A Method for Detecting Aircraft Small Targets in Remote Sensing Images by Using CNNs Fused With Handcrafted Features

Lijian Yu, Xiyang Zhi, Shuqing Zhang, Shikai Jiang, Jianming Hu, Wei Zhang, and Yuanxin Huang

***Abstract*— Aircraft target detection is a challenging task in remote sensing images, especially for aircraft small target detection. The most advanced object detection framework currently processes all information in the image uniformly through a deep neural network. In the past, in the process of detecting aircraft small targets, the feature extraction process was carefully designed, and handcrafted features were derived from expert knowledge or historical data, which included prior knowledge that was conducive to object detection. Embedding prior features into deep neural networks can enhance the saliency of target information and improve the detection performance of the model. Accordingly, this letter proposes a handcrafted feature fusion stream (HFFS) for embedding prior knowledge. We obtain handcrafted features based on the grayscale co-occurrence matrix and edge extraction operator and generate an attention map in deep convolutional neural networks (CNNs) to achieve the fusion of handcrafted feature maps and high-level feature maps in deep convolutional networks. The experimental results show that using HFFS on the baseline model improves the detection performance of the model for aircraft small targets. Compared with the baseline model, our detection model achieves improvements of 1.1% AR, 1.6% AP@0.5, and 1.6% AP@0.5:0.95 in the proposed dataset.**

***Index Terms*— Aircraft target, feature fusion, handcrafted feature, remote sensing, small object.**

I. Introduction

REMOTE sensing technology is constantly developing and has been widely applied in various fields such as military, agriculture, and transportation. Remote sensing image object detection, as a fundamental issue, is the foundation of various remote sensing applications. Before the introduction of RCNN in 2014, in the mainstream methods of remote sensing image object detection, the quality of manually designed features played a decisive role in the detection performance of the final model. The process of handcrafted feature design is deeply influenced by visual saliency, and low-level vision and more detailed manual features are commonly used to improve positioning accuracy [1]. With the development of deep learning algorithms, large-scale datasets, and high-performance computing hardware, many methods based on deep learning have been applied to remote sensing image object detection.

Manuscript received 1 March 2024; revised 26 April 2024; accepted 16 May 2024. Date of publication 21 May 2024; date of current version 30 May 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61975043 and Grant 62101160. *(Corresponding author: Xiyang Zhi.)*

The authors are with the Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: yulijian1994@163.com; zhixiyang@hit.edu.cn; SQ.z@hit.edu.cn; jiangshikai@hit.edu.cn; hujianming@hit.edu.cn; wzhang@hit.edu.cn; huangyxopt@163.com).

Digital Object Identifier 10.1109/LGRS.2024.3403548

Small targets are very common in remote sensing object detection. On publicly available datasets such as COCO [2], there is still a significant performance gap between small objects and normal objects. Co-DETR [3], one of the state-of-the-art detectors, as an example, the mean average accuracy (mAP) of small objects on the COCO validation set obtained by Co-DETR is only 45.1%, lagging behind medium- and large-size objects (64.7% and 76.4%, respectively). Artificial design of feature extraction operators can emphasize the target, so we believe that prior features can be used to construct attention mechanisms to help improve the performance of DCNN in detecting aircraft small targets in remote sensing images.

In this article, we investigate the method of integrating handcrafted features into deep convolutional networks and propose a new module called handcrafted features fusion stream (HFFS). The HFFS module, as a branch, is obtained by combining handcrafted feature fusion (HFF) layers and convolutional layers. The HFFS module performs shallow processing on the handcrafted feature map and integrates information with the deep layers of the baseline network. Our HFFS module can be inserted into other convolutional neural network (CNN)-based neural networks. We used YOLOv8, one of the mainstream models, as the baseline model. The experimental results showed that compared to the baseline network, the model with HFFS improved its detection performance for aircraft small targets.

The main contributions of this article are summarized as follows.

1) A detection model HFF-YOLO based on feature fusion is proposed, which uses handcrafted features to improve the detection performance.
2) An HFFS branch structure is proposed, which combines handcrafted features and automatic learning features through an attentional mechanism.
3) A new aircraft small target dataset has been proposed, which includes multiple aircraft categories and scene categories.

II. Related Work

A. Aircraft Detection in Remote Sensing Images

Before 2014, the main steps of remote sensing object detection methods included proposal box generation, feature vector extraction, and region classification. Common handcrafted features include geometric features, texture features,

moment features, scanning statistical features, scale-invariant features [4], and HOG features [5]. These feature extraction operators are a summary of a large number of experiments and practical experiences and have been widely applied and confirmed in practice. At present, the mainstream algorithm for object detection in remote sensing images is based on CNN, and the feature extraction process is completely automated by deep neural networks. Anchor-based methods, such as R-CNN family methods [6], YOLO family methods [7], [8], and RetinaNet [9], achieve aircraft detection in remote sensing images by predicting anchor box belonging to the target. Some methods predict key points belonging to the bounding box for aircraft detection, such as [10], However, there are dense and symmetrically distributed aircraft in remote sensing images, and clustering these key points is difficult.

For aircraft small target detection, Wang et al. [11] modeled the bounding boxes as 2-D Gaussian distributions and proposed a new evaluation metric using Wasserstein distance for small object detection. Yuan et al. [12] proposed coarse-to-fine RPN (CRPN) and the conventional detection head with a feature imitation (FI) branch to ensure high-quality proposals and region representations for small objects.

YOLOv8 was released in January 2023 by Ultralytics [13], the company that developed YOLOv5. YOLOv8 uses a C2f module (cross-stage partial bottleneck with two convolutions), decoupled head, CIoU [14], and DFL. [15] loss functions for bounding box loss and binary cross entropy for classification loss. These improvements have improved object detection performance, particularly when dealing with smaller objects.

B. Feature Fusion in Object Detection

There is currently limited research on the fusion of manually designed features and deep learning automatically extracted features. Zhang and Zhang [16] have conducted a preliminary discussion on the possibility of injecting traditional handcrafted features into CNN, and experiments have shown that injecting handcrafted features into CNN models can effectively improve classification accuracy.

The fusion of automatically extracted features at different scales has been extensively studied and used in several detection methods, such as FPN [17], PANet [18], DetectoRS [19], CO-DETR [3], AAHRH [20], YOLOv5 [21], and YOLOv8 [21]. The main idea of feature fusion in these methods is to fuse features of different scales by feature pyramid or by attention weighting. The objects incorporated in these studies are all automatically extracted features.

The mainstream object detection methods tend to learn features in data automatically, but the fusion of automatically extracted features and handcrafted features is relatively less studied. However, in some cases, combining handcrafted features with those learned from deep learning models can further improve detection performance. In this letter, feature fusion is carried out from another perspective. Handcrafted features are introduced into the fusion step, and this method can coexist with FPN, PAN, and other methods for fusing automatically extracted features.

Zhang et al. [22] proposed a feature fusion method that combines HOG features with principal component analysis and deep learning fully connected layers to improve the ship classification performance of the model. Saba et al. [23] concatenated the features obtained from VGG-19 with HOG and LBP features and used multiple classifiers for brain tuner classification. Liu et al. [24] proposed a spectral–spatial fusion method to improve the classification accuracy for multispectral and hyperspectral images.

The closest research to this letter is literature [25], where researchers proposed a two-stream CNN architecture for semantic segmentation, called GSCNN. Utilize multitasking learning to achieve the supervision of handcrafted features on object boundary prediction, thereby improving the accuracy of semantic segmentation. Our proposed HFFS is also a two-stream architecture that utilizes the saliency of handcrafted features to construct an attention mechanism and improve the detection performance of the baseline model.

III. Proposed Method

This letter proposes a method for fusing handcrafted features with automatic learning features in deep convolutional networks. We adopt the YOLOv8 as the baseline network, and each module is described in detail next.

A. Network Structure

We use YOLOv8 as the baseline model. We propose an aircraft small target detection network HFF-YOLO that can fuse features, as shown in Fig. 1. In addition to the original image as input, we use the feature extraction operator to extract the feature maps of the handcrafted feature and downsample to the same scale of the deep learning feature map after normalization. There are three HFF layers in the HFFS structure, corresponding to the number of detection heads in the baseline model. The feature maps after each layer of HFF processing are used as the partial input of the next fusion and the partial input of the detection head. The network architecture can be seen as two streams: one is the original deep convolutional network feature processing stream, and the other is the stream that processes handcrafted features. The output of the handcrafted feature stream is concatenated with the input of the original detection head as the new input of the detection head.

B. Handcrafted Features

The size of traditional features is often different. In the fusion model proposed in this letter, it is required that the traditional feature map is 2-D for convolution and other operations. In remote sensing images, there are fewer details and textures of aircraft small targets, but the contour texture is still preserved. Therefore, we obtain handcrafted feature maps based on gray-level co-occurrence matrix (GLCM) [26] and edge extract operators. The GLCM is a matrix that describes the grayscale relationship between a certain pixel in an image and adjacent pixels or pixels within a certain distance. Based on GLCM, various texture description features have been derived.

This letter selects contrast, dissimilarity, homogeneity, angular second moment (ASM), entropy, and standard deviation.

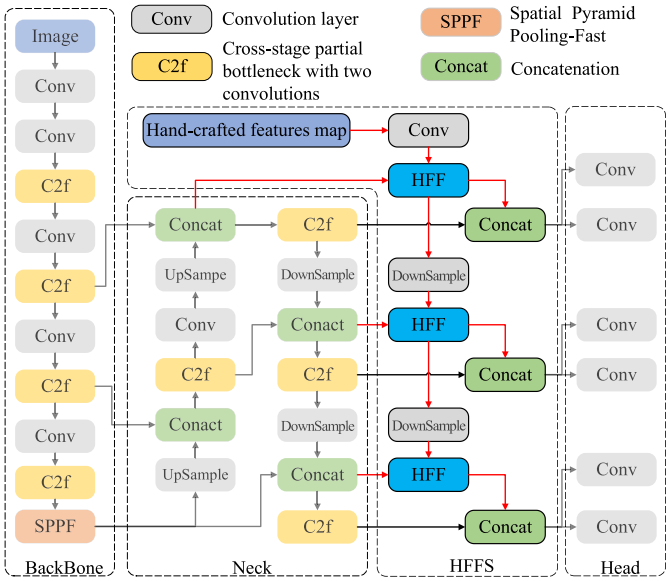

Fig. 1. Structure of HFF-YOLO.

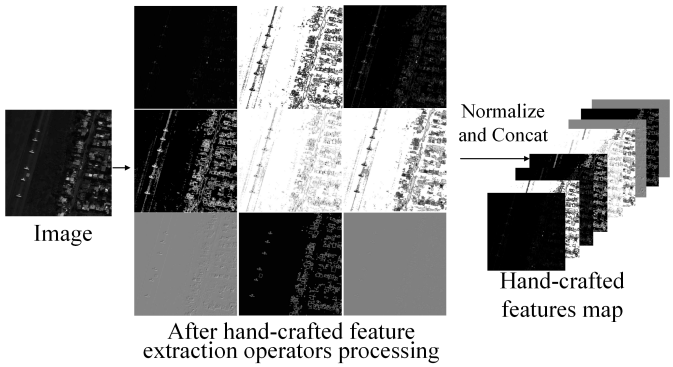

Fig. 2. Handcrafted features map.

The edge extract operator is extracted through Sobel, Canny, and Laplacian operators.

The handcrafted features map is shown in Fig. 2. The feature map of the final input HFF is the normalized result.

C. HFF Model

In the traditional self-attention mechanism, the input of query *(Q)*, key *(K)*, and value *(V)* comes from a unified feature map. In asymmetric nonlocal neural network (ANN) [27], the input of Q, K, and V in the AFNB module proposed by the author comes from different layers, and experiments have shown that this structure better integrates features of different scales. Inspired by this, we use the feature map from handcrafted feature stream as input for Q and feature map from baseline as input for K and V. We define the feature maps from the handcrafted feature stream as A_f and the feature maps from the baseline as B_f, and the standard 1×1 convolutional layer is defined as $C_{1\times1}$. Then, Q, K, and V are calculated as follows:

$$Q = C_{1\times1}(A_f), \quad K = C_{1\times1}(B_f), \quad V = C_{1\times1}(B_f). \tag{1}$$

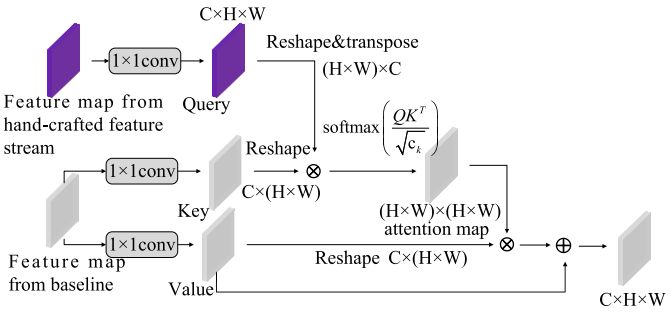

Fig. 3. Structure of HFF.

The attention map is obtained using the standard self-attention calculation method, as shown in (2) [28], where c_k represents the number of channels in K

$$\alpha = \text{soft}\max\left(\text{QK}^{\text{T}}/\sqrt{\text{c}_k}\right). \tag{2}$$

Simultaneously introducing residual connections to reduce overfitting and gradient vanishing issues, the output of the final HFF module can be expressed as

$$\text{out} = \text{Reshape}(V) * \alpha + V. \tag{3}$$

The higher the attention score of a certain point, the higher the degree of correlation between the point and Q. The HFF module can achieve attention on parts of high-level activation that are more relevant to prior knowledge. The HFF module structure is shown in Fig. 3.

IV. Experiments

A. Experimental Data and Evaluation Metrics

The AI-TOD [29] is a dataset used for small object detection, we select images that contain aircraft targets. Considering that the actual aircraft target still has flight status, we have added some small target data for flight status. The image source for the newly added data comes from WorldView, which includes different types of backgrounds such as forests, clouds, and cities. The final dataset used for this experiment consists of approximately 1600 images, with an object instance count of approximately 2100 and an image size of 640 × 640 pixels. Most of the targets in the dataset are smaller than 32 × 32 pixels.

We use average recall (AR) and average precision (AP) under different intersection over union (IoU) as evaluation indicators to measure the detector performance. AR and AP are calculated based on precision *(P)* and recall *(R)*

$$\text{AP} = \int_0^1 P(R)d(R), \quad \text{AR} = 2\int_{0.5}^1 R(\text{IoU})d(\text{IoU}). \tag{4}$$

B. Training and Test Details

When training the model, Adam [30] was used as the optimizer, with a batch size of 2 and a total of 100 epochs. Both the baseline model and the improved model use the OneCycleLR [31] method to accelerate the convergence of the training process. The initial learning rate is 0.001, and the maximum learning rate is 0.1. Train using an E5-2630 v3 CPU and a single RTX-4080 GPU. All detection models are trained and tested under the same hardware conditions.

TABLE I
COMPARISON OF DETECTION RESULTS ON THE DATASET

Network	Fusion method	AR@0.5:0.95	AP@0.5	AP@0.5:0.95	AP_{VS}@0.5:0.95	AP_S@0.5:0.95	AP_M@0.5:0.95
Faster_rcnn [6]	FPN	53.1%	70.4%	50.8%	5.5%	58.8%	78.0%
Cascade R-CNN [32]	FPN	54.3%	71.7%	52.0%	6.5%	59.4%	79.6%
Co-DETR [3]	FPN	36.2%	58.3%	30.3%	2.5%	24.7%	56.1%
Retinanet [9]	FPN	56.4%	72.9%	52.4%	3.6%	56.3%	82.1%
CFINet [12]	FPN	53.5%	72.5%	50.6%	8.0%	54.2%	77.0%
DetectoRS [19]	RFP	54.7%	72.3%	52.6%	7.3%	60.1%	80.2%
NWD [11]	RFP	56.2%	73.6%	53.7%	8.7%	61.5%	80.5%
PANet [18]	PAN	52.5%	71.4%	50.4%	4.9%	57.1%	78.5%
YOLOv5 [21]	PAN	58.0%	72.9%	55.7%	12.0%	62.6%	80.8%
YOLOv8 [21]	PAN	64.4%	77.8%	60.9%	17.1%	68.4%	85.7%
HFF-YOLO (3)	PAN+HFFS	63.4%	77.3%	60.6%	17.4%	68.7%	83.8%
HFF-YOLO (5)	PAN+HFFS	64.2%	77.9%	61.2%	17.9%	68.5%	85.4%
HFF-YOLO (7)	PAN+HFFS	64.9%	78.6%	62.0%	18.6%	69.9%	85.3%
HFF-YOLO (9)	PAN+HFFS	**65.5%**	**79.4%**	**62.5%**	**19.5%**	**69.9%**	**86.0%**

The number after "HFF-YOLO" represents the number of hand-crafted features selected.

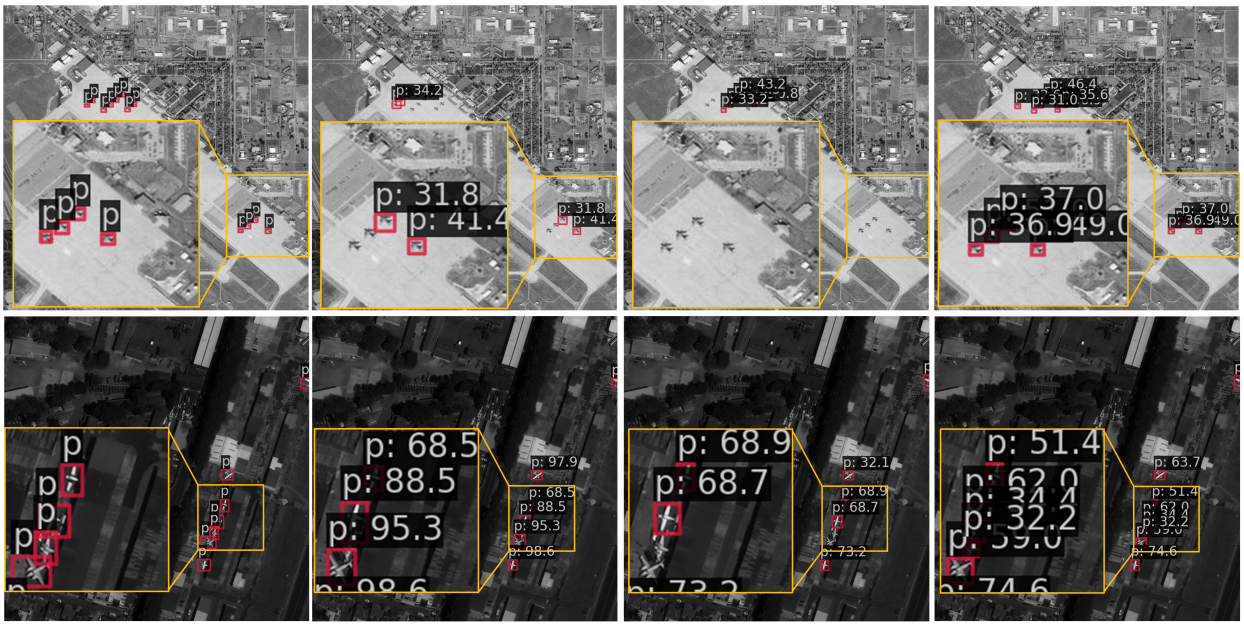

Fig. 4. Example results on the test set. (a) Ground truth. (b) NWD. (c) YOLOv8. (d) HFF-YOLO.

C. Main Results and Analysis

We compared the proposed method with current mainstream methods and the baseline model using the dataset proposed in this article, and the experimental results are shown in Table I.

The experimental results show that compared with the baseline model, the model detection performance improved after adding the HFFS module, and it also has competitive performance compared to current mainstream models. YOLOv8 uses the idea of PAN for feature fusion, but it does not conflict with the method proposed in this article. Based on YOLOv8, the method proposed in this article can further improve the detection performance. YOLOv8 improves the optimal results by 1.1% in AR, 1.6% in AP@0.5, and 1.6% in AP@0.5:0.95. Compared with the mainstream model, HFF-YOLO not only fuses automatically extracted features of different scales but also uses an attention mechanism to introduce handcrafted features into the feature fusion process so that the model has the best detection ability for small targets. When the number of handcrafted features exceeds 7, HFF-YOLO achieves better performance than other models.

In mainstream datasets, such as the COCO dataset, any pixel count below 32×32 is considered a small target. In order to evaluate the detection performance of the algorithm on small targets in more detail, this article regards the target pixels in the range of $(0, 16 \times 16)$ as a very small target, $(16 \times 16, 32 \times 32)$ as a small target, $(32 \times 32, 96 \times 96)$ as a medium target, and $(96 \times 96, +\infty)$ as a large target. It is important to note that there are no large targets in the data used in this experiment. Table I also shows the performance of different algorithms on very small target, small target, and medium target in AP@0.5 0.95. In Table I, AP_{VS}@0.5:0.95, AP_S@0.5:0.95, and AP_M@0.5:0.95 correspond to very small target, small target, and medium target, respectively. The detection performance of HFF-YOLO on very small targets is better than that of the baseline model, improving by 2.4%.

To further analyze the effectiveness of the method proposed in this article, the results of YOLOv8 and HFF-YOLO are compared. As shown in Fig. 4, due to the attention of handcrafted features, the features of small targets are not easily submerged by background information, and HFF-YOLO can detect targets even when the targets are very small.

More experiments and results are available at https://github.com/heitongYulj/HFF-YOLO.

V. Conclusion

In the past, there have been few studies on how to use handcrafted features to enhance aircraft small tlárget detection. We propose a module HFFS that integrates handcrafted features and deep learning automatic features, which is embedded in the backend of deep convolutional networks in the form of branches. HFFS utilizes the prior properties of handcrafted features to achieve attention to high-level features of deep convolutional networks, which is a knowledge and data-driven approach. Compared with the baseline model, our detection model shows higher performance in the aircraft small target detection dataset proposed in this letter, with an improvement of 1.1% AR, 1.6% AP@0.5, and 1.6% AP@0.5:0.95. For very small targets in the dataset, AP@0.5:0.95 increased by 2.4%.

References

[1] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proc. IEEE*, vol. 111, no. 3, pp. 257–276, Mar. 2023, doi: 10.1109/JPROC.2023.3238524.

[2] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[3] Z. Zong, G. Song, and Y. Liu, "DETRs with collaborative hybrid assignments training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6725–6735, doi: 10.1109/ICCV51070.2023.00621.

[4] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.* Kerkyra, Greece: IEEE, Sep. 1999, pp. 1150–1157, doi: 10.1109/ICCV.1999.790410.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[7] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[8] Y. Zhang, M. Ye, G. Zhu, Y. Liu, P. Guo, and J. Yan, "FFCA-YOLO for small object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–15, 2024, doi: 10.1109/TGRS.2024.3363057.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[10] Z. Zhu, X. Sun, W. Diao, K. Chen, G. Xu, and K. Fu, "AOPDet: Automatic organized points detector for precisely localizing objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5606816, doi: 10.1109/TGRS.2021.3093557.

[11] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.

[12] X. Yuan, G. Cheng, K. Yan, Q. Zeng, and J. Han, "Small object detection via coarse-to-fine proposal generation and imitation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6294–6304, doi: 10.1109/ICCV51070.2023.00581.

[13] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," 2023, *arXiv:2304.00501*.

[14] 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 *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 12993–13000, doi: 10.1609/aaai.v34i07.6999.

[15] X. Li et al., "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Proc. NeurIPS*. Red Hook, NY, USA: Curran Associates, 2020, pp. 21002–21012.

[16] T. Zhang and X. Zhang, "Injection of traditional hand-crafted features into modern CNN-based models for SAR ship classification: What, why, where, and how," *Remote Sens.*, vol. 13, no. 11, p. 2091, May 2021, doi: 10.3390/rs13112091.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[19] S. Qiao, L.-C. Chen, and A. Yuille, "DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10208–10219.

[20] T. Shi et al., "Adaptive feature fusion with attention-guided small target detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5623116, doi: 10.1109/TGRS.2023.3323409.

[21] G. Jocher, A. Chaurasia, and J. Qiu. (2023). *Ultralytics YOLO*. Accessed: Feb. 29, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[22] T. Zhang et al., "HOG-ShipCLSNet: A novel deep learning network with HOG feature fusion for SAR ship classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5210322, doi: 10.1109/TGRS.2021.3082759.

[23] T. Saba, A. Sameh Mohamed, M. El-Affendi, J. Amin, and M. Sharif, "Brain tumor detection using fusion of hand crafted and deep learning features," *Cogn. Syst. Res.*, vol. 59, pp. 221–230, Jan. 2020, doi: 10.1016/j.cogsys.2019.09.007.

[24] S. Liu et al., "A shallow-to-deep feature fusion network for VHR remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2022, Art. no. 5410213, doi: 10.1109/TGRS.2022.3179288.

[25] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-SCNN: Gated shape CNNs for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5228–5237, doi: 10.1109/ICCV.2019.00533.

[26] M. Hall-Beyer. (2017). *GLCM Texture: A Tutorial V. 3.0 March 2017*. Accessed: Feb. 29, 2024. [Online]. Available: http://hdl.handle.net/1880/51900

[27] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 593–602, doi: 10.1109/ICCV.2019.00068.

[28] D. Wang et al., "Advancing plain vision transformer toward remote sensing foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607315, doi: 10.1109/TGRS.2022.3222818.

[29] J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia, "Tiny object detection in aerial images," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 3791–3798, doi: 10.1109/ICPR48806.2021.9413340.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[31] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*. Bellingham, WA, USA: SPIE, May 2019, pp. 369–386, doi: 10.1117/12.2520589.

[32] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162, doi: 10.1109/CVPR.2018.00644.