Integrally Mixing Pyramid Representations for Anchor-Free Object Detection in Aerial Imagery

Cong Zhang, *Graduate Student Member, IEEE*, Jun Xiao, Cuixin Yang, Jingchun Zhou, Kin-Man Lam, *Senior Member, IEEE*, and Qi Wang, *Senior Member, IEEE*

***Abstract*— Anchor-free object detectors have recently received increasing research attention in the field of aerial scene object detection, due to their high flexibility and practicality. Anchor-free detectors typically depend on the feature pyramid network (FPN) to alleviate the challenge of significant variations in object scales in aerial contexts. Despite establishing a multiscale feature pyramid, existing FPN-based methods treat each aerial object as an indivisible entity solely managed by a single-scale representation. However, they fail to take into account the distinct characteristics of various components within an instance. To this end, this letter proposes a novel anchor-free detector, namely IMPR-Det, which can integrally mix multiscale pyramid representations for different components of an instance, thus boosting the fine-grained object representation capability. Specifically, IMPR-Det fundamentally introduces a more advanced detection head with an adaptive routing mechanism for pixel-level multiscale feature assignment, instead of previous instance-level assignment. Experimental results demonstrate the superiority of the proposed method over its counterparts, in terms of both accuracy and efficiency, for object detection in aerial images.**

***Index Terms*— Adaptive detection head, aerial images, anchor-free object detection, deep learning, pyramid representations.**

I. Introduction

OBJECT detection in aerial imagery is a fundamental task for both computer vision and remote sensing communities, which aims to locate objects of interest and identify their categories from a bird's eye view. With the rapid development of deep learning and the increasing availability of large-scale aerial images, modern convolutional neural network (CNN)-based aerial object detectors have shown promising potential for real-world applications, such as environmental monitoring, urban planning, and intelligent transportation [1], [2], [3], [4].

Regarding whether anchor boxes should be generated in advance, recent aerial object detection methods can be mainly divided into two categories: anchor-based and anchor-free detectors. The former methods [3], [4], [5], [6], [7], [8], [9], [10], [11] are usually based on Faster R-CNN [12] or RetinaNet [13]. They first tile numerous rectangles as predefined anchors on the image and then predict the category and the refined coordinates of these anchors to produce the final detection results. For example, ASSD [8] aimed to alleviate the issue of spatial- and scale-misalignment between the predefined anchors and extracted features for improved accuracy. Although anchor-based detectors have achieved impressive progress, their detection performance has shown to be highly vulnerable to anchor-related configurations, significantly limiting their applicability [14], [15], [16], [17].

To overcome this problem, recent anchor-free aerial detectors [18], [19], [20], [21], [22] have eliminated proposal and anchor generation and directly localize objects by point-level calculations, which have garnered tremendous research attention due to their high generalizability and efficiency. However, in the absence of anchor-box prior constraints, existing anchor-free detectors suffer from a degradation in detection accuracy, especially in challenging aerial scenarios. On the one hand, compared to natural-scene objects, aerial instances usually exhibit more pronounced variations in scale across diverse categories [23], [24]. This issue further exacerbates the representation vulnerability of anchor-free detectors. On the other hand, feature pyramid network (FPN) [13] can alleviate the aforementioned issue due to its capability of establishing hierarchical representations [14]. This enables the anchor-free detection head to allocate different pyramid levels according to the object sizes. Then, each instance is separately represented and managed by the corresponding pyramid level with specific feature resolution. However, existing methods always treat each instance as a holistic indivisible region, neglecting the distinct characteristics of different components (or subregions) within an instance. Therefore, as illustrated in Fig. 1(a), in the conventional anchor-free detection head, only the single-scale representation from the FPN is utilized and contributes to the final results.

This letter introduces a conceptually novel strategy, namely Integrally Mixing Pyramid Representations, to address the above issues. Based on this, we propose a novel anchor-free detector, IMPR-Det. As depicted in Fig. 1, different from previous anchor-free detectors, IMPR-Det incorporates an adaptive detection head, which can adaptively select multiscale representations from multiple pyramid stages, rather than a single stage, for distinct components of an instance. For example, in Fig. 1(b), the features representing the wings, nose, tail, and fuselage of an airplane are adaptively derived from different pyramid levels of FPN, instead of the same one as in Fig. 1(a). This approach significantly benefits representation generalization and diversity. Then, the integral combination of all pixel-level representations related to the instance constitutes the output feature for predictions. Specifically, to achieve this objective, inspired by the dynamic network mechanism [25], we devise an adaptive routing space in the anchor-free detection head of our IMPR-Det, where each node is endowed with three optional forward paths. In this way, different components of an object can be adaptively

Manuscript received 8 April 2024; accepted 8 May 2024. Date of current version 30 May 2024. *(Corresponding author: Cong Zhang.)*

Cong Zhang, Jun Xiao, Cuixin Yang, and Kin-Man Lam are with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: cong-clarence.zhang@connect.polyu.hk).

Jingchun Zhou is with the School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China.

Qi Wang is with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China.

Digital Object Identifier 10.1109/LGRS.2024.3404481

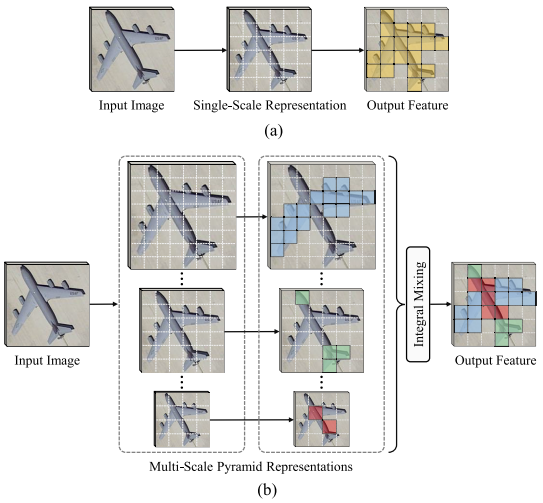

Fig. 1. Comparisons between (a) traditional detection head used in previous anchor-free object detectors and (b) adaptive detection head in the proposed IMPR-Det. The detection predictions in the former are only based on single-scale representations, whereas the latter leverages pyramid representations across multiple scales, which are integrally mixed to form a pixel-level combination to serve as the output feature for prediction.

associated with appropriate pyramid representations. Moreover, our proposed adaptive routing mechanism enjoys two crucial advantages: *data-dependency* and *fine-granularity*. "Data-dependency" means that it can dynamically adjust the activated paths in accordance with the content of various input images. "Fine-granularity" refers to the routing process being pixel-centric within potential instances, thus facilitating fine-grained object representations that are highly compatible with anchor-free detectors. Extensive experiments on different datasets have demonstrated the superiority of the proposed method in terms of both accuracy and efficiency. Overall, the contributions of this work can be summarized as follows.

1) A novel anchor-free detector, namely IMPR-Det, is proposed for object detection in aerial images. It is characterized by a lightweight adaptive detection head. Fundamentally, its design principle lies in integrally mixing multiscale pyramid features from the FPN, which aims to boost the object representation capability and benefit both detection accuracy and computational efficiency.
2) An adaptive routing mechanism is proposed to equip the detection head by dynamically determining the optimal pyramid representation levels for different components or subregions of an aerial instance. This mechanism can consistently strengthen anchor-free object detectors.

II. Methodology

In this section, we first provide an overview of IMPR-Det and introduce the proposed adaptive detection head. Then, we elaborate on its core mechanism, that is, adaptive routing, for fine-grained pyramid representation aggregation. Finally, we detail the optimization process of our proposed IMPR-Det.

A. Anchor-Free Adaptive Detection Head

As depicted in Fig. 2, our IMPR-Det consists of three crucial components: 1) ResNet as the backbone to extract multiple features at different scales; 2) FPN to establish a feature pyramid; and 3) a novel anchor-free adaptive detection head for final detections, composed of several convolutions and

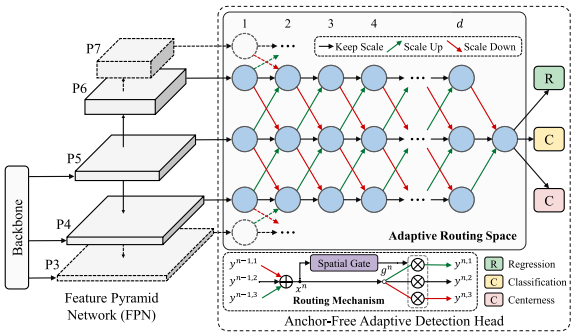

Fig. 2. Illustration of the adaptive routing process in the anchor-free detection head of our proposed IMPR-Det. P_3–P_7 represent the adjacent FPN features with different scales. Each blue-filled circle refers to an adaptive routing node, with up to three alternative subsequent paths, that is, keeping scale, scaling up, and scaling down. Taking P_5 as an example, the routing nodes corresponding to P_4 and P_6 are visible, while those for P_2 and P_7 are invisible.

shared across all FPN scales. Previous anchor-free detection heads [1], [8], [20], [23], [26] simply operate on the FPN features at each scale separately. However, at any moment, only one single-scale pyramid representation with insufficient feature responses is responsible for instance-level predictions. This essentially limits semantic interactions between low- and high-resolution features. Therefore, to address this issue, we propose a novel adaptive detection head, shown in Fig. 2, which aims to integrally mix multiscale pyramid features *at the pixel level* to enhance the fine-grained representation capability. Specifically, for the pyramid feature P_i from the ith FPN stage, we first define an adaptive routing space in the detection head, where d represents its depth and is set to 2 by default. Notably, d is usually crucial to the performance of anchor-free detectors, whose effects will be investigated in the experiments. Then, each routing node can adaptively select its next path for each location or pixel in P_i from three different scale variations, that is, keeping scale, scaling up, and scaling down. Finally, following [13] and [14], each feature pyramid level i will be equipped with the proposed adaptive head, and the final detections can be yielded by merging predictions at all levels. The configuration of the FPN also remains the same as the previous practices, and $\{P_3, P_4, P_5, P_6, P_7\}$ are adopted to construct the feature pyramid, that is, $i \in \{3, 4, 5, 6, 7\}$.

B. Adaptive Routing

Integrally mixing multiscale pyramid representations requires determining the originating FPN level for each pixel. To facilitate this, we propose an adaptive routing mechanism to predict fine-grained element-wise routing paths. As illustrated in Fig. 2, given a node n within the routing space, we define the accumulation of multiple input features as $x^n = \{x_m^n\}_{m=1}^M$ and a set of paths as $\mathcal{P} = \{p_l^n(\cdot)|l \in \{1, \ldots, L\}\}$, where M refers to the total count of pixel-level locations while L indicates the number of spatial gates for each node, typically equal to the number of predefined selectable adjacent paths. The output of the lth spatial gate $\mathcal{G}_l^n(\cdot)$ for the location m can be formulated as follows:

$$g_m^{n,l} = \mathcal{G}_l^n\left(x_m^n; \theta_l^n\right) \tag{1}$$

where θ_l^n represents the parameters of the partial network associated with the lth gate, shared by all locations. The gating output $g_m^{n,l}$ can dynamically regulate the path state (on or off), and we further propagate it into the continuous space, that is, $g_m^{n,l} \in [0, 1]$, instead of the discrete space. In this way,

TABLE I

COMPARISON OF IMPR-DET IN DIFFERENT CONFIGURATIONS WITH OTHER DETECTORS, ‡ REPRESENTING OUR REIMPLEMENTED VERSION WITH THE FPN

Detectors	FPN	Adaptive	mAP_{50}	mAP_{75}	$mAP_{50:95}$	Params	FLOPs (Avg)	FLOPs (Max)	FLOPs (Min)
Faster R-CNN [12]	×	×	54.7	30.7	30.5	70.3M	116.5G	116.5G	116.5G
Faster R-CNN‡ [12]	✓	×	61.7	34.7	34.5	41.2M	134.5G	134.5G	134.5G
RetinaNet [13]	✓	×	57.3	39.6	36.9	36.5M	133.1G	133.1G	133.1G
ATSS [15]	✓	×	59.8	32.2	32.2	**31.9M**	126.5G	126.5G	126.5G
FCOS [14] (Baseline)	✓	×	59.6	34.9	34.5	**31.9M**	123.6G	123.6G	123.6G
IMPR-Det-Fixed ($d = 2$)	✓	×	60.6	35.2	35.3	32.0M	91.5G	91.5G	91.5G
IMPR-Det ($d = 2$)	✓	✓	62.9	39.1	38.0	32.0M	**73.6G**	**90.5G**	**66.4G**
IMPR-Det ($d = 4$)	✓	✓	64.6	41.1	39.7	36.9M	87.7G	105.9G	73.7G
IMPR-Det ($d = 8$)	✓	✓	**64.8**	**43.7**	**41.4**	46.7M	105.3G	135.0G	86.7G

TABLE II

COMPARISON OF THE COEFFICIENT λ UNDER DIFFERENT SETTINGS

Detectors	λ	mAP_{50}	mAP_{75}	$mAP_{50:95}$	FLOPs (Avg)	FLOPs (Max)	FLOPs (Min)
FCOS (Baseline)	−−	59.6	34.9	34.5	123.6G	123.6G	123.6G
IMPR-Det	0	**63.5**	38.7	37.9	185.0G	196.4G	170.4G
	0.1	62.9	**39.1**	**38.0**	73.6G	90.5G	66.4G
	0.5	62.6	37.0	37.0	64.4G	74.1G	62.7G
	0.8	62.6	36.4	36.6	**63.5G**	**69.8G**	**62.4G**

the gating output can indicate the estimated probability of the path being activated, and weight the pyramid representations processed by different paths, formulated as follows:

$$y_m^n = \left\{ y_m^{n,l} \middle| y_m^{n,l} = g_m^{n,l} \cdot p_l^n\left(x_m^n\right), l \in \Omega_m \right\} \tag{2}$$

where $\Omega_m = \{l | g_m^{n,l} > 0, l \in \{1, \ldots, L\}\}$. It is worth noting that our proposed routing mechanism allows for the simultaneous activation of multiple paths to serve a single location.

Additionally, the spatial gate $\mathcal{G}_l^n(\cdot)$ is instantiated by a lightweight convolutional network incorporating a softmax activation function. As mentioned above, three different forward paths are provided for each routing node. The scaling factor for both the "scaling up" and "scaling down" paths is restricted to 2, while the difference lies in whether up-sampling or down-sampling is performed according to the gating output. Furthermore, we introduce spatially sparse convolution to each path for higher inference efficiency. Concretely, a global spatial mask, denoted as $\mathcal{S}^{n,l}$, is first produced by performing max-pooling with a kernel size of 3×3 on the gating output $g^{n,l}$. Then, all positive values in $\mathcal{S}^{n,l}$ are further quantified to one, generating the quantified version $\widetilde{\mathcal{S}}^{n,l}$, based on which, sparse convolution operates only on those positive locations, significantly reducing computational complexity. In this way, pyramid representations are integrally aggregated for final predictions, with higher discrimination and robustness.

C. Optimization With Computational Resource Constraint

In essence, the proposed adaptive routing mechanism tends to activate more or even all paths to improve the final detection performance. However, aerial scenarios typically entail limited computational resources, thereby encouraging a tradeoff between detection accuracy and efficiency. To this end, we introduce the computational resource constraint to disable as many paths as possible with minimal accuracy degradation. Specifically, the computational complexity for the lth path in the node n is denoted as $\mathcal{C}^{n,l}$, and the total resource constraint of the single node n is formulated as follows:

$$\mathcal{R}^n = \frac{1}{M} \sum_m \sum_l \mathcal{C}^{n,l} \mathcal{S}_m^{n,l} \tag{3}$$

which can be further considered as a regularization term to guide the optimization process during the training phase. Concretely, as shown in Fig. 2, following the design of the anchor-free object detection frameworks [14], the proposed IMPR-Det involves three loss functions for classification, regression, and centerness of bounding boxes, denoted as $\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{reg}}$, and $\mathcal{L}_{\text{cen}}$, respectively. Moreover, the resource constraint $\mathcal{R}^n$ is also adopted to form a new resource loss, denoted as $\mathcal{L}_R$, to restrict the computational load, formulated as follows:

$$\mathcal{L}_R = \frac{\sum_n \mathcal{R}^n}{\sum_n \mathcal{C}^n} \in [0, 1], \quad \text{where } \mathcal{C}^n = \sum_l \mathcal{C}^{n,l}. \tag{4}$$

Ultimately, the whole detector can be end-to-end optimized during training with a joint loss $\mathcal{L}$, as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{cen}} + \lambda \mathcal{L}_R \tag{5}$$

where the coefficient $\lambda \in [0, 1]$ is introduced to flexibly regulate the computational resource consumption for achieving a desired balance between effectiveness and efficiency.

III. EXPERIMENTS AND ANALYSIS

A. Datasets, Evaluation Metrics, and Implementation Setup

To evaluate the proposed IMPR-Det, we utilize two representative datasets, namely DIOR [28] and NWPU VHR-10 [27], in the experiments. DIOR [28] is currently the largest dataset for object detection in aerial scenarios, consisting of 23 463 images in 20 categories, that is, airplane (APL), airport (APO), baseball field (BF), basketball court (BC), bridge (BR), chimney (CH), dam (DAM), expressway toll station (ETS), expressway service area (ESA), golf field (GF), ground track field (GTF), harbor (HA), overpass (OP), ship (SH), stadium (STA), storage tank (STO), tennis court (TC), train station (TS), vehicle (VE), and windmill (WM). In the experiments, 1/4, 1/4, and 1/2 of the original images were randomly selected for training, validation, and testing, respectively. NWPU VHR-10 [27] is a 10-class aerial object detection dataset, containing a total of 800 images, which are randomly divided into two subsets for training and testing, with a ratio of 3:1.

The well-known anchor-free detector FCOS [14] is leveraged as the baseline in the experiments. Unless otherwise specified, we employ ResNet50 as the default backbone network for all detectors, and all the backbones are pretrained on the ImageNet dataset. The mAPs with different IoU thresholds,

TABLE III
DETECTION ACCURACY COMPARISON OF VARIOUS AERIAL OBJECT DETECTORS ON THE DIOR DATASET

Methods	Backbone	APL	APO	BF	BC	BR	CH	DAM	ESA	ETS	GF	GTF	HA	OP	SH	STA	STO	TC	TS	VE	WM	mAP_{50}
Anchor-Based																						
RetinaNet [13]	ResNet101	53.3	77.0	69.3	85.0	44.1	73.2	62.4	78.6	62.8	78.6	76.6	49.9	59.6	71.1	68.4	45.8	81.3	55.2	44.4	85.5	66.1
CSFF [3]	ResNet101	57.2	79.6	70.1	87.4	46.1	76.6	62.7	82.6	73.2	78.2	81.6	50.7	59.5	73.3	63.4	58.5	85.9	61.9	42.9	86.9	68.0
SCRDet++ [9]	ResNet50	64.3	79.0	73.2	85.7	45.8	76.0	68.4	79.3	68.9	77.7	77.9	56.7	62.2	70.4	67.7	60.4	80.9	63.7	44.4	84.6	69.4
SB-MSN [6]	ResNeXt101	79.6	82.2	76.4	89.8	45.6	78.2	64.8	58.9	59.3	79.2	82.4	51.8	60.8	74.4	79.7	66.4	85.6	65.4	45.1	79.9	70.3
GLNet [4]	ResNet101	62.9	83.2	72.0	81.1	50.5	79.3	67.4	86.2	70.9	81.8	83.0	51.8	62.6	72.0	75.3	53.7	81.3	65.5	43.4	89.2	70.7
ASSD [8]	VGG16	85.6	82.4	75.8	89.5	40.7	77.6	64.7	67.1	61.7	80.8	78.6	62.0	58.0	84.9	76.7	65.3	87.9	62.4	44.5	76.3	**71.1**
Anchor-Free																						
CornerNet [16]	Hourglass104	58.8	**84.2**	72.0	80.8	46.4	75.3	64.3	81.6	76.3	79.5	79.5	26.1	60.6	37.6	70.7	45.2	84.0	57.1	43.0	75.9	64.9
O^2-DNet [18]	Hourglass104	61.2	80.1	73.7	81.4	45.2	75.8	64.8	81.2	76.5	79.5	79.7	47.2	59.3	72.6	70.5	53.7	82.6	55.9	49.1	77.8	68.4
CenterNet [17]	Hourglass104	84.6	73.8	90.7	82.2	46.7	80.3	54.6	72.6	74.1	68.4	83.7	27.1	60.7	51.5	91.4	54.7	87.2	50.6	55.4	85.0	68.8
SRAF-Net [20]	ResNet50	85.8	80.6	91.6	87.2	39.0	81.3	52.6	81.7	68.4	71.9	82.0	14.2	55.2	59.5	80.5	55.4	93.3	50.1	56.9	92.9	69.0
MSFC-Net [21]	ResNeSt101	85.8	76.2	74.4	90.1	44.2	78.1	55.5	60.9	59.5	76.9	73.7	49.6	57.2	89.6	69.2	76.5	86.7	51.8	55.2	84.3	70.1
IMPR-Det (Ours)	ResNet50	82.2	78.2	81.4	88.3	42.9	76.1	66.9	81.6	62.1	75.4	76.2	52.8	57.6	76.0	85.3	71.8	88.2	53.7	44.7	84.3	**71.3**

that is, mAP_{50}, mAP_{75}, and $mAP_{50:95}$ are utilized as the metrics to evaluate the detection accuracy, while the number of parameters (abbreviated as Params) and floating-point operations (FLOPs) are adopted to compare the computational efficiency of different detectors. In addition, we mainly exploit the $1\times$ training schedule and the AdamW optimizer with an initial learning rate of 0.0001, which is decreased by 0.1 at the 6th and 8th epochs. All the experiments were conducted on two NVIDIA GeForce RTX 3090 GPUs with a batch size of 8.

B. Ablation Studies

In this section, ablation studies are conducted on the DIOR validation set to demonstrate the effectiveness of some critical components or design principles of the proposed IMPR-Det.

1) Effect of the Proposed Anchor-Free Adaptive Detection Head: To illustrate the superiority of the proposed IMPR-Det, especially the effectiveness of its adaptive detection head equipped with the advanced routing mechanism, we compare IMPR-Det with its degraded version that contains a fixed head (namely IMPR-Det-Fixed), as well as previous counterparts in Table I. Most comparative methods constructed feature pyramids by the FPN. It can be observed that the FPN plays a crucial role in this task, while our IMPR-Det significantly outperforms other representative detectors, in terms of both accuracy and efficiency. In addition, since each routing node in IMPR-Det-Fixed is fixedly connected to three adjacent pyramid paths, instead of adaptively adjusting with different inputs by routing, it requires more computational resources than IMPR-Det. However, this setting cannot yield additional performance gains, which further demonstrates the effectiveness of our adaptive routing mechanism. Fig. 3 compares the multiscale representations of different detectors. It reveals that our proposed adaptive routing dynamically allocates pixel-level features from different pyramid scales to subregions or components of a single instance. The integrally mixed representations imply enhanced semantic correlations and suppressed background noise.

2) Effect of Routing Depth: Table I tabulates the effect of the depth in the detection head routing mechanism. As the depth d increases, the detection accuracy gradually improves, albeit at the cost of increased model size and computational complexity. Notably, our advanced adaptive routing mechanism keeps the average FLOPs of IMPR-Det substantially lower than the baseline, even at a maximum of d. Considering the tradeoff between detection accuracy and efficiency in aerial scenarios, we default to utilizing the most lightweight head, that is, $d = 2$, in the experiments.

3) Effect of Computational Resource Constraint: Table II explores the effect of the proposed computational resource constraint strategy by varying the key coefficient λ during optimization. It can be observed that regardless of the value assigned to λ (excluding 0), IMPR-Det consistently achieves superior accuracy compared to the baseline, while maintaining a remarkably lower computational cost. Furthermore, as suggested in (4) and (5), a larger λ typically places more emphasis on minimizing the computational load, thereby leading to fewer inference computations. For instance, when λ is set to 0.1, our IMPR-Det achieves competitive and promising results, indicating an improved 3.3% mAP_{50} over the baseline, with much fewer computations. As λ increases to 0.8, the required computational resources further decrease to 51.4% of the baseline, with the detection accuracy slightly degrading yet still significantly better. Considering this accuracy-efficiency tradeoff, we set λ to 0.1 in the experiments by default.

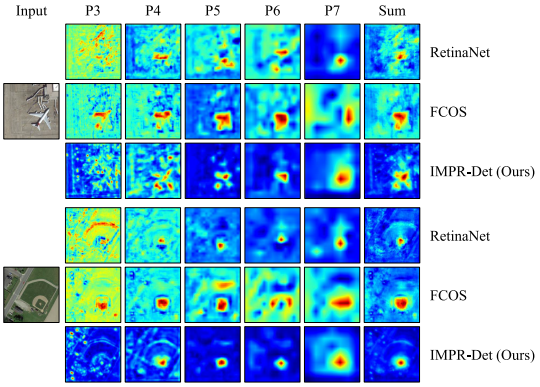

Fig. 3. Visualization of pyramid representations with diverse FPN scales generated by different object detectors, where the foreground representations of our IMPR-Det are more accurate and complete, while the background interference is significantly suppressed.

C. Comparison With State-of-the-Art Aerial Object Detectors

To further verify the superiority of the proposed IMPR-Det, we compare it with other state-of-the-art aerial object detectors, including both anchor-based and anchor-free methods, on two datasets in Tables III and IV, respectively. It can be observed from Table III that our IMPR-Det achieves the best accuracy of 71.3% mAP_{50}, surpassing the recent anchor-free aerial detector MSFC-Net by about 1.2% mAP_{50}, with a more lightweight backbone. Similarly, in Table IV, IMPR-Det consistently outperforms other counterparts across most categories, finally achieving a high accuracy of 95.0% mAP_{50}. The above experimental results demonstrate the effectiveness of the proposed strategy of "integrally mixing pyramid representations," which significantly enhances the discrimination

TABLE IV
DETECTION ACCURACY COMPARISON OF VARIOUS AERIAL OBJECT DETECTORS ON THE NWPU VHR-10 DATASET

Methods	Backbone	Airplane	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Track Filed	Harbor	Bridge	Vehicle	mAP_{50}
Anchor-Based												
RICNN [27]	AlexNet	88.4	77.3	85.3	88.1	40.8	58.5	86.7	68.6	61.5	71.1	72.6
MSCA [5]	VGG16	99.5	80.0	90.4	90.5	90.6	77.3	100.0	76.1	65.9	80.6	85.1
FMSSD [7]	VGG16	99.7	89.9	90.3	98.2	86.0	96.8	99.6	75.6	80.1	88.2	89.2
CAD-Net [10]	ResNet101	97.0	77.9	95.6	93.6	87.6	87.1	99.6	100.0	86.2	89.9	91.5
MEDNet [11]	ResNet101	99.2	94.4	82.2	98.5	95.4	95.2	98.3	88.1	75.1	89.3	**91.6**
Anchor-Free												
SRAF-Net [20]	ResNet50	94.6	83.8	72.8	97.0	88.4	92.3	99.0	63.5	54.0	89.2	83.5
DKD [22]	ResNet18	98.4	88.3	77.0	96.7	87.5	74.4	97.7	95.6	92.1	94.0	90.2
CANet [19]	ResNet101	100.0	86.0	99.3	97.3	97.8	84.8	98.4	90.4	89.2	90.3	93.3
DA^2FNet [1]	Hourglass104	99.9	89.5	87.2	98.7	95.7	95.2	99.5	99.0	89.0	93.6	94.6
IMPR-Det (Ours)	ResNet50	99.9	91.1	94.1	98.3	85.7	87.2	99.4	96.7	90.1	86.2	92.9
IMPR-Det (Ours)	ResNet101	99.9	96.5	95.7	99.8	98.6	92.3	100.0	90.3	85.8	91.3	**95.0**

and generalization of object representations, thereby benefiting anchor-free detection in aerial images.

IV. CONCLUSION

This letter proposes an advanced anchor-free detector IMPR-Det for object detection in aerial images. IMPR-Det is characterized by its conceptually novel design principles: integrally mixing pyramid representations from multiple scales of the FPN for different subregions of an aerial instance, instead of the single-scale representation corresponding to the whole indivisible instance. This concept is instantiated by a novel detection head equipped with an adaptive routing mechanism. Extensive experimental results have validated the superiority of our method in terms of both accuracy and efficiency.

REFERENCES

[1] Y. Guo, X. Tong, X. Xu, S. Liu, Y. Feng, and H. Xie, "An anchor-free network with density map and attention mechanism for multiscale object detection in aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[2] C. Zhang, K.-M. Lam, T. Liu, Y.-L. Chan, and Q. Wang, "Structured adversarial self-supervised learning for robust object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5613720.

[3] G. Cheng, Y. Si, H. Hong, X. Yao, and L. Guo, "Cross-scale feature fusion for object detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 431–435, Mar. 2021.

[4] Z. Teng, Y. Duan, Y. Liu, B. Zhang, and J. Fan, "Global to local: Clip-LSTM-based object detection from remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603113.

[5] J. Chen, L. Wan, J. Zhu, G. Xu, and M. Deng, "Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 4, pp. 681–685, Apr. 2020.

[6] W. Han et al., "Improving training instance quality in aerial image object detection with a sampling-balance-based multistage network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10575–10589, Dec. 2021.

[7] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[8] T. Xu, X. Sun, W. Diao, L. Zhao, K. Fu, and H. Wang, "ASSD: Feature aligned single-shot detection for multiscale objects in aerial imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5607117.

[9] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, and T. He, "SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2384–2399, Feb. 2023.

[10] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019.

[11] Q. Lin, J. Zhao, B. Du, G. Fu, and Z. Yuan, "MEDNet: Multiexpert detection network with unsupervised clustering of training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4703114.

[12] 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.

[13] 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. 2980–2988.

[14] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.

[15] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9759–9768.

[16] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[17] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[18] H. Wei, Y. Zhang, Z. Chang, H. Li, H. Wang, and X. Sun, "Oriented objects as pairs of middle lines," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 268–279, Nov. 2020.

[19] L. Shi, L. Kuang, X. Xu, B. Pan, and Z. Shi, "CANet: Centerness-aware network for object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603613.

[20] J. Liu, S. Li, C. Zhou, X. Cao, Y. Gao, and B. Wang, "SRAF-net: A scene-relevant anchor-free object detection network in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5405914.

[21] T. Zhang, Y. Zhuang, G. Wang, S. Dong, H. Chen, and L. Li, "Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5608720.

[22] Y. Zhang, Z. Yan, X. Sun, W. Diao, K. Fu, and L. Wang, "Learning efficient and accurate detectors with dynamic knowledge distillation in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5613819.

[23] C. Zhang, K.-M. Lam, and Q. Wang, "CoF-net: A progressive coarse-to-fine framework for object detection in remote-sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5600617.

[24] C. Zhang, J. Su, Y. Ju, K.-M. Lam, and Q. Wang, "Efficient inductive vision transformer for oriented object detection in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5616320.

[25] L. Song et al., "Fine-grained dynamic head for object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 11131–11141.

[26] J. Fu, X. Sun, Z. Wang, and K. Fu, "An anchor-free method based on feature balancing and refinement network for multiscale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1331–1344, Feb. 2021.

[27] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[28] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.