_{50:95}$
Conv (base)	56.14	44.41	46.96	28.46	92.14	86.77	91.42	36.26
PinConv Yang et al. [2025]	58.35	47.67	49.94	30.67	91.26	86.75	91.48	35.84
LSConv Wang et al. [2025]	59.90	47.13	49.76	30.54	90.68	85.31	90.18	33.64
IDConv Yu et al. [2024]	56.24	45.45	47.50	28.95	91.85	85.21	90.89	35.13
WTConv Finder et al. [2024]	55.89	44.36	46.60	28.13	91.13	85.38	90.74	35.29
FADC Chen et al. [2024]	59.96	47.76	51.00	31.19	90.54	84.72	89.88	33.86
EagleConv(7,7,7)	61.31	48.71	51.63	31.83	92.48	89.33	**93.41**	**37.86**
EagleConv(11,11,11)	61.72	**48.77**	51.72	31.91	92.52	89.14	93.10	37.29
EagleConv(17,17,17)	**62.03**	48.53	**52.20**	**32.22**	**92.69**	**89.71**	93.11	37.45

Table 2: Results of integrating each module into YOLOv11s on AI-TOD. All Settings are the same as Table 1.

Module	P	R	mAP_{50}	$mAP_{50:95}$
Conv	60.37	52.62	52.57	24.38
PinConv	60.72	53.30	53.26	24.34
LSConv	57.93	54.83	53.79	24.46
IDConv	62.58	54.78	54.68	25.77
WTConv	63.92	50.19	51.36	23.75
FADC	65.94	24.68	32.46	15.02
EagleConv-7	60.43	**57.50**	54.74	25.04
EagleConv-11	64.37	49.48	55.85	25.74
EagleConv-17	**69.40**	55.87	**60.75**	**28.24**

metrics. The size of the model is measured by the number of parameters (Params), and the inference efficiency is reported in the FPS.

4.2 RESULTS

Table 1, 2 compares the accuracy of standard and last convolution modules with EagleConv in YOLOv11s. It can be seen that the proposed EagleConv surpasses all counterparts in terms of accuracy, proving its effectiveness.

To assess deployment efficiency, we compare YOLOv11s, YOLOv11s+EagleConv, and other convolution modules on the same platform. With identical input resolution, batch size, and software settings, we measure parameters, computation, latency, memory bandwidth, and FPS, as reported in Table 3. Although EagleConv uses a large 17×17 kernel, it still exceeds 150 FPS, showing that large kernels do not noticeably reduce inference efficiency.

In extremely small object detection on the AI-TOD dataset, performance should not be evaluated only by absolute mAP gains. AP for very tiny $[0, 8]$ px, tiny $(8, 16]$ px, small $(16, 32]$ px and medium > 32 px objects increases from 25.66%, 52.82%, 56.13% and 69.35% to 30.76%, 57.69%, 58.98% and 73.58%, respectively. These improvements indicate that EagleConv better preserves weak textures, edges, and semantic cues while maintaining lightweight and real-time properties for drone and remote-sensing deployment.

Besides, we construct a COCO2017 Lin et al. [2014] subset with 5489 training images and 234 validation images to test EagleConv with YOLOv11s. EagleConv improves Precision from 45.28% to 54.24%, mAP_{50} from 38.06% to 39.21% and $\text{mAP}_{50:95}$ from 24.11% to 25.06% (with Recall 38.33% to 35.40%), suggesting that the dual-fovea mechanism is effective for extracting general features.

To evaluate the universality of EagleConv, we integrate it into YOLOv5x and Faster R-CNN, then compare the resulting models with Eagle-YOLO Liao et al. [2023], TPH-YOLOv5 Zhu et al. [2021], and the original Faster R-CNN Ren et al. [2017]. Under the Eagle-YOLO setting, YOLOv5x+EagleConv attains 38.8% mAP, surpassing Eagle-YOLO (32.91%) and TPH-YOLOv5 (30.94%), indicating that early frequency-selective modeling preserves small-object features and complements detector-level redesign. In addition, for Eagle-Faster R-CNN, we insert EagleConv before C2, C3, and C4 enter the FPN while keeping ResNet50, FPN, RPN, and the RoI Head unchanged, which enriches multi-scale features, strengthens object responses, and suppresses background clutter. Under identical training, it improves mAP_{50}/$\text{mAP}_{50:95}$ from 32.40%/18.28% to 33.88%/19.00%.

To verify the stability and reproducibility of EagleConv against random initialization, we repeat the experiments on the VisDrone2019 benchmark with five random seeds $[42, 3407, 1024, 2026, 8888]$, where the number of runs is limited by computational constraints. The reported results in terms of mean ± standard deviation demonstrate that the

Table 3: Efficiency comparison between different kernel sizes of YOLOv11s and EagleConv and other convolutions.

Method	**Params (M)**	**GFLOPs**	**Latency (ms/img)**	**Memory bandwidth (GB/s)**	**FPS**
YOLOv11s	9.43	21.7	5.92 ± 0.67	96.13	169
FADC	10.89	22.9	11.85 ± 0.50	67.17	84
PinConv	10.21	34.9	8.16 ± 1.18	109.66	122
LSConv	9.64	24.4	7.28 ± 0.04	112.50	137
WTConv	9.53	22.1	7.20 ± 0.94	104.60	138
IDConv	9.43	21.6	6.26 ± 0.8	93.88	159
EagleConv (7,7,7)	8.74	23.4	6.48 ± 0.48	117.55	154
EagleConv (17,17,17)	8.82	24.7	6.41 ± 0.52	115.99	155

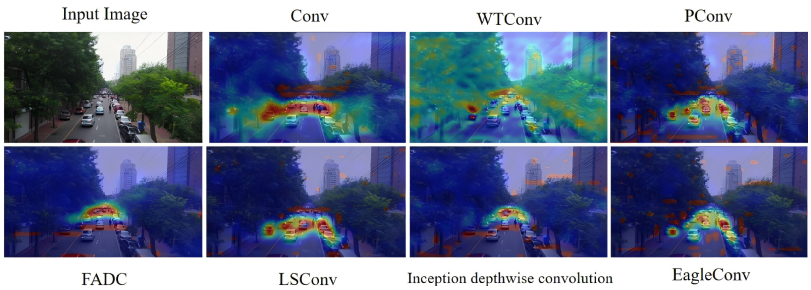


Figure 4: Feature response visualization of different convolution modules. EagleConv focuses strongly on targets while suppressing background activations.

Table 4: Computational complexity and parameter count of EagleConv.

Component	Complexity	Params.
DDF conv.	$\mathcal{O}(\rho C N^2 S)$	$\rho C N^2$ or 0
PW fusion	$\mathcal{O}(C^2 S)$	C^2
BN	$\mathcal{O}(CS)$	$2C$
Avg. pooling	$\mathcal{O}(CS)$	0
Projection	$\mathcal{O}(CC_{\text{out}}k^2S')$	$CC_{\text{out}}k^2 + 2C_{\text{out}}$
Gate	-	1

Table 5: Ablation study of EagleConv components. All values are percentages.

Module	P	R	mAP_{50}	$\text{mAP}_{50:95}$
No Gate	57.83	47.91	49.73	30.50
No DDF (DWConv)	60.75	**48.74**	51.33	31.24
Only Gaussian	60.27	47.47	50.65	31.10
Full EagleConv	**61.31**	48.71	**51.63**	**31.83**

proposed EagleConv is robust to different random initializations, achieving 51.95 ± 0.32 mAP_{50} and 32.03 ± 0.25 $\text{mAP}_{50:95}$ with YOLOv11s. The consistently small variances confirm that the performance gains are robust and not attributable to favorable random seeds.

In summary, EagleConv (17,17,17) achieves the best performance on VisDrone2019-DET, reaching 52.20% mAP_{50}, and also delivers competitive performance on USOD with 93.11% mAP_{50}. However, its advantage decreases on AI-TOD, where many targets fall below the effective kernel resolution. Fig. 4 further shows that EagleConv concentrates on targets and suppresses clutter, producing more discriminative features.

4.3 ABLATION STUDY

We analyze EagleConv complexity and parameters. For an input feature map $C \times H \times W$ with C_{out} output channels, let $S = HW$ and $S' = H'W'$, where $H' \times W'$ is the downsampled resolution. The dual-foveated and output projection kernels are $N \times N$ and $k \times k$, and ρ denotes the channel ratio processed by depthwise dual-foveated convolution (DDF). Table 4 gives component-wise complexity and parameters. With $\rho = 0.5$, $N = 15$, $k = 1$, and a fixed dual-foveated kernel, EagleConv has about $C^2 + CC_{\text{out}} + 2C + 2C_{\text{out}} + 1$ parameters, much fewer than $CC_{\text{out}}N^2$ for a same-receptive-field standard convolution. Its cost is mainly pointwise convolution, $\mathcal{O}(C^2HW)$, making it suitable for lightweight networks. We further validate EagleConv through VisDrone2019 ablations on YOLOv11s, using full EagleConv as the baseline. Under identical training settings, we vary the dual-foveated kernel size K and remove the gate, simplify the kernel, or remove

Table 6: Sensitivity analysis of kernel size and ablation study on ρ (both on VisDrone2019-DET; input size 960×960).

Kernel size sensitivity						
Kernel Size	**P**	**R**	**mAP$_{50}$**	**mAP$_{50\text{-}95}$**	**Params (M)**	**GFLOPs**
K=5	60.30%	48.64%	50.64%	31.10%	8.73	23.3
K=7	61.31%	48.71%	51.63%	31.83%	8.74	23.4
K=9	60.01%	48.75%	51.30%	31.64%	8.75	23.6
K=11	61.72%	48.77%	51.72%	31.91%	8.76	23.8
K=13	60.09%	49.31%	51.71%	31.58%	8.78	24.1
K=15	60.92%	48.64%	51.97%	31.95%	8.80	24.4
K=17	**62.03%**	48.53%	**52.20%**	**32.22%**	8.82	24.7
K=19	60.75%	49.02%	51.92%	31.90%	8.84	25.1
K=21	61.03%	48.95%	51.98%	31.91%	8.87	25.5
K=23	60.57%	**49.91%**	51.94%	31.83%	8.89	26.0
Hyperparameter ρ (kernel size $k = 7$)						
ρ	**P**	**R**	**mAP$_{50}$**	**mAP$_{50\text{-}95}$**	**Params (M)**	**GFLOPs**
0.01	60.27%	48.03%	51.09%	31.30%	8.72	23.2
0.25	59.28%	**48.77%**	50.77%	31.35%	8.73	23.3
0.5	61.31%	48.71%	51.63%	31.83%	8.74	23.4
0.75	**61.60%**	48.38%	**51.76%**	**31.91%**	8.75	23.6
1	60.78%	48.17%	50.90%	31.19%	8.76	23.7

center-surround antagonism to assess small-object SNR and semantic preservation. All ablation experiments use input size 960, batch size 6, 300 epochs, and patience 100.

Kernel size. We vary K from 5 to 23 with step 2 on VisDrone2019 and summarize the results in Table 6. Small kernels ($K < 9$) cannot form a stable three-ring pattern, resulting in weaker background suppression and lower tiny-object SNR. Increasing K improves mAP$_{50}$ from 50.64% to 52.20%, but increases computational cost.

Kernel Structure. Table 5 shows that Full EagleConv achieves the best overall performance. Removing the gate causes the largest drop, reducing precision and mAP$_{50}$ by 3.48% and 1.90%, respectively. Replacing DDF with DW-Conv mainly reduces mAP$_{50:95}$ by 0.59%, while using only a Gaussian kernel reduces mAP$_{50}$ by 0.98%. These results indicate that EagleConv gains come mainly from gated frequency-selective modeling and center-surround antagonism, rather than merely using more parameters or a larger receptive field.

Hyperparameter ρ. We vary $\rho \in \{0.01, 0.25, 0.5, 0.75, 1\}$ with $K = 7$ and report the results in Table 6. The model performance first improves and then declines as ρ increases. Although $\rho = 0.75$ achieves highest $mAP_{50:95}$ of 31.91%, $\rho = 0.5$ yields a comparable $mAP_{50:95}$ of 31.83% and high Recall of 48.71%, with only a slight increase in parameter count. We select $\rho = 0.5$ to balance accuracy and efficiency.

5 CONCLUSION

In summary, EagleConv inspired by a strong-center-intermediate-suppression-weak-periphery pattern consistently improves small-object detection across three benchmarks. Across three benchmarks and multiple detectors, it improves small-object detection while remaining lightweight and real-time. Future work will study its scalability to more architectures, robustness under challenging conditions (e.g., adverse weather and low light) with better dynamic-background handling, and extension to other tasks such as semantic segmentation and object tracking.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China (Grant No. 62402249 and U25A20401), in part by the Science and Technology Plan Project of Inner Mongolia A. R. of China (Grant No. 2026YFHH0321 and 2024KJHZ0003), in part by the "Unveiling the List and Appointing the Best" Project under the Science and Technology "Breakthrough" Project of of Inner Mongolia A. R. of China (Grant No. 2025KJTW0010 and 2025KJTW0021), and in part by the Industry-Commissioned Project (Grant No. NMGMZ-ZB-20250714).

References

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

Jierun Chen, Shiu hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, and S.-H. Gary Chan. Run, don't walk: Chasing higher FLOPS for faster neural networks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12021–12031, 2023.

Linwei Chen, Lin Gu, Dezhi Zheng, and Ying Fu. Frequency-adaptive dilated convolution for semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3414–3425, 2024.

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*, 45:13467–13488, 2023.

Lu Chi, Borui Jiang, and Yadong Mu. Fast Fourier convolution. In *Advances in Neural Information Processing Systems*, pages 4479–4488, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13435–13444, 2023.

Dawei Du, Pengfei Zhu, and Longyin Wen et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results. In *2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 213–226, 2019.

Shahaf E Finder, Roy Amoyal, Eran Treister, and Oren Freifeld. Wavelet convolutions for large receptive fields. In *European Conference on Computer Vision*, pages 363–380, 2024.

Abimael Guzman-Pando and Mario I. Chacon-Murguia. DeepFoveaNet: Deep fovea Eagle-Eye bioinspired model to detect moving objects. *IEEE Transactions on Image Processing*, 30:7090–7100, 2021.

Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. GhostNet: More features from cheap operations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1577–1586, 2020.

Lifei Hao, Baoqi Huang, Bing Jia, and Guoqiang Mao. On the fine-grained crowd analysis via passive WiFi sensing. *IEEE Transactions on Mobile Computing*, 23:6697–6711, 2024.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications, 2017.

Peiyun Hu and Deva Ramanan. Finding tiny faces. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1530, 2017.

Baoqi Huang, Minmin Liu, Zhendong Xu, and Bing Jia. On the performance analysis of WiFi based localization. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4369–4373, 2018.

Baoqi Huang, Guoqiang Mao, Yong Qin, and Yun Wei. Pedestrian flow estimation through passive WiFi sensing. *IEEE Transactions on Mobile Computing*, 20:1529–1542, 2021.

Bing Jia, Baoqi Huang, Hepeng Gao, and Wuyungerile Li. Dimension reduction in radio maps based on the supervised kernel principal component analysis. *Soft Computing*, 22:7697–7703, 2018.

Rahima Khanam and Muhammad Hussain. YOLOv11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1951–1959, 2017.

Lyuchao Liao, Linsen Luo, Jinya Su, Zhu Xiao, Fumin Zou, and Yuyuan Lin. Eagle-YOLO: An eagle-inspired YOLO for object detection in unmanned aerial vehicles scenarios. *Mathematics*, 11, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

Kai Liu, Zhihang Fu, Sheng Jin, Ze Chen, Fan Zhou, Rongxin Jiang, Yaowu Chen, and Jieping Ye. ESOD: Efficient small object detection on high-resolution images. *IEEE Transactions on Image Processing*, 34:183–195, 2025.

Kang Liu, Ju Huang, and Xuelong Li. Eagle-Eye-Inspired attention for object detection in remote sensing. *Remote Sensing*, 14, 2022.

Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. FcaNet: Frequency channel attention networks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 763–772, 2021.

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*, 39:1137–1149, 2017.

Yulong Shi, Mingwei Sun, Yongshuai Wang, Jiahao Ma, and Zengqiang Chen. EViT: An eagle vision transformer with Bi-Fovea self-attention. *IEEE Transactions on Cybernetics*, 55:1288–1300, 2025.

Ao Wang, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. LSNet: See large, focus small. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9718–9729, 2025.

Jian Wang, Xinqi Li, Jiafu Chen, Lihui Zhou, Linyang Guo, Zihao He, Hao Zhou, and Zechen Zhang. DPH-YOLOv8: Improved YOLOv8 based on double prediction heads for the UAV image object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

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, 2021.

Tongyan Wu, Haibin Duan, and Zhigang Zeng. Biological Eagle-Eye-Based correlation filter learning for fast UAV tracking. *IEEE Transactions on Instrumentation and Measurement*, 73:1–12, 2024.

Chenhongyi Yang, Zehao Huang, and Naiyan Wang. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13658–13667, 2022.

Jiangnan Yang, Shuangli Liu, Jingjun Wu, Xinyu Su, Nan Hai, and Xueli Huang. Pinwheel-shaped convolution and scale-based dynamic loss for infrared small target detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:9202–9210, 2025.

Weihao Yu, Pan Zhou, Shuicheng Yan, and Xinchao Wang. InceptionNeXt: When inception meets ConvNeXt. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5672–5683, 2024.

Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7324–7334, 2019.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, 2018.

Yin Zhang, Mu Ye, Guiyi Zhu, Yong Liu, Pengyu Guo, and Junhua Yan. FFCA-YOLO for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

Zichao Zhang, Shubo Wang, Jian Chen, and Yu Han. A bionic dynamic path planning algorithm of the micro UAV based on the fusion of deep neural network optimization/filtering and Hawk-Eye vision. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53: 3728–3740, 2023.

Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson Lau. BiFormer: Vision transformer with bi-level routing attention. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10323–10333, 2023.

Xingkui Zhu, Shuchang Lyu, Xu Wang, and Qi Zhao. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops(ICCVW)*, pages 2778–2788, 2021.