ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10889402

Delving Into Coarse-Fine Feature Interaction Alignment for UAV Object Detection

Yanchao Bi
School of Computer Science and Technology
Shandong Jianzhu University
Jinan, P.R. China
2022110101@stu.sdjzu.edu.com

Yang Ning
School of Computer Science and Technology
Shandong Jianzhu University
Jinan, P.R. China
ningyang20@sdjzu.edu.cn

Xiushan Nie*
School of Computer Science and Technology
Shandong Jianzhu University
Jinan, P.R. China
niexsh@hotmail.com

***Abstract*—Due to limited features and dense object layouts, object detection in UAV images is challenging. Given that existing feature fusion methods have not fully explored the relationship between fine- and coarse-grained features, direct feature fusion can result in poor correlation between them, hindering the representative capability of fine-grained semantic information. To alleviate this issue, we introduce a method of Coarse-fine Feature Interaction Alignment (CFIA), which enhances the correlation between coarse-grained and fine-grained features across multi-scale feature maps through their interactive alignment. Firstly, we present the Wavelet-based High-frequency Preserving Down-sampling (WHPD), utilizing wavelet transform to extract high-frequency information to enhance object boundaries, minimizing crucial fine-grained information loss. Secondly, we propose the Feature Refinement and Interaction Alignment Strategy (FRIAS), which achieves feature interaction alignment by establishing the association of feature maps between coarse-grained and fine-grained features. This enhances the representative capability of feature maps at various scales for detecting small objects. Extensive experiments on the VisDrone, CARPK, and Drone-vs-Bird datasets have demonstrated the effectiveness of the CFIA method, which is highly competitive with state-of-the-art methods. The code is available at https://github.com/b-yanchao/CFIA.git.**

***Index Terms*—Drone-view small object detection, Fine-grained information mining, Coarse-fine-grained feature alignment.**

I. Introduction

Object detection for Unmanned Aerial Vehicles (UAVs) is crucial in various scenarios, including military applications, autonomous driving, and intelligent inspection [1]–[4]. The prevalence of small objects in UAV imagery challenges extracting sufficient effective features, resulting in a notable performance gap between small objects and routine objects [5].

To alleviate this gap, multi-scale feature fusion methods effectively utilize the unique receptive field properties at each level by combining feature maps across various scales [6]–[8]. However, the backbone loses much fine-grained information through successive feature map down-scaling, leading to poor correlation between the acquired coarse-grained and the fine-grained information. Directly fusing multi-level features can lead to difficulty in precisely representing the semantic information of fine-grained features [9]–[11]. Determining effective feature alignment methods to reduce the difficulty of separating foreground objects from the background is key to achieving the above goals [12], [13].

Therefore, we deeply explore the relationship between coarse and fine-grained features and propose the Coarse-fine Feature Interaction Alignment (CFIA) method, enabling multi-scale object detection in UAV imagery through coarse-fine feature alignment. Specifically, we first propose a Wavelet-based High-frequency Preserving Down-sampling (WHPD) to capture high-frequency details via wavelet transformation [14] to enhance the object boundary information, reducing the critical fine-grained information lost of down-sampling operation. To enhance the interaction between coarse and fine-grained features, we present the Feature Refinement and Interaction Alignment Strategy (FRIAS), comprising two modules. The Coarse-to-Fine Feature Alignment Module (CFAM) merges coarse-grained features into shallow feature maps via a progressive fusion approach. And the Fine-to-Coarse Contextual Semantic Information Mining Module (FCSIM) learning different receptive field coarse-grained information relevant to fine-grained features directly in the shallow feature map. By interactively aligning, we enhance the correlation between coarse-grained and fine-grained information, strengthening the representative capability of the feature maps for small objects.

As a simple and effective design, the CFIA method can be integrated with a variety of convolution-based detectors, resulting in significant performance improvements. In summary, our work makes contributions in three main aspects:

- We introduce a Wavelet-based High-frequency Preserving Down-sampling by enhancing the object boundary information to reduce the critical fine-grained information lost by the down-sampling operation.
- We propose a Feature Refinement and Interaction Alignment Strategy, which enhances the correlation between coarse- and fine-grained features through the interactive alignment of the Coarse-fine Feature Alignment Module and the Fine-coarse Contextual Semantic Information Mining Module, strengthening the representative capa-

*Corresponding author.

This work is supported in part by the Shandong Provincial Natural Science Foundation for Distinguished Young Scholars (ZR2021JQ26), National Natural Science Foundation of China (62176141), Major science and technology innovation project of Shandong Province (2021CXGC11204), Taishan Scholar Project of Shandong Province (tsqn202103088), Natural Science Foundation of Shandong Province (ZR202103010201).

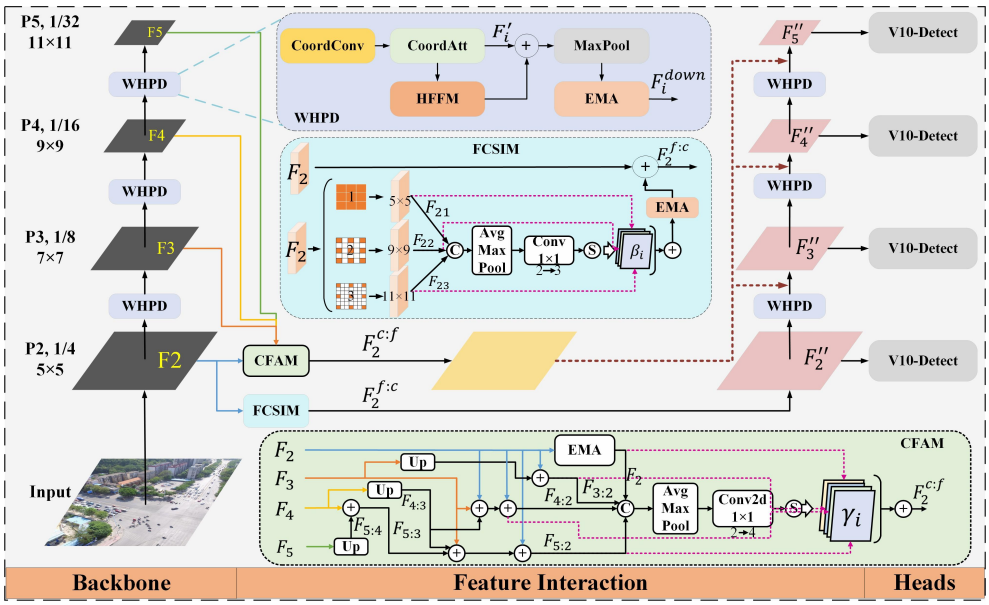

Fig. 1. The architecture of the proposed CFIA method.

bility of the feature map for small objects.

- We experimented with the CFIA method on the VisDrone, CARPK, and Drone-vs-Bird datasets, and achieved improvements of 2.3%, 2.3%, and 3.8% in mAP, respectively, compared to the baseline with a 640 × 640 input resolution. These improvements are competitive with state-of-the-art methods.

II. Method

A. Overall Architecture

Figure 1 illustrates an overview of the proposed method. Deliver a UAV image into the backbone, where it first undergoes the WHPD to enhance the object boundary by capturing high-frequency information, enhancing the correlation of the acquired coarse-grained features with the fine-grained features. Then, the CFAM merges coarse-grained features into shallow feature maps via a progressive fusion method. The FCSIM learns different receptive field coarse-grained information relevant to fine-grained features directly within the shallow feature map. By interactively aligning CFAM and FCSIM, we enhance the correlation between coarse-grained and fine-grained features, obtaining multi-scale feature maps with stronger expressive ability for small objects, and effectively improving the performance of UAV object detection.

B. Wavelet-based High-frequency Preserving Down-sampling

The existing backbone tends to lose fine-grained features due to down-sampling operations, resulting in a poor correlation between the acquired coarse-grained information and the fine-grained information. This makes it difficult to accurately express the semantic information of fine-grained features during subsequent feature fusion.

To alleviate the above problems, we propose a novel down-sampling method, Wavelet-based High-frequency Preserving Down-sampling (WHPD), to enhance the correlation between the acquired coarse-grained information and the fine-grained features by reducing the loss of critical fine-grained information. Specifically, as shown in Fig. 1, in order to make the network better use of the spatial structure to obtain more high-frequency information of small objects and reduce the influence of routine object boundaries. We first enhance the feature F_i using coordinate attention and coordinate convolution to obtain a feature that is more sensitive to positional information, which we denote as $F_i^{'}$:

$$F_i^{'} = CoordAttn(CoordConv(F_i)), \tag{1}$$

where $CoordConv(\cdot)$ denotes Coordinate Convolution [15], which learns the positional relationships between objects and their surroundings by incorporating positional information into the convolutional operation. Besides, $CoordAttn(\cdot)$ represents the Coordinate Attention mechanism [16], which assists the model in understanding complex contextual relationships by focusing on specific regions within the feature map.

Subsequently, high-frequency region information is obtained with the aid of wavelet transformation to enhance the boundary areas of small objects, reducing the loss of critical fine-grained information of these objects, and resulting in the down-sampled feature F_i^{down}:

$$F_i^{down} = EMA(MaxPool(F_i^{'} + HFFM(F_i^{'}))), \tag{2}$$

where $HFFM(\cdot)$ denotes the wavelet transform operation [14], which can obtain the high frequency key information. $MaxPool(\cdot)$ is the maximum pooling operation [17], which is more helpful to keep the important edge feature information of the small object. $EMA(\cdot)$ denotes the efficient multi-scale attention [18], which can make the feature maps pay more attention to the foreground object region to reduce the interference of the background region.

The proposed WHPD enhances the edge information of objects to reduce the loss of critical fine-grained information for small objects, obtaining coarse-grained semantic information that is more relevant to fine-grained features.

C. Feature Refinement and Interaction Alignment Strategy

Considering that directly integrating finer-grained information with more relevant coarse-grained information into shallow feature maps may lead to feature confusion due to conflicts, potentially overwhelming the semantic information of fine-grained features and increasing the risk of overwhelming or losing small objects [9]. Therefore, by deeply exploring the relationship between coarse-grained and fine-grained features, we propose a novel Feature Refinement and Interaction Alignment Strategy (FRIAS). This strategy enhances the representational capacity for understanding and collaboration between coarse-grained and fine-grained features, achieving interaction and alignment across multiple levels of features.

Specifically, as shown in Fig. 1, the strategy contains a Coarse-to-fine Feature Alignment Module (CFAM) and a Fine-to-coarse Contextual Semantic Information Mining Module (FCSIM). To establish the association of coarse-grained features with fine-grained features, the CFAM module obtains the feature map for transitioning from coarse-grained to fine-grained features by interactively fusing deep feature maps (F_3, F_4, and F_5) with those at the F_2 level, complementing them progressively at different scales. The formula is as follows:

$$F_{i:2} = \begin{cases} Up(F_3) + F_2, & i = 3, \\ Up(Up(F_4) + F_3) + F_2, & i = 4, \\ Up(Up(Up(F_5) + F_4) + Up(F_4) + F_3) + F_2, & i = 5, \end{cases} \tag{3}$$

where $Up(\cdot)$ denotes the up-sampling operation, specifically bilinear interpolation [19]. The transition from deep to shallow feature maps is achieved by interacting and complementing step by step, resulting in the F_2 level feature map that contains coarse-grained information from different receptive fields. Inspired by LSKNet [20], the spatial selection mechanism mask γ_i can be obtained as follows:

$$\gamma_i = \sigma(F^{2\to 4}([P_{avg}(F_{i:2}); P_{max}(F_{i:2})])), i \in [2, 5], \tag{4}$$

where σ, P_{avg}, and P_{max} denote the sigmoid function, the average pooling, and the maximum pooling operation, respectively. The symbol "[;]" denotes the concatenate operation.

Finally, we add appropriate receptive field coarse-grained contextual information for objects of different scales to the F_2 feature map, associating coarse-grained features with fine-grained features, resulting in $F_2^{c:f}$:

$$F_2^{c:f} = EMA(F_2) \times \gamma_1 + F_{3:2} \times \gamma_2 + F_{4:2} \times \gamma_3 + F_{5:2} \times \gamma_4. \tag{5}$$

To establish the association between fine-grained and coarse-grained features, the FCSIM module directly learns fine-grained relevant semantic information from the F_2 feature map. Specifically, as shown in Fig. 1, the input image undergoes down-sampling through convolutions with a 3×3 kernel and a stride of 2, progressively halving the dimensions of the feature map. This process yields feature maps at levels designated as F_1, F_2, F_3, F_4, and F_5. Each of these levels corresponds to receptive field sizes of 3×3, 5×5, 7×7, 9×9, and 11×11, respectively [7]. Typically, the F_1 feature map is not used due to its excessively large dimensions.

Thus, when using dilation convolution [21] with kernel size 3×3 on the F_2 feature map to obtain coarse-grained features F_{2d} with the same receptive field as the F_4 and F_5 feature maps, the required dilation rates are 2 and 3, respectively. And the convolution with a dilation rate of 1 is used to prevent grid effects. Note that the kernel size of 3×3 has been experimentally verified to offer the optimal accuracy and the fastest speed.

Similarly, we first obtain the mask β_i of semantic information F_{2d} relevant to fine-grained features. Then, we utilize the spatial selection mechanism [20] to associate the fine-grained features with coarse-grained features, resulting in $F_2^{f:c}$:

$$F_2^{f:c} = EMA(F_{21} \times \beta_1 + F_{22} \times \beta_2 + F_{23} \times \beta_3) + F_2. \tag{6}$$

Finally, by interactively aligning $F_2^{c:f}$ and $F_2^{f:c}$, we achieve the complementarity between coarse-grained and fine-grained features, resulting in more representative capability feature maps for small objects across different scales, denoted as F_i':

$$F_i^{''} = \begin{cases} F_2^{f:c}, i = 2 \\ WHPD(F_2^{f:c}) + MaxPool(F_2^{c:f}), 2 < i \leq 5, \end{cases} \tag{7}$$

where $WHPD(\cdot)$ denotes Equations 1 and 2 of the Wavelet-based High-frequency Preserving Down-sampling in II-B.

Through extensive experiments, we found that the interactively aligning between coarse-grained and fine-grained features enhanced the correlation between them. This improvement boosts the feature map's expressive ability for small objects and enhances UAV object detection performance.

III. Experiments

Combining our CFIA method with YOLOv10 [22], the latest YOLO model, we tested it on mainstream benchmarks for UAV images: VisDrone [23], CARPK [24], and Drone-vs-Bird [25]. VisDrone consists of 7,019 high-resolution images (2000×1500) with small, dense objects in 10 categories, using 6,471 for training, 548 for validation, and 1,610 for testing. CARPK contains 989 training images and 459 test images captured by drones, presenting challenges such as small and densely distributed objects. Drone-vs-Bird includes 1,387 training and 434 test images, rich in UAV and environmental data. We evaluate performance using mean Average Precision (mAP) [26], where the IoU of mAP_{50} is 0.5, and the IoU of mAP is the average across thresholds from 0.5 to 0.95.

A. Implementation Details

We implement our network based on PyTorch [27]. All models utilize YOLOv10m as the baseline and are trained for 200 epochs. Our method employs the same loss function as YOLOv10m, which includes object classification loss and bounding box regression loss. The classification loss incorporates BCELoss [22] and FocalLoss [28], while the regression

TABLE I
PERFORMANCE COMPARISON WITH STATE-OF-THE-ART METHODS ON VISDRONE, CARPK, AND DRONE-VS-BIRD. THE '-' STANDS FOR THE RESULT THAT IS NOT REPORTED.

Datasets	Method	#P(M)↓	mAP↑	mAP_{50}↑
VisDrone [23]	SOD-UAV [30]	32.20	26.80	45.70
	CZ FCOS Det [31]	-	33.91	56.20
	GFL V1+CEASC [32]	-	28.70	50.70
	FCOS+FGE+SAW [33]	-	-	51.50
	HRDNet [34]	152.2	31.40	53.30
	SDPDet [35]	-	34.20	57.80
	STF-YOLO [36]	46.74	36.73	-
	YOLOV10 [22]	16.40	35.60	56.10
	CFIA(ours)	**15.80**	**37.40**	**58.40**
CARPK [24]	CZ Det [31]	-	-	92.18
	QueryDet [37]	-	-	93.96
	YOLOv5 [38]	90.96	62.30	95.30
	Car-Det [38]	-	63.10	95.80
	YOLOv10 [22]	16.40	65.70	96.00
	CFIA(ours)	**15.80**	**68.00**	**96.10**
Drone-vs-Bird [25]	YOLOv5 [38]	90.96	-	74.60
	DETR+MNMS [39]	-	41.90	82.20
	YOLOv7 [40]	37.20	49.20	93.00
	YOLOv10 [22]	16.40	50.30	91.70
	CFIA(ours)	**15.80**	**54.10**	**94.50**

loss uses CIoULoss [29]. The input resolution is set to 640 × 640 and 1280 × 1280 on VisDrone, and to 640 × 640 on CARPK and Drone-vs-Bird. All models use the Adam optimizer for training, with an initial learning rate of 0.01 and a decay rate of 1e-5. We trained and tested the models on four NVIDIA RTX 2080 Ti GPUs, using a batch size of 8.

B. Comparison with State-of-the-Art Methods

Table I shows the comparison of our method CFIA with the state-of-the-art methods on three mainstream datasets. On the VisDrone, our method using a 1280 × 1280 input resolution compare to the state-of-the-art methods SDPDet [35], STF-YOLO [36], and YOLOv10 [22], and it was found that not only are the parameters of our method lower but also obtain the competitive performance(mAP: 35.6%→37.40%, mAP_{50}: 56.1%→58.40%). Compared to the SOD-UAV [30] with 640 × 640, our CFIA method has fewer parameters (32.2M→15.8M) and superior performance (26.8%→28.4%). We also tested our method on CARPK and Drone-vs-Bird with a 640 × 640 input resolution, achieving significant improvement over baseline (mAP: 65.7%→68.00%, 50.30%→54.10%), which is highly competitive with the state-of-the-art methods.

C. Ablation Studies

We validated our CFIA method on the VisDrone2019 validation set, analyzing the effectiveness of each component. We use YOLOv10m with a 640 × 640 input as the baseline and use mAP and mAP_{50} as evaluation metrics (See Table II).

As demonstrated in Table II, the implementation of the Feature Refinement and Interaction Alignment Strategy for aligning coarse-fine features not only reduces the model's parameters (16.4M→13.9M) but also markedly enhances model performance (mAP: 26.1%→27.5%, mAP_{50}: 42.5%→45.1%).

TABLE II
ABLATION OF EACH COMPONENT ON VISDRONE VALIDATION SET. **WHPD** STANDS FOR WAVELET-BASED HIGH-FREQUENCY PRESERVING DOWN-SAMPLING. **FRIAS** STANDS FOR FEATURE REFINEMENT AND INTERACTION ALIGNMENT STRATEGY.

Baseline	FRIAS	WHPD	#P(M)↓	GFLOPs	mAP↑	mAP_{50}↑
✓			16.4	64.00	26.1	42.5
✓	✓		13.9	90.70	27.5	45.1
✓	✓	✓	15.8	111.1	**28.4**	**46.7**

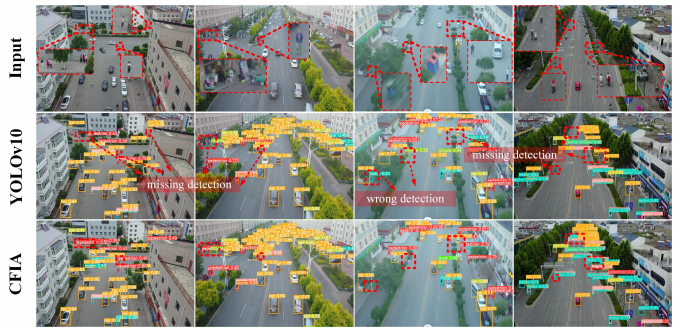

Fig. 2. Comparison of the performance of YOLOv10m and our CFIA method on the VisDrone test dataset.

By further integrating the WHPD to extract coarse-grained semantic information more closely aligned with fine-grained features, we achieve additional improvements in model performance (mAP: 27.5%→28.4%, mAP_{50}: 45.1%→46.7%). Compared to the baseline, our CFIA method has fewer parameters. Although there is a minor increase in GFLOPs, the significant performance improvements justify this trade-off.

D. Visualization

Figure 2 demonstrates the effectiveness of our method compared to the baseline on the VisDrone test dataset. It can be observed from the red boxes in the first and third columns that YOLOv10 missed detection of some distant "people" and wrong detection of "trash cans" as "motor". In contrast, our method enhances the expression ability of the feature map for small objects by aligning coarse and fine-grained features, and more small objects can be discovered, which effectively improves the detection performance of small objects.

IV. CONCLUSION

In this study, we introduce a novel plug-and-play approach to effectively enhance UAV object detection by interactively aligning coarse and fine-grained features. Firstly, we propose the WHPD to obtain high-frequency information to enhance the object boundary information, reducing the critical fine-grained information lost. Then, we propose FRIAS for interactively aligning coarse-grained and fine-grained features, which enhances the feature map's representative capability for small objects by enhancing their correlation. The large number of experiments demonstrates the effectiveness of our CFIA method in UAV object detection, which not only reduces the model parameters but also obtains competitive performance in comparison with other state-of-the-art methods.

REFERENCES

[1] A. Masadeh, M. Alhafnawi, H. A. B. Salameh, A. Musa, and Y. Jararweh, "Reinforcement learning-based security/safety uav system for intrusion detection under dynamic and uncertain target movement," *IEEE Transactions on Engineering Management*, 2022.

[2] I. Martinez-Alpiste, G. Golcarenarenji, Q. Wang, and J. M. Alcaraz-Calero, "Search and rescue operation using uavs: A case study," *Expert Systems with Applications*, vol. 178, p. 114937, 2021.

[3] S. Asadzadeh, W. J. de Oliveira, and C. R. de Souza Filho, "Uav-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives," *Journal of Petroleum Science and Engineering*, vol. 208, p. 109633, 2022.

[4] C. Lang, G. Cheng, B. Tu, C. Li, and J. Han, "Base and meta: A new perspective on few-shot segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 669–10 686, 2023.

[5] J.-H. Kim, N. Kim, and C. S. Won, "High-speed drone detection based on yolo-v8," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.

[6] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[8] J. Li, P. Tian, R. Song, H. Xu, Y. Li, and Q. Du, "Pcvit: A pyramid convolutional vision transformer detector for object detection in remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[9] G. Yang, J. Lei, H. Tian, Z. Feng, and R. Liang, "Asymptotic feature pyramid network for labeling pixels and regions," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[10] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[11] X. Ma, M. Ma, C. Hu, Z. Song, Z. Zhao, T. Feng, and W. Zhang, "Log-can: local-global class-aware network for semantic segmentation of remote sensing images," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[12] Z. Zhou and Y. Zhu, "Rafpn: Relation-aware feature pyramid network for dense image prediction," *IEEE Transactions on Multimedia*, 2024.

[13] M. Hong, S. Li, Y. Yang, F. Zhu, Q. Zhao, and L. Lu, "Sspnet: Scale selection pyramid network for tiny person detection from uav images," *IEEE geoscience and remote sensing letters*, vol. 19, pp. 1–5, 2021.

[14] D. Zhang and D. Zhang, "Wavelet transform," *Fundamentals of image data mining: Analysis, Features, Classification and Retrieval*, pp. 35–44, 2019.

[15] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," *Advances in neural information processing systems*, vol. 31, 2018.

[16] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[17] L. Sun, Z. Chen, Q. J. Wu, H. Zhao, W. He, and X. Yan, "Ampnet: Average-and max-pool networks for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4321–4333, 2021.

[18] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] E. J. Kirkland and E. J. Kirkland, "Bilinear interpolation," *Advanced computing in electron microscopy*, pp. 261–263, 2010.

[20] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 794–16 805.

[21] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[22] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *arXiv preprint arXiv:2405.14458*, 2024.

[23] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. 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*, 2019, pp. 0–0.

[24] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4145–4153.

[25] A. Coluccia, A. Fascista, L. Sommer, A. Schumann, A. Dimou, D. Zarpalas, and N. Sharma, "Drone-vs-bird detection grand challenge at icassp2023," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–2.

[26] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2012 (voc2012) results. 2012 http://www. pascal-network. org/challenges," in *VOC/voc2012/workshop/index. html*, 2012.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[28] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 002–21 012, 2020.

[29] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[30] Y. Li, Y. Wang, Z. Ma, X. Wang, and Y. Tang, "Sod-uav: Small object detection for unmanned aerial vehicle images via improved yolov7," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 7610–7614.

[31] A. Meethal, E. Granger, and M. Pedersoli, "Cascaded zoom-in detector for high resolution aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2046–2055.

[32] B. Du, Y. Huang, J. Chen, and D. 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*, 2023, pp. 13 435–13 444.

[33] S. Huang, S. Ren, W. Wu, and Q. Liu, "Discriminative features enhancement for low-altitude uav object detection," *Pattern Recognition*, vol. 147, p. 110041, 2024.

[34] Z. Liu, G. Gao, L. Sun, and Z. Fang, "Hrdnet: High-resolution detection network for small objects," in *2021 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2021, pp. 1–6.

[35] N. Yin, C. Liu, R. Tian, and X. Qian, "Sdpdet: Learning scale-separated dynamic proposals for end-to-end drone-view detection," *IEEE Transactions on Multimedia*, 2024.

[36] Y. Hui, J. Wang, and B. Li, "Stf-yolo: A small target detection algorithm for uav remote sensing images based on improved swintransformer and class weighted classification decoupling head," *Measurement*, vol. 224, p. 113936, 2024.

[37] C. Yang, Z. Huang, and N. 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*, 2022, pp. 13 668–13 677.

[38] D.-L. Nguyen, X.-T. Vo, A. Priadana, and K.-H. Jo, "Car detector based on yolov5 for parking management," in *Conference on Information Technology and its Applications*. Springer, 2023, pp. 102–113.

[39] M. Kassab, R. A. Zitar, F. Barbaresco, and A. E. F. Seghrouchni, "Drone detection with improved precision in traditional machine learning and less complexity in single shot detectors," *IEEE Transactions on Aerospace and Electronic Systems*, 2024.

[40] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464–7475.