# Large-Scale High-Altitude UAV-Based Vehicle Detection via Pyramid Dual Pooling Attention Path Aggregation Network

Zilu Ying, Jianhong Zhou, Yikui Zhai, *Senior Member, IEEE*, Hao Quan, Wenba Li, *Student Member, IEEE*, Angelo Genovese, *Senior Member, IEEE*, Vincenzo Piuri, *Fellow, IEEE*, and Fabio Scotti, *Senior Member, IEEE*

***Abstract*— UAVs can collect vehicle data in high-altitude scenes, playing a significant role in intelligent urban management due to their wide of view. Nevertheless, the current datasets for UAV-based vehicle detection are acquired at altitude below 150 meters. This contrasts with the data perspective obtained from high-altitude scenes, potentially leading to incongruities in data distribution. Consequently, it is challenging to apply these datasets effectively in high-altitude scenes, and there is an ongoing obstacle. To resolve this challenge, we developed a comprehensive vehicle dataset named LH-UAV-Vehicle, specifically collected at flight altitudes ranging from 250 to 400 meters. Collecting data at higher flight altitudes offers a broader perspective, but it concurrently introduces complexity and diversity in the background, which consequently impacts vehicle localization and recognition accuracy. In response, we proposed the pyramid dual pooling attention path aggregation network (PDPA-PAN), an innovative framework that improves detection performance in high-altitude scenes by combining spatial and semantic information. Object attention integration in both spatial and channel dimensions is aimed by the pyramid dual pooling attention module (PDPAM), which is achieved through the parallel integration of two distinct attention mechanisms. Furthermore, we have individually developed the pyramid pooling attention module (PPAM) and the dual pooling attention module (DPAM). The PPAM emphasizes channel attention, while the DPAM prioritizes spatial attention. This design aims to enhance vehicle information and suppress background interference more effectively. Extensive experiments conducted on the LH-UAV-Vehicle conclusively demonstrate the efficacy of the proposed vehicle detection method. Our code and dataset can be found at https://github.com/yikuizhai/PDPA-PAN.**

***Index Terms*— UAV, high-altitude scenes, vehicle detection, attention mechanism.**

## I. Introduction

IN RECENT years, the rapid advancements in unmanned aerial vehicle (UAV) technology, as well as UAV communication technologies [1], [2] have generated considerable interest in camera-equipped UAVs. The capability to conduct real-time data analysis is possessed by UAVs equipped with embedded devices, making them applicable in a wide range of practical scenarios, which include intelligent city management [3], security monitoring [4], [5], panoramic perception analysis [6], [7]. In intelligent city management, orbital satellites can be substituted by UAVs for performing real-time surveys of urban development progress. The high-resolution imagery captured by UAV can be effectively employed to oversee illegal construction activities within the city and optimize smart city traffic management [8], among other diverse applications.

In recent research, a comprehensive understanding of urban structures and features has been attained through the panoramic segmentation of UAV images [6]. Another study [7] is dedicated to addressing the detection of UAV images at varying flight altitudes. These endeavors showcase the substantial potential of UAVs in practical applications, with a specific emphasis on intelligent city management. During the process of urban development survey, significantly higher spatial resolution images can be provided by UAV imagery compared to those captured by orbital satellites. Specifically, the ability to easily distinguish objects that remain unidentified in orbital satellite imagery is provided by UAV imagery, which represents a key advantage over orbital satellites. However, this advantage poses a challenge that requires the use of techniques such as image similarity comparison [9] or change detection [10] techniques to effectively monitor urban development changes across different time periods using UAV imagery. Nevertheless, the amplification of specific mobile non-building objects, specifically vehicles, within

Manuscript received 27 August 2023; revised 8 January 2024; accepted 1 May 2024. This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515011576; in part by Guangdong Science and Technology Planning Project under Grant 2021A0505030080 and Grant 2021A0505060011; in part by Guangdong Higher Education Innovation and Strengthening School Project under Grant 2020ZDZX3031, Grant 2022ZDZX1032, and Grant 2023ZDZX1029; in part by the Wuyi University Hong Kong and Macao Joint Research and Development Fund under Grant 2022WGALH19; and in part by Guangdong Jiangmen Science and Technology Research Project under Grant 2220002000246 and Grant 2023760300070008390. The Associate Editor for this article was M. H. Anisi. *(Zilu Ying, Jianhong Zhou, Yikui Zhai, and Hao Quan contributed equally to this work.) (Corresponding author: Yikui Zhai.)*

Zilu Ying, Jianhong Zhou, Yikui Zhai, and Wenba Li are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, Guangdong 529020, China (e-mail: ziluy@163.com; jackychou_lab@126.com; yikuizhai@163.com; wenbalee@163.com).

Hao Quan is with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy (e-mail: hao.quan@polimi.it).

Angelo Genovese, Vincenzo Piuri, and Fabio Scotti are with the Department of Computer Science and the Dipartimento di Informatica, Universita Degli Studi di Milano, 20133 Milan, Italy (e-mail: angelo.genovese@unimi.it; vincenzo.piuri@unimi.it; fabio.scotti@unimi.it).

Digital Object Identifier 10.1109/TITS.2024.3396915

the image can be resulted from the elevated resolution of UAV imagery, which can potentially impact the performance of the mentioned algorithms. To address this challenge, various commonly employed strategies are included such as object preprocessing, region filtering, anomaly detection, and manual selection. However, accurate vehicle detectors are relied upon by all of these strategies, excluding time-consuming manual approaches. Moreover, precise and intricate information about foreground objects is offered by the high-resolution images captured by UAVs, while simultaneously introducing heightened complexity to the background.

Extensive research has been conducted on vehicle detection tasks [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], resulting in various aerial datasets suitable for vehicle detection, such as UAVDT [23], UAV123 [24], Stanford Drone [25], VisDrone [26], and DLR 3K [27]. Additionally, some studies are dedicated to exploring vehicle detection in challenging low-light conditions, such as the case of DroneVehicle [28]. Unfortunately, the mentioned UAV-based datasets were collected at flight altitudes below 150 meters. However, for urban development survey tasks, altitudes exceeding 350 meters must be operated by UAVs to effectively monitor the progress of urbanization and capture a sufficiently expansive field of view. Moreover, the aforementioned datasets were collected in typical scenarios, covering parking lots, bus stations, city streets, residential areas, and highways. In contrast, for urban development survey tasks, data collection is often performed by UAVs in specific scenes, such as urban-rural intersections, construction sites, and riverbank embankments. In these specific scenes, the background is more complex, and a common problem is the imbalance between foreground and background. Furthermore, certain changes may be introduced to the original images used for urban development survey projects during the preprocessing, thereby impacting the transferability of other vehicle detection datasets to this task. To address this gap, a UAV-based vehicle detection dataset was assembled at high flight altitudes, comprising 18,663 images and featuring annotations for 180,358 rotated bounding boxes. Moreover, a broader range of scenes, such as towns, construction sites, riverbank embankments, and various common vehicle detection scenarios, is encompassed by our LH-UAV-Vehicle dataset. All images were captured at flight altitudes ranging from 250 meters to 400 meters, and they are all aerial images. In this work, vehicle detection in this specific UAV-based high-altitude aerial scene is concentrated on, with the aim of enhancing the localization accuracy of vehicle detection to mitigate the influence of the vehicle's random movement attributes on UAV-based urban management.

Greater challenges are encountered in vehicle detection in high-altitude scenes as demonstrated by our LH-UAV-Vehicle images. Fig. 1 (a) reveals the imbalance between background and foreground information in more complex scenes, further exacerbating the difficulty of the vehicle detection task. Moreover, some UAV images captured at a flight altitude of 400 meters are illustrated in Fig. 1 (b), showing smaller and denser vehicles. In such cases, compared to UAV images taken at flight altitudes below 150 meters as shown in Fig. 1 (c), the information density of individual targets in the image is higher, making it more difficult to discern fine-grained details and accurately locate vehicle targets. Therefore, small object detection in the high-altitude scenes of UAV poses a significant challenge, often encountering issues such as low resolution and ambiguous features. The pyramidal feature hierarchy, introduced by [29], leverages feature maps with varying spatial resolutions to concurrently attend to information at multiple scales. This facilitates enhanced capture of targets across different scales within the image, with a particular emphasis on small targets. Furthermore, feature pyramid network (FPN) [30] emerges as one of the most recommended techniques to enhance vehicle detection accuracy. A set of feature maps with diverse spatial resolutions is generated by the method through a gradual fusion of high-level and shallow-level features, thereby enabling effective vehicle detection at various scales, particularly in UAV scenarios. Path aggregation network (PAN) [31] is a significant improvement over FPN in instance segmentation tasks. A combination of bottom-up and top-down structures is utilized to integrate spatial and semantic information from various feature layers, leading to enhanced localization accuracy. Additionally, [32], [33], and other researchers have further extension of PAN, leading to improved performance of the object detection model. In this investigation, we expand FPN to PAN by employing appropriate upsampling and downsampling methods. The model's perception of objects across different scales can be augmented, consequently enhancing the overall performance of the vehicle detection model.

Attention mechanisms have been extensively studied in previous research [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44]. The purpose of highlighting crucial feature positions and further enhancing their significance is served by attention while suppressing redundant ones. While [36], [38], [40], [42], [43], [44], and [53] have the potential to improve the performance of the model, optimal results can only be achieved by retraining the adapted backbone network for each of these techniques. This is because the inherent pre-trained weights of the backbone network are often disrupted by methods that involve embedding or substituting crucial components to facilitate attention. To address this challenge, method called pyramid dual pooling attention path aggregation network (PDPA-PAN) is introduced. The impact of complex backgrounds on the model is effectively mitigated by this innovative approach without disrupting the original features, thereby improving the model's ability to accurately detect and prioritize small vehicle targets. Refined features are extracted through attention by comprehensively integrating spatial relationships and abstract features from features of various sizes in this method. Specifically, two distinct attention modules are designed in this paper: pyramid pooling attention module (PPAM) and dual pooling attention module (DPAM), which are combined using pyramid dual pooling attention module (PDPAM). Furthermore, FPN is extended by incorporating a PAN and integrating it with PDPAM to emphasize the crucial features of


Fig. 1. Challenges are encountered in high-altitude scenes by UAV-based vehicle detection. (a) Diverse and complex backgrounds. (b) Densely scenarios in high-altitude scenes. (c) UAV images taken at flight altitudes below 150 meters. The images in both (a) and (b) are sourced from the LH-UAV-Vehicle dataset.

vehicle objects. The recognition and localization performance of vehicle detection are significantly improved by this integration.

The key contributions of this paper can be summarized as follows:

1) The LH-UAV-Vehicle dataset for vehicle detection was created by us, which includes 18,663 images containing a total of 180,358 instances of vehicle objects captured in various scenes. In contrast to other UAV datasets, a significant number of high-altitude scene images are encompassed by our dataset, thereby encompassing more intricate scenarios.
2) A pyramid dual pooling attention path aggregation network (PDPA-PAN) is introduced, allowing precise vehicle localization through the seamless integration of spatial and semantic information from diverse feature layers. Furthermore, the vehicle detection performance of our model is enhanced by employing a parallel strategy to integrate a hybrid attention mechanism.
3) To address the information loss resulting from global pooling, a novel pyramid pooling attention module (PPAM) is presented. Through a sequential process of fusing and computing attention, the model's parameter count is remarkably reduced, thereby enhancing detection performance. Moreover, a dual pooling attention module (DPAM) specifically designed for aerial vehicle detection is introduced by us. The extraction of finer features with an expanded receptive field is enabled by this module, which incorporates supplementary feature maps and considers convolutional kernels of various sizes.
4) Extensive experiments on the proposed LH-UAV-Vehicle dataset were conducted, clearly demonstrating the effectiveness of our approach. The baseline model was outperformed by our method with a 1.35% improvement in mAP50, and superior performance against state-of-the-arts was achieved.

The structure of the remaining paper is outlined as follows: A brief summary of related work is provided in Sec. II. The LH-UAV-Vehicle dataset is introduced in Sec. III. Sec. IV presents our proposed method (PDPA-PAN), and its effectiveness is evaluated through extensive experiments in Sec. V. Finally, conclusions are drawn in Sec. VI.

## II. Related Work

In this section, the currently available datasets capable of supporting vehicle detection tasks are first examined. Next, a concise overview of state-of-the-art vehicle detection algorithms is provided. Finally, various attention mechanism methods applicable for vehicle detection are reviewed.

### A. Aerial Scene Vehicle Detection Datasets

In recent years, several aerial scene datasets have been proposed by researchers to advance the progress of aerial view vehicle detection. Some datasets are collected through orbital satellites, such as the DLR 3K [27] dataset, which consists of 20 aerial images captured at a flight altitude of 1000 meters using the Canon Eos 1Ds Mark III camera, with a resolution of 5816×3744 pixels. Urban and residential areas are encompassed by the captured scenes, primarily comprising two categories: cars with 9300 instances and trucks with 160 instances. Other categories are often disregarded due to their limited number of instances. The VEDAI [45] dataset comprises 1200 small-sized images obtained through the cropping of remote sensing images collected by orbital satellites. A sparse distribution of vehicle objects is exhibited by the dataset, which consists of nine distinct categories of small vehicles, while encompassing a diverse range of complex scenes. Additionally, two versions of the dataset are comprised: 512×512 and 1024×1024, where VEDAI-512 is a downscaled version of VEDAI-1024. The DOTA [46] dataset is comprised of 2806 aerial images captured from platforms in various cities, resulting in image sizes ranging from 800×800 pixels to 4000×4000 pixels. 15 object categories are encompassed by it, consisting of vehicles of different scales and orientations. However, due to the high shooting altitude, fine-grained information about vehicles cannot be identified. As a result, two vehicle categories, small vehicles and large vehicles, are the only ones comprised in the dataset.

There are also efforts to gather datasets through UAVs for capturing aerial images, such as the Stanford Drone [25] dataset, which was captured by a 3DR Solo UAV equipped with a 4K camera at a height of 80 meters, focusing on eight distinct scenes within the Stanford University campus. It comprises approximately 60 annotated videos with a video resolution of 1400×1904 pixels and includes car and bus categories. The CARPK [4] dataset was collected by a UAV in four different parking lots at a flight altitude of approximately 40 meters. UAV123 [24] is a low-altitude UAV dataset with flight heights ranging from 5 meters to 25 meters. It contains 123 video sequences and 110,000 fully annotated images, including vehicle categories. The image resolutions vary from 1280×720 pixels to 3840×2160 pixels, primarily used for single-object tracking tasks. The UAVDT [23] dataset was collected by UAVs at flight altitudes ranging from 10 meters to 70 meters in complex scenes. 100 videos and 8,000 images with a resolution of 1080×5400 pixels are comprised in it. In addition to axis-aligned bounding box annotations, useful attribute information such as flight height, occlusion status, and weather conditions is also included. VisDrone [26] was collected by UAVs in different cities and under various weather conditions using different UAV platforms. 263 videos with a total of 179,000 frames and 10,209 images, with resolutions ranging from 2000×1500 pixels to 3840×2160 pixels. Various common car types, pedestrians, bicycles, and tricycles are encompassed within the dataset, which contains 10 object categories. The DroneVehicle [28] dataset is primarily used for UAV-based vehicle detection under low-light conditions. A total of 28,439 pairs of RGB-infrared cross-modal images are contained within, with flight altitudes ranging from 50 meters to 150 meters. The dataset incorporates five common vehicle categories and encompasses scenes featuring low-light conditions such as heavy fog and nighttime.

Compared to the aforementioned aerial scene datasets, the LH-UAV-Vehicle dataset we propose has been collected at altitudes ranging from 250 meters to 400 meters. Complex scenes with foreground-background imbalance and dense small objects are involved in it, making it suitable for UAV-based vehicle detection research.

### B. Vehicle Detection

The aim of vehicle detection is to precisely locate each vehicle within a given image and classify its category. In recent years, with the rapid advancement of object detection, aerial scene vehicle detection has been significantly benefited and has emerged as a crucial subfield in target detection for remote sensing images. Notably, extensive applications have been found by vehicle detection algorithms such as RetinaNet [47] and Faster R-CNN [48], regardless of the scene type, be it natural or aerial. A one-stage anchor-based object detection network called RetinaNet is used. Each vehicle object is predicted by independently transferring the output feature maps from the feature extractor to both the classification subnet and the regression subnet. The severe imbalance issue between foreground and background classes caused by dense predictions directly on feature maps is mitigated by the notable contribution of introducing focal loss. On the contrary, Faster R-CNN is a two-stage detector. In the first stage, proposals are generated using the region proposal network (RPN). These proposals are then classified into foreground and background regions and transferred to the second stage for bounding box refinements. In the second stage, regions of interest corresponding to the proposals are extracted from the feature map using region of interest (RoI) pooling. The final vehicle detection results are obtained by utilizing both the classification subnet and the regression subnet.

In the context of aerial scenes, many objects exhibit substantial variations in orientation and scale, making it challenging to accurately locate them using axis-aligned bounding boxes. To address this issue, the fine-grained detection of rotated objects is enhanced by $R^3$Det [49], which transforms axis-aligned bounding boxes into oriented bounding boxes, building on the foundation of RetinaNet. Additionally, the feature refinement module is devised to reconstruct the entire feature map pixel-by-pixel for precise feature alignment in order to tackle the misalignment between axis-aligned and oriented features. RoITransformer [50] is a three-stage detector that extends Faster R-CNN by incorporating the RoITransformer module, responsible for converting axis-aligned proposals generated by RPN into oriented proposals. A novel approach to address the misalignment issue between axis-aligned and oriented features is introduced by $S^2$ANet [51]. Oriented vehicle objects are directly predicted based on the feature map output from the feature encoder, followed by the transfer of the predicted position information

to the feature alignment module, which utilizes alignConv for precise feature alignment. Moreover, oriented detection module is proposed to separate classification features from regression features, ensuring sensitivity to rotation information while maintaining high classification accuracy. In contrast to RoITransformer, Oriented R-CNN [52] generates oriented proposals by employing a straightforward center point offset prediction for the proposal regions generated in the initial stage of Faster R-CNN. In the second stage, rotated RoIAlign is employed to extract corresponding regions of interest, which are then transferred to the classification subnet and the regression subnet to obtain the final results. Only 1/3000 of the parameter count of the RoITransformer module is required by the oriented RPN of Oriented R-CNN, while better detection results are achieved. Furthermore, intersection over union (IOU) calculations are significantly impacted by even slight variations in the oriented bounding box, especially for vehicles with changing orientations. Thus, in order to achieve precise oriented vehicle detection performance, a two-stage anchor-based model is adopted for implementing UAV-based vehicle detection in this paper.

### C. Attention Mechanisms

Relevant regions within an image can effectively be captured by attention mechanisms in computer vision, enabling object regions of interest to be specifically concentrated on by the model and attentional focus to be established on the feature map [36]. Within the domain of vehicle detection, a critical role is played by attention mechanisms in effectively emphasizing vehicle objects amidst numerous irrelevant background regions. Consequently, the precise localization and recognition of vehicles are facilitated. Channel attention, spatial attention, hybrid attention, and self-attention are encompassed by notable attention mechanisms. Channel attention, establishing correlations between channels in the feature map using Squeeze and excitation operations, is represented by SENet [36]. Subsequently, channel-wise weights are applied to enhance critical features while suppressing less significant ones. To tackle spatial information loss caused by global average pooling in SENet, SPANet [40] introduces spatial pyramid pooling to compensate for the missing spatial information and obtain improved channel-wise weights.

SKNet [42] investigates the correlations between convolutional kernels and proposes selective kernel convolution. Through parallel convolutions of varying sizes, SKNet obtains feature maps with distinct receptive fields and fuses multi-scale semantic information to learn channel correlations in different feature maps. The resulting channel-wise weights are then applied to the parallel feature maps, and the final feature map is obtained through an add operation. LSKNet [43] combines large convolutions and selective kernel convolution, leading to large selective kernel convolution. Unlike SKNet, LSKNet replaces parallel convolutions with sequential convolutions and utilizes larger receptive field dilated convolution kernels to capture larger receptive fields. LSKNet employs spatial attention instead of channel attention.

scSE [37], BAM [39], and CBAM [38] are prominent examples of hybrid attention mechanisms, effectively integrating both channel attention and spatial attention to optimize feature maps. SENet is expanded upon by scSE through the introduction of spatial attention and channel attention modules. The final output is yielded by scSE through an add operation, fusing feature maps extracted using both attention mechanisms. Element-wise multiplication is utilized by BAM to combine channel-wise weights and spatial-wise weights, resulting in the generation of mixed attention weights. The input feature map is then subjected to the application of the mixed attention weights to emphasize key features. In contrast to scSE and BAM, a sequential combination strategy is adopted by CBAM to integrate both attention mechanisms. Two sequential combinations and one parallel combination are explored in [38], with experimental results indicating that superior performance is achieved by channel attention followed by spatial attention.

Computer vision tasks are addressed with remarkable success by applying self-attention mechanisms to them by Vision transformer [53]. In order to reduce reliance on external information, inherent semantic information within the feature map is leveraged by self-attention, which is a variant of attention mechanisms. Spatial window multihead self-attention and channel group self-attention are introduced by DaVit [44], which combines self-attention with spatial attention and channel attention. The two self-attention mechanisms are then combined through serial strategy. These attention mechanisms are typically embedded as plug-ins within the blocks of the backbone network [36], [38], [40], or crucial components in the backbone network are replaced with components featuring attention mechanisms [42], [43], [44], [53].

However, the original pre-trained weights of the backbone network may be disrupted by embedding attention mechanisms, while the replacement approach necessitates that the model be retrained to obtain the corresponding initial weights. To tackle this issue, the pyramid dual pooling attention path aggregation network (PDPA-PAN) is presented as a novel approach that effectively integrates attention mechanisms and NeckNet into a cohesive framework. This integration is seamlessly accomplished without any disruption to the feature extraction process in the backbone network.

## III. LH-UAV-Vehicle Dataset

In this section, we presents a comprehensive overview of data split and preprocessing as well as information related to the LH-UAV-Vehicle dataset. Table I presents a comprehensive comparison between the provided dataset and other relevant benchmark datasets in the field of vehicle detection.

### A. Images Collection

The dataset LH-UAV-Vehicle consists of 18,663 optical images, all of which were captured using DJ Mavic2 UAVs equipped with cameras. To obtain a diverse range of remote sensing images, governmental permission was obtained to capture images at altitudes ranging from 250 to 400 meters in various towns and urban areas located in

TABLE I

COMPARISON OF THE STATE-OF-THE-ART BENCHMARKS AND DATASETS. NOTE THAT, THE OCCURRENCE OF '-' REPRESENTS INSTANCES THAT ARE UNCOUNTABLE ACCURATELY DUE TO THE INCLUSION OF VIDEOS IN THE DATASET

| Object Detection Datasets | Top View | Oriented Bounding Box | UAV Platform | Complex Scence | Altitude(m) | Images | Instances | Image Size |
|---|---|---|---|---|---|---|---|---|
| DLR 3K [27] | ✓ | ✓ | | | 1000 | 20 | 14,235 | 5616×3744 |
| VEDAI [45] | ✓ | ✓ | | ✓ | more than 2000 | 1200 | 2950 | 512×512 or 1024×1024 |
| DOTA [46] | ✓ | ✓ | | ✓ | more than 2000 | 2806 | 188,282 | 800×800 to 4000×4000 |
| Stanford Drone [25] | ✓ | | ✓ | | 80 | 60 videos | - | 1400×1904 |
| CARPK [4] | ✓ | | ✓ | | 40 | 1448 | 89,777 | 1280×720 |
| UAV123 [24] | | | ✓ | ✓ | 35 | 123 videos | - | 1280×720 to 3840×2160 |
| UAVDT [23] | | | ✓ | ✓ | 5 to 25 | 80000 | 840,000 | 1080×540 |
| VisDrone [26] | | | ✓ | ✓ | 10 to 70 | 10209 images and 263 videos | - | 840×712 |
| DroneVehicle [28] | | ✓ | ✓ | ✓ | 80 to 120 | 28,439 | 452,570 | 840×712 |
| LH-UAV-Vehicle **(ours)** | ✓ | ✓ | ✓ | ✓ | 250 to 400 | 18,663 | 180,358 | 640×640 to 1024×1024 |

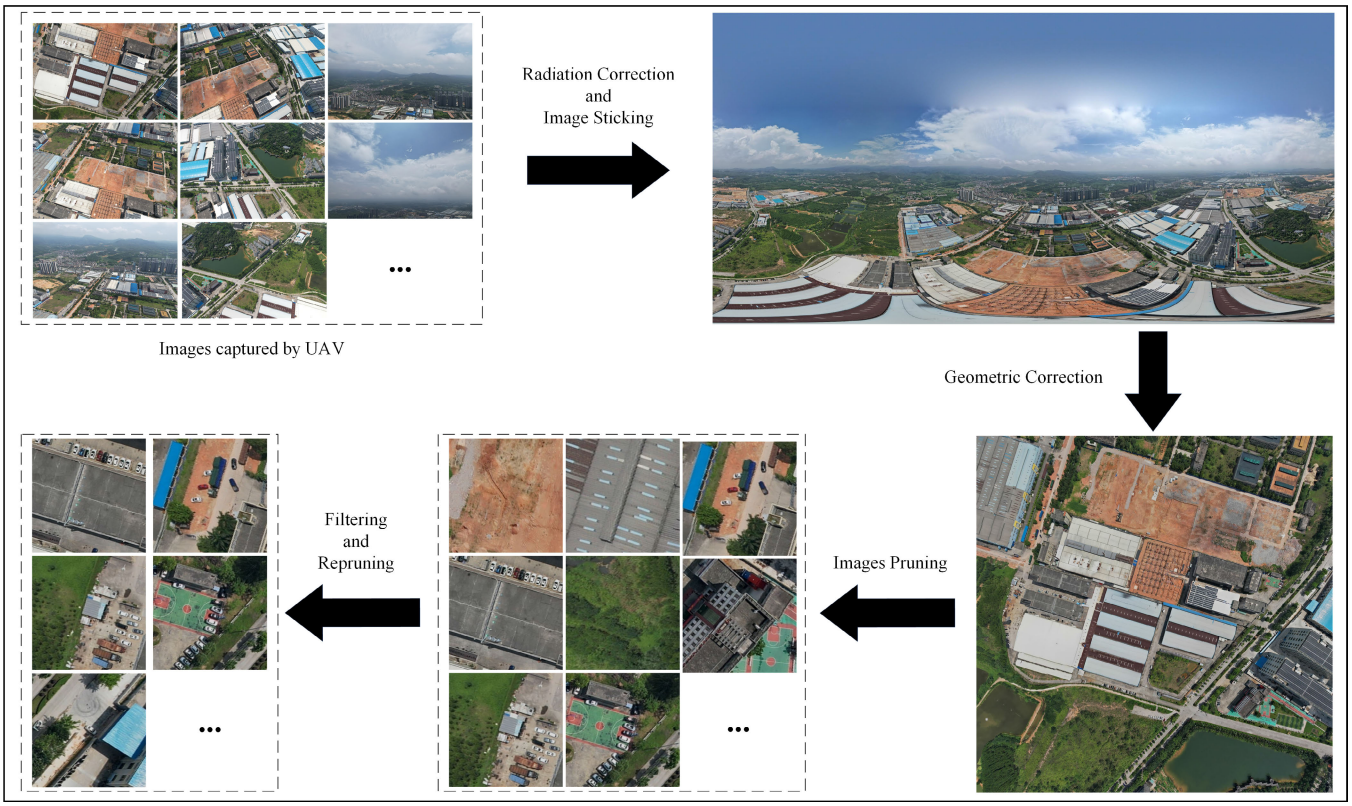


Fig. 2. A structured sequence of steps is involved in the preprocessing procedure for the LH-UAV-Vehicle dataset.

Shantou City and Jiangmen City, Guangdong Province, China. A diverse range of environments is included in the captured scenes, such as townships, construction sites, riverbank embankments, urban roads, residential areas, parking lots, and highways. The specific instance categories are comprised of cars, vans, buses, and trucks. It is important to note that this dataset was collected across various scenes and flight altitudes using DJ Mavic2 UAVs. The collection process entailed determining the UAV's flight altitude and global positioning system (GPS) location, capturing multiple remote sensing images from various directions at designated locations, and generating panoramic images using a synthesis algorithm. Using this method, we obtained over 400 panoramic images. Additionally, We have enriched LH-UAV-Vehicle from additional datasets. As a result, our dataset comprises 180,358 manually annotated bounding boxes.

### *B. Images Preprocessing*

Due to the images being captured at a flight altitude of 250-400 meters, the collection process is susceptible to various factors, including scene illumination, camera angle, and wind velocity. Consequently, it is essential to apply suitable data processing techniques to mitigate the influence of external factors. The specific procedure for data preprocessing is depicted in Fig. 2. Initially, potential

interference stemming from variations in illumination and atmospheric conditions, affecting images acquired by UAV, is effectively mitigated through precise radiation correction techniques. Subsequently, seamless panoramic compositions are generated via sophisticated image stitching algorithms. After creating panoramic representations, a rigorous geometric correction process is applied to rectify distortions and stretching artifacts caused by misalignment of the images. The derivation of meticulously pseudo-orthorectified images is resulted in, laying the foundation for subsequent analyses. For the purpose of spatial consistency and facilitating future assessments, the previously captured images are systematically divided into standardized patches that measure 1024×1024 pixels. A comprehensive evaluation of the segmented image patches is entailed in the ultimate step of the preprocessing. Images of insufficient quality and those lacking discernible objects are effectively identified and removed through this assessment. In addition, a specialized recropping process is applied to images that exhibit significant boundary deformation, resulting in the generation of 640×640 pixel image patches.

*1) Radiation Correction:* The implementation of radiometric calibration is deemed imperative in order to enhance image quality and minimize variations in pixel grayscale resulting from lighting and atmospheric conditions. Prominent techniques for radiometric calibration encompass empirical calibration, physical model construction, horizontal plane projection, spectral matching, and radiometric calibration based on pseudo-invariant features. Radiometric calibration discrepancies are caused by variations in shooting angles during the UAV image stitching process. However, consistent spectral characteristics are exhibited by pseudo-invariant features across varying times and conditions. Therefore, the calibration of image radiometric values is facilitated through the adoption of a pseudo-invariant feature-based approach, leading to enhanced uniformity in brightness and color, as well as seamless stitching. Furthermore, the practicality of employing radiometric calibration based on pseudo-invariant features is achieved by eliminating the requirement for additional radiometric data acquisition devices. Therefore, a radiometric calibration method based on pseudo-invariant features was employed. Initially, representative buildings and roads were selected for feature matching, from which radiometric calibration coefficients were computed. Subsequently, these coefficients were applied to conduct radiometric calibration on the UAV images. Prior to subsequent processing, radiometric calibration was performed on the pre-stitched images using this method.

*2) Geometric Correction:* Misalignment of the relative positions of the captured objects is often experienced due to a multitude of factors, including systematic and non-systematic influences like camera azimuth parameters, UAV flight altitude, and variations in terrain. Pixel distortion and stretching are caused by this misalignment relative to their true positions. Therefore, during the data acquisition process, the camera's attitude information is concurrently captured by the UAV along with the images. Subsequently, alignment of the captured images with the recorded UAV's attitude information is carried out, followed by the execution of the necessary operations to achieve geometric correction. As a result, pseudo-orthorectified images with dimensions of 8192×8192 pixels are generated.

*3) Images Pruning:* The direct utilization of remotely sensed images, which have undergone both radiation and geometric correction, for annotation or model training is deemed impractical. Firstly, images with a resolution as high as 8192×8192 pixels are found to be challenging for models to handle. Secondly, significant challenges are posed by the manual annotation of large images containing densely distributed small objects. Therefore, the additional processing of pseudo-orthorectified images is deemed necessary. A crucial role is played by data pruning in dataset construction. In the initial stage, the 8192×8192 pseudo-orthorectified images are cropped into 1024×1024 image patchs. Next, a manual inspection is conducted on all image patchs to eliminate images with poor quality (e.g., excessively blurry and indistinguishable) and those lacking objects. Additionally, images exhibiting severe boundary deformation are recropped into 640×640 image patchs. Ultimately, after additional selection, the original unannotated data is obtained.

### C. Semi-Automatic Annotation Method

In the field of object detection, the prevailing method for data annotation involves annotating image objects using rectangular bounding boxes. There are three common annotation formats for bounding boxes: $(x_c, y_c, w, h)$, $(x_{left}, y_{left}, w, h)$, and $(x_{left}, y_{left}, x_{right}, y_{right})$. Here, $(x_c, y_c)$ denotes the center position, $(x_{left}, y_{left})$ represents the top-left corner position, $(x_{right}, y_{right})$ denotes the bottom-right corner position, while $w$ and $h$ correspond to the width and height, respectively. These annotation methods are commonly used in typical object detection scenarios, including autonomous driving and security monitoring. However, the accurate calibration of objects using standard rectangular boxes is challenged by the presence of object orientation uncertainty in pseudo-orthorectified images. Therefore, during the annotation process, the object contours are precisely and compactly outlined using arbitrary quadrilaterals, and their minimum bounding rectangles are subsequently calculated to obtain the final annotation information. Reference [46] is utilized to store annotation information for rotated bounding boxes of vehicles. The horizontal and vertical coordinates of four points are recorded in a clockwise order in this process. Stringent annotation criteria are employed when addressing noisy data. Our approach involves excluding vehicles with truncation rates or occlusion rates surpassing 75% from annotation. The rationale behind this choice is to uphold the dataset's visual data integrity, achieved by omitting vehicles with significant truncation or occlusion. This enhances the dependability of resources for diverse applications.

Semi-automatic annotation refers to the process of pseudo-annotating images using a pre-trained model prior to manual annotation. Our dataset was collected over multiple intervals, and the utilization of the semi-automatic annotation method can greatly improve annotation efficiency while also mitigating the risk of missing annotations often encountered with manual

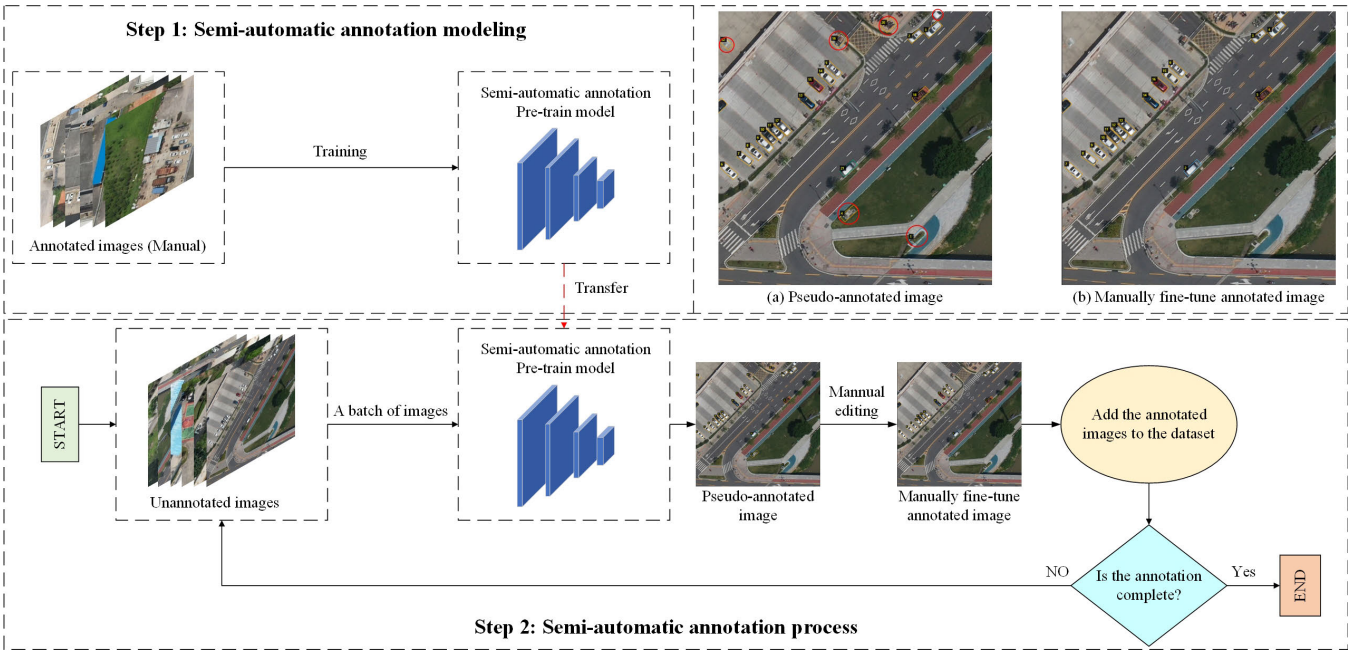


Fig. 3. The processing procedure of the semi-automatic annotation method. Step 1: Semi-automatic annotation modeling. Step 2: Semi-automatic annotation process. (a) Pseudo-annotated image. (b) Manually fine-tune annotated image.

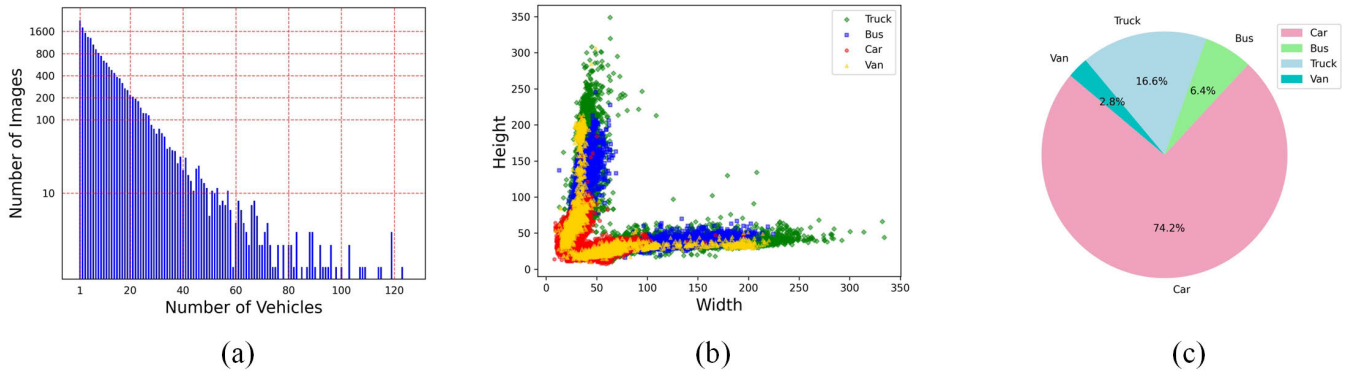


Fig. 4. The visualization of the LH-UAV-Vehicle attributes. (a) The distribution of all vehicle instances. (b) The scatter plot of the width and height of all vehicle instances. (c) The proportion of pixels occupied by different vehicle instances of an image in the foreground.

annotation alone. The specific process is depicted in Fig. 3. Before the initiation of the semi-automatic annotation, the model was initially fine-tuned using a limited amount of annotated data to adapt it for the task of semi-automatic annotation, as illustrated in step 1 of Fig. 3. Subsequently, the semi-automatic annotation model was utilized to create pseudo-annotations on the images. Naturally, not all pre-generated annotations are precise; certain inaccuracies are present, encompassing erroneous and omitted annotations, as depicted in step 2 of Fig. 3. These discrepancies, visually marked with red solid circles in Fig. 3 (a), necessitate careful adjustment and rectification. Following this, discrepancies present in the pseudo-annotated images are meticulously rectified and refined through manual intervention. Additionally, any overlooked annotations are meticulously introduced to produce the annotated images, as depicted in Fig. 3 (b). Finally, the annotated images were integrated into the fine-tuning data to optimize the semi-automatic annotation model. This iterative process continued until all images were annotated. It is essential to note that the annotations are performed in batches rather than optimizing the semi-automatic annotation model after annotating each individual image. The recently acquired data is partitioned into multiple batches. Following batch-wise semi-automatic annotation and manual refinement, the data is employed for the optimization of the semi-automatic annotation model. Moreover, *labelImg* is utilized for both manual annotation and fine-tune annotation tasks.

### *D. Statistics and Attributes*

Four vehicle categories that are commonly of interest in UAV applications, namely cars, buses, trucks, and vans, have been selected for the LH-UAV-Vehicle dataset. Calculations were conducted within the LH-UAV-Vehicle dataset to determine the total number of annotated bounding boxes for each category. The corresponding statistical results can be found in Table II, and the detailed visualization of the LH-UAV-Vehicle attributes are illustrated in Fig. 4. The dataset

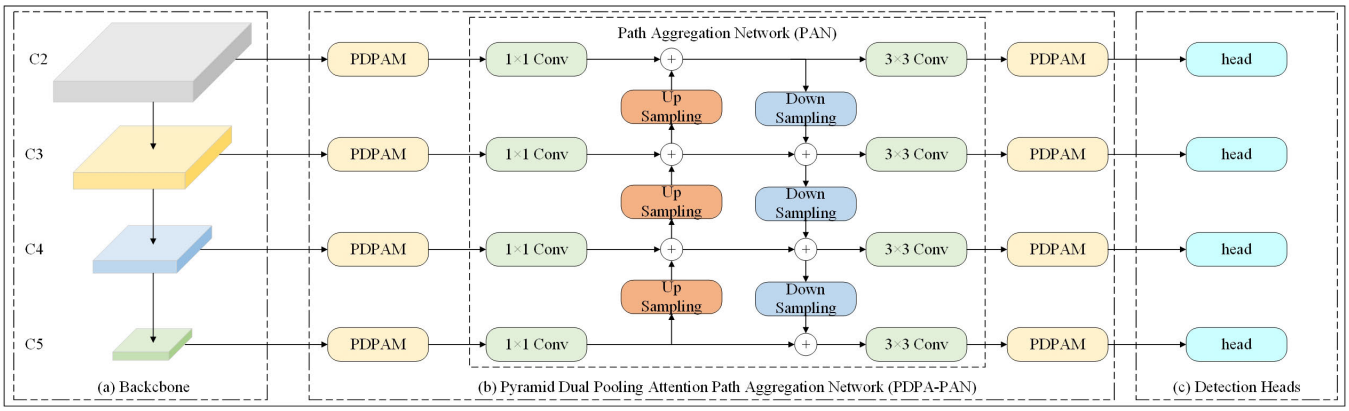


Fig. 5. The overall framework of our proposed vehicle detection method. The PDPA-PAN model comprises two components: PAN and PDPAM. To begin, the feature maps (C2-C5) produced by the backbone are processed by PDPAM to extract specific information from the corresponding feature layers. Subsequently, the hierarchical architecture of the PAN component, which incorporates both bottom-up and top-down pathways, is utilized to comprehensively fuse semantic and spatial information from distinct feature layers. Lastly, after undergoing extensive fusion, the feature maps are subjected to another round of processing by PDPAM in order to extract finer features. These refined feature maps are then directed to their respective detection heads.

TABLE II
STATISTICAL ANALYSIS OF THE OBJECT NUMBER FOLLOWING THE PARTITIONING OF THE LH-UAV-VEHICLE DATASET

| Set of Type | car | bus | truck | van |
|---|---|---|---|---|
| Training Set | 106,534 | 3069 | 9502 | 2465 |
| Validation Set | 26,956 | 826 | 2320 | 608 |
| Test Set | 23,059 | 287 | 3654 | 1078 |

consists of 18,663 optical images that were captured by UAVs. Comprehensive annotations using arbitrary quadrilaterals have been provided for these four categories. Specifically, there are 156,549 instances of cars, 15,476 instances of trucks, 4,182 instances of buses, and 4,151 instances of vans. The images were captured during daylight and encompass a wide range of scenes, including parking lots, residential areas, urban roads and streets, construction sites, and rural environments. Moreover, within the LH-UAV-Vehicle dataset, an average of 9.66 vehicles are contained in each image, with a maximum count of 126 vehicles. In summary, a diverse range of complex scenes and high flight altitudes are covered by the LH-UAV-Vehicle dataset. A significant number of finely annotated vehicle objects are included, addressing the gap in detecting vehicles in remote sensing images captured by UAVs at altitudes of 250 meters and above.

### *E. Data Split*

In order to achieve a more balanced distribution match between the training and testing data and to effectively assess the model's performance on the validation and testing sets, the data will be partitioned based on the time of collection. Our dataset was collected during two distinct time periods. The images from one period will be exclusively used as the training set, comprising approximately 65% of the original images. The images from the other period will be utilized to create the validation and testing sets, with the validation set accounting for 17% of the original images and the testing set accounting for 18% of the original images. Moreover, the training set, validation set, and testing set collectively cover a diverse range of complex scenes, comprising weather influences, low resolution, dense vehicular distribution, and scenarios where vehicles are predominantly occluded. This approach is undertaken to precisely evaluate the model's holistic performance across different intricate scenarios. All original images and annotation information for the training, validation and testing sets will be publicly released.

## IV. METHODS

The introduction of a novel vehicle detection method named pyramid dual pooling attention path aggregation network (PDPA-PAN) is presented in this paper. PDPA-PAN is comprised of two pyramid dual pooling attention module (PDPAM) and a path aggregation network (PAN) that is an extension of feature pyramid network (FPN). Each PDPAM consists of a pyramid pooling attention module (PPAM) and a dual pooling attention module (DPAM). Essential information from all layers is integrated by the network, enabling improved focus on vehicle objects. The overall framework of the proposed method is illustrated in Fig. 5.

### *A. Pyramid Pooling Attention Module*

In order to address the spatial information bias in channel attention maps that arises from the direct application of global pooling for channel information extraction, multiple adaptive pooling operations with varying output sizes are utilized. Channel attention maps encompassing diverse spatial information can be aggregated using this approach. Both max pooling and average pooling have been explored in recent studies on attention. The extraction of critical features associated with the foreground objects is enhanced by max pooling, facilitating a more detailed channel attention mechanism. In contrast, background interference is effectively mitigated and information redundancy is reduced by average pooling. As a result, the simultaneous utilization of max pooling and average pooling with varying dimensions is proposed. Nonetheless, escalated computational complexity

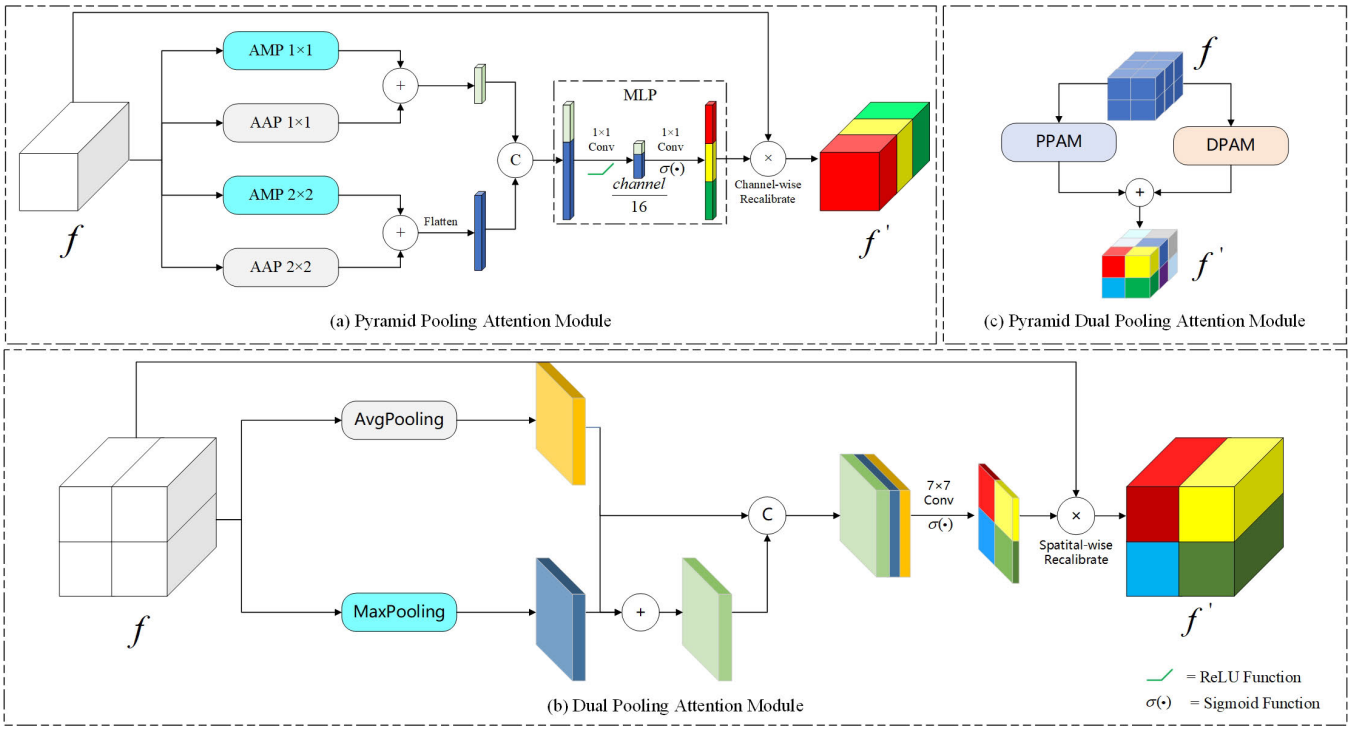


Fig. 6. The conceptual illustration of the proposed modules. (a) The proposed channel attention mechanism, PPAM, integrates dual pooling stages of varying sizes for four pooling layers. Initially, input feature maps are aggregated using four distinct pooling strategies to capture spatial information comprehensively. Subsequently, outputs with identical dimensions are fused through summation, flattened into feature vectors, and concatenated. The merged feature vectors undergo MLP processing to compute channel-wise weights. Finally, the derived channel-wise weights are applied to input feature maps, generating refined feature maps through channel attention. (b) The proposed spatial attention mechanism, DPAM, utilizes the distinctive strengths of both maximum pooling and average pooling. Initially, the input feature maps undergo channel-axis pooling, employing these methods to capture channel-related information. The results from both pooling approaches are additively fused, generating supplementary feature maps. Subsequently, the three resulting feature maps are concatenated and subjected to a convolutional operation for computing spatial-wise weights. Lastly, the spatial-wise weights derived are applied to the input feature maps. (c) The proposed structure that integrates two distinct attention mechanisms is denoted as PDPAM. PDPAM combines the PPAM and DPAM attention mechanisms using a parallel approach.

may be caused by larger channel attention maps. Therefore, the addition operation is employed to fuse features and reduce subsequent computations by merging pooling outputs of the same size but different types. Experimental results reveal that the integration of channel attention modules with diverse spatial information sizes substantially enhances the network's representational capacity. The effectiveness of this design choice is highlighted when compared to a channel attention module with a single size (refer to Table III).

In the beginning, the spatial information from feature map $f$ is aggregated by employing both average pooling and max pooling operations with different sizes. Four distinct spatial contextual representations are generated by this process: $f^c_{avg1}$, $f^c_{avg2}$, $f^c_{max1}$, and $f^c_{max2}$. These representations correspond to adaptive average pooling outputs with dimensions of $1\times1$ and $2\times2$, as well as adaptive maximum pooling outputs with dimensions of $1\times1$ and $2\times2$, respectively. Subsequently, we perform addition operations between $f^c_{avg1}$ and $f^c_{max1}$ to obtain the fused feature $f^c_1$, and between $f^c_{avg2}$ and $f^c_{max2}$ to obtain the fused feature $f^c_2$. Afterwards, we flatten $f^c_2$ and concatenate it with $f^c_1$ using a concatenation operation. The channel attention map $M_c$ is generated by feeding the concatenated result into a multi-layer perceptron (MLP). Two fully connected layers are comprised by the MLP. The number of hidden nodes is reduced to one-sixteenth of the original size, as suggested by [36], in order to mitigate parameter overhead. Lastly, the input feature map $f$ is applied with the channel attention map $M_c$. In summary, the detailed process is shown in Fig. 6 (a), the computation process of pyramid pooling attention module (PPAM) can be summarized as follows:

$$\begin{aligned} f'_c &= f \times MLP\left(concat\left(AMP1\left(f\right)\right.\right. \\ &\quad \left.\left. + AAP1\left(f\right), AMP2\left(f\right) + AAP2\left(f\right)\right)\right) \\ &= f \times MLP\left(concat\left(f^c_{avg1} + f^c_{max1}, f^c_{avg2} + f^c_{max2}\right)\right) \\ &= f \times MLP\left(concat\left(f^c_1 + f^c_2\right)\right) \\ &= f \times M_c \end{aligned} \tag{1}$$

where $AMP$ denotes adaptive maximum pooling, $AAP$ denotes adaptive average pooling, and the numbers 1 and 2 following each pooling method correspond to the output sizes of the adaptive pooling. $Concat$ signifies the concatenation operation.

### B. Dual Pooling Attention Module

Likewise, a spatial attention map is generated by considering the positional relationships within the feature map. When processing images, the relationships between different positions are essential for visual comprehension, particularly in scenarios where images feature intricate scenes or objects at specific locations. Solely relying on channel attention might not adequately capture the relationships between

positions. This limitation could introduce bias in the model's overall scene understanding due to inadequate attention across different positions. Hence, we introduce spatial attention as a complementary technique, modulating feature responses at each position based on their computed significance. Presently, [34] and [35] concurrently employ average pooling and max pooling for spatial attention modeling. This is attributed to the fact that average pooling mitigates background interference, whereas max pooling accentuates crucial features of foreground objefcts. This combination facilitates a more comprehensive evaluation of information importance across various positions, resulting in a more accurate generation of spatial attention maps. For further enhancement of spatial attention quality, we introduce an additional feature map while concurrently considering max pooling and average pooling. The goal is to capture comprehensive information influenced by both average and max pooling operations, consequently improving attention quality at specific positions. Following this step, the two pooled feature maps and the additional feature map are concatenated to extract spatial attention. Subsequently, a convolutional layer is applied to generate the spatial attention map, effectively modulating spatial information by enhancing or suppressing. The specific operations will be further explained in the subsequent sections.

Firstly, the channel information from the input feature map is aggregated utilizing two distinct pooling operations, resulting in two feature maps: $f_a^s$ and $f_m^s$. Subsequently, an additional feature, $f_e^s$, is obtained by summing the two feature maps. The three feature maps are concatenated to form $f_{mae}^s$. Afterwards, the spatial attention map, $M_s$, is computed by applying a convolutional layer to the $f_{mae}^s$ feature map. In summary, the detailed process is depicted in Fig. 6 (b). The computation process of dual pooling attention module (DPAM) can be summarized as follows:

$$\begin{aligned} f_s' &= f \times Conv_{7\times7}\left(concat\left(MP(f), AP(f), AP(f)\right.\right. \\ &\quad \left.\left.+MP(f)\right)\right) \\ &= f \times Conv_{7\times7}\left(concat\left(f_m^s, f_a^s, f_e^s\right)\right) \\ &= f \times Conv_{7\times7}\left(f_{mae}^s\right) \\ &= f \times M_s \end{aligned} \tag{2}$$

where $MP$ represents maximum pooling, $AP$ represents average pooling, $Conv7\times7$ denotes a convolution operation using a 7×7 filter size, and *concat* indicates the concatenation operation.

### C. Pyramid Dual Pooling Attention Path Aggregation Network

In the context of vehicle detection, the extraction of semantic information from deep features through shallow features is allowed by the bottom-up structure of feature pyramid network (FPN). Multi-scale reasoning is facilitated by this, which in turn results in the enhancement of the performance of the detector. Furthermore, extensive interactions between deep semantic information and shallow spatial information across different feature layers are facilitated by the combined bottom-up and top-down structures of path aggregation network (PAN). Enhanced performance in vehicle detection tasks is led by this. The PAN paradigm has been extended by previous works [33], [36]. Therefore, PAN and pyramid dual pooling attention module (PDPAM) are combined in our approach, and deep semantic and shallow spatial information are integrated through a mixed attention mechanism, resulting in the extraction of refined features for accurate vehicle detection. The comprehensive framework diagram of pyramid dual pooling attention path aggregation networ (PDPA-PAN) is illustrated in Fig. 6 (c) The detailed description of PDPA-PAN will be provided below.

The nearest-neighbor interpolation upsampling, utilized in FPN, was substituted with bilinear interpolation upsampling, and a top-down pathway was simultaneously incorporated, thereby extending FPN to the PAN architecture. Within the downsampling process, resolution was diminished by implementing a 3×3 pure convolution operation. In contrast to contemporary methodologies, the adoption of concatenation operation, frequently utilized in their feature fusion, was refrained from; instead, the original addition operation employed in both PAN and FPN for feature fusion was preserved. Two types of attention modules proposed by us are integrated in PDPAM, which can be deployed in either parallel or sequential fashion. It was demonstrated by our experiments that superior results were yielded when parallel placement of PDPAM was employed in necknet, as depicted in Fig. 5. The specific workflow of PDPA-PAN can be described as follows: firstly, the feature maps generated by the backbone at each layer undergo attention extraction using PDPAM, followed by dimension adjustment through 1×1 convolution. Subsequently, the proposed PAN is utilized to comprehensively fuse spatial and semantic information, which is followed by feature refinement through 3×3 convolution. Finally, attention is applied using PDPAM to obtain refined feature maps, which are then separately fed into the shared detection head.

## V. Experimental and Results

The proposed network for vehicle detection will be examined in this section. First, ablation studies were performed on the LA-UAV-Vehicle dataset to assess the effectiveness of the proposed method. Next, an in-depth analysis of the operations that impact the method's performance was conducted. Lastly, the method's performance was compared with state-of-the-art techniques, utilizing various detection metrics, and the visual detection results were discussed in detail.

### A. Experimental Setting

*1) Implementation Details:* ResNet50-PAN, initialized with a pre-trained ResNet50 [59] model on the ImageNet [57] dataset, was chosen as the backbone network for this study. To enhance diversity, random rotations with a probability of 0.5 at 90/180/270 degrees, and random horizontal/vertical flips with a probability of 0.5 were applied to each input image. The complete network was trained using the SGD+Momentum optimizer for a single schedule (12 epochs). Learning rate decay is incorporated during the 7th and 10th epochs,

employing a decay factor of 0.1. An initial learning rate was utilized by the baseline of the ablation study, with a batch-size of 2. The chosen values for weight decay and momentum were 0.0001 and 0.9, respectively. All experiments were performed utilizing the *Jittor* framework on a workstation that featured an NVIDIA RTX3090 GPU.

*2) Baselines:* Our proposed method was compared with 8 state-of-the-art approaches: $R^3$Det [49], RsDet [54], ATSS [55], $S^2$ANet [51], Faster R-CNN(OBB) [48], RoITransformer [50], Gliding [56], and Oriented RCNN [52]. To ensure a fair comparison, detection results for both our method and the baselines were acquired using identical experimental settings, with the exception of the initial learning rate. This is because performance bias in different detectors may be caused by variations in the initial learning rate.

*3) Experimental Data:* The experiments in this paper were conducted using the LH-UAV-Vehicle dataset. Training was conducted using the training set, and evaluation was performed on the test set. Moreover, the LH-UAV-Vehicle dataset consists of images in two sizes: 640×640 and 1024×1024. Images with a size of 1024×1024 undergo cropping to 640×640 with an overlap ratio of 128. Importantly, all training and validation sets within the LH-UAV-Vehicle dataset will be accessible as open-source data.

*4) Evaluation Metric:* Multiple evaluation metrics were adopted, comprising mean average precision 50 (mAP50), mean average precision 75 (mAP75), and mean average precision 50:95 (mAP50:95). Specifically, a prediction is considered correct by mAP50 when the intersection over union (IOU) between the prediction and the ground-truth is higher than 0.5. A prediction is considered correct by mAP75 when the IOU is greater than 0.75. The mean value of all metrics ranging from mAP50 to mAP95, with a sampling interval of 0.05, is calculated to obtain mAP50:95. Among these metrics, the model's recognition capability is gauged by mAP50, the model's localization capability is assessed by mAP75, and the model's overall performance is evaluated by mAP50:95.

## B. Ablation Study

In this section, the effectiveness of our proposed pyramid dual pooling attention path aggregation network (PDPA-PAN) is validated by using Oriented RCNN as the baseline and conducting relevant experiments on the LH-UAV-Vehicle dataset. Additionally, three distinct backbones were taken into consideration: ResNet18 [59], ResNet50 [59], ResNet101 [59], and SwinTransformer-Tiny [60]. For brevity in the table, ResNet18 is abbreviated as R-18, ResNet50 as R-50, ResNet101 as R-101, and SwinTransformer-Tiny as Swin-T. Notably, the average of multiple experiments is represented by all our experimental results.

*1) Pyramid Pooling Attention Module:* The effectiveness of the designed pyramid pooling attention module (PPAM) for vehicle detection was validated by conducting experiments utilizing various channel attention mechanisms illustrated in Fig. 7, and their impact on the detection performance was assessed. In this experiment, channel attention was exclusively incorporated before the 1×1 convolutional layer of the PAN,

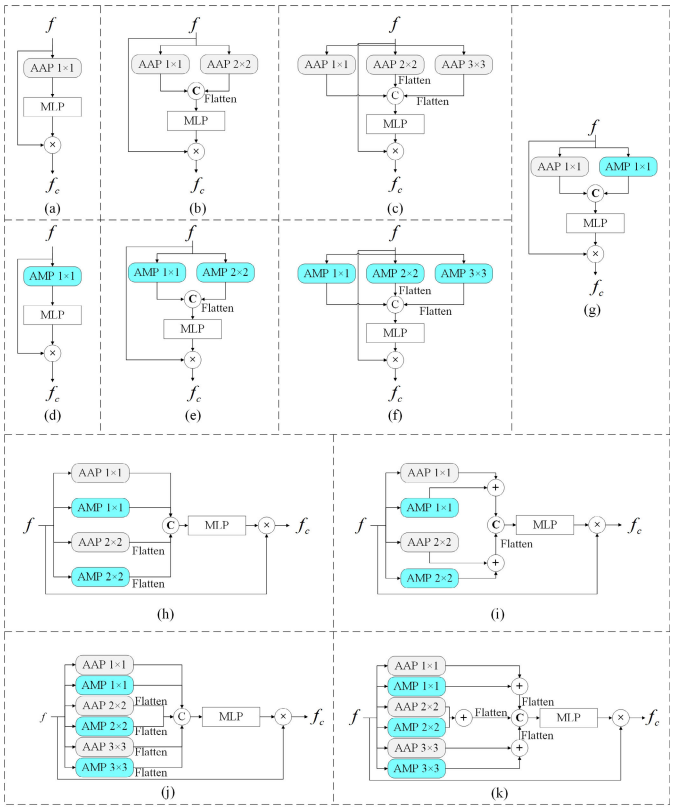


Fig. 7. The various channel attention modules are presented in the following sequence: (a) $C_{A1}$ module, (b) $C_{A2}$ module, (c) $C_{A3}$ module, (d) $C_{M1}$ module, (e) $C_{M2}$ module, (f) $C_{M3}$ module, (g) $C_{Mix1}$ module, (h) $C_{Mix2}$ module, (i) $C_{Mix3}$ module, (j) Proposed $PPAM_2$ module, and (k) Proposed $PPAM_3$ module.

and the results of the experiments are presented in Table III. It is suggested by the results that the model's vehicle detection performance is enhanced by utilizing spatial pyramid pooling of various sizes for spatial information compensation. As illustrated in (a), (b) and (c) in Fig. 7,and (d), (e), and (f), the performance improves as more stacked pyramid pooling of different sizes is introduced. Furthermore, it is shown by the introduction of channel attention in PAN with a single pooling type that max pooling slightly outperforms average pooling. Notably, it is depicted in Fig. 7 that the combination of both max and average pooling yields better performance than using a single pooling type alone, as shown in (h), (i), and (j). However, performance can be improved by combining pooling operations of different sizes and types, but it comes at the expense of an increased number of model parameters. To tackle this issue, the fusion of features obtained from different pooling types of the same size was investigated by employing an additive operation before computing channel attention weights. Subsequently, these features were concatenated to extract attention, as illustrated in (j) and (k) in Fig. 7. Interestingly, this approach of fusing before calculating channel attention weights is adopted, which not only reduces the model's parameter count to some extent but also slightly improves its performance. After carefully balancing the model's parameter count and performance, we selected $PPAM_2$ as our channel attention

TABLE III
THE COMPARISON OF DIFFERENT CHANNEL ATTENTION METHODS

| Different Channel Attention | Backbone | mAP50(%) | mAP75(%) | mAP50:95(%) | Parameters(M) | FLOPs(G) |
|---|---|---|---|---|---|---|
| baseline | R-50-PAN | 84.08 | 75.84 | 59.50 | 42.89872 | 96.01633 |
| $C_{A1}$ | R-50-PAN | 84.17 | 76.13 | 59.69 | 43.59504 | 96.02932 |
| $C_{A2}$ | R-50-PAN | 84.36 | 76.32 | 59.81 | 53.34352 | 96.05135 |
| $C_{A3}$ | R-50-PAN | 84.65 | 76.54 | 60.15 | 203.74864 | 96.21405 |
| $C_{M1}$ | R-50-PAN | 84.31 | 76.15 | 59.81 | 43.59504 | 96.02932 |
| $C_{M2}$ | R-50-PAN | 84.53 | 76.28 | 59.93 | 53.34352 | 96.05135 |
| $C_{M3}$ | R-50-PAN | 84.57 | 76.57 | 60.24 | 203.74864 | 96.21405 |
| $C_{Mix1}$ | R-50-PAN | 84.52 | 76.32 | 60.10 | 44.98768 | 96.40300 |
| $C_{Mix2}$ | R-50-PAN | 84.71 | 76.57 | 60.28 | 81.19632 | 96.10378 |
| $C_{Mix3}$ | R-50-PAN | 84.85 | 76.58 | 60.42 | 676.672 | 96.71885 |
| Proposed $PPAM_2$ | R-50-PAN | 84.84 | 76.58 | 60.33 | 53.35120 | 96.07594 |
| Proposed $PPAM_3$ | R-50-PAN | **84.92** | **76.76** | **60.50** | 203.74864 | 96.25092 |

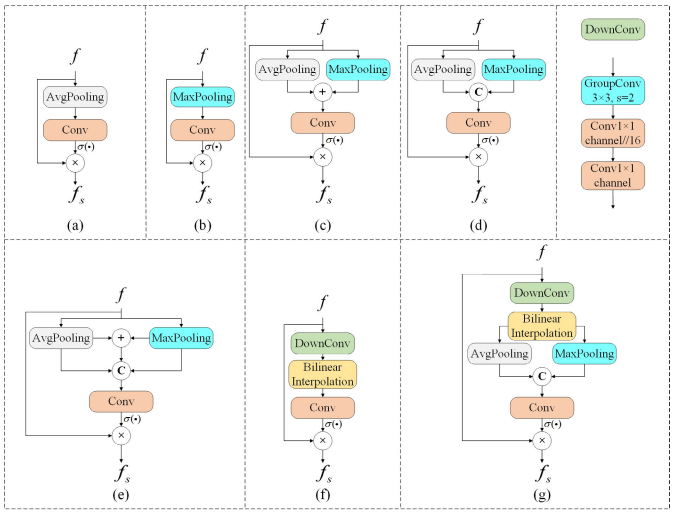


Fig. 8. The different spatial attention modules are presented in the following sequence: (a) $S_{avg}$ module, (b) $S_{max}$ module, (c) $S_{add}$ module, (d) $S_{concat}$ module, (e) Proposed DPAM module, (f) GEAM module, and (g) $GEAM_{mix}$.

module. Compared to $C_{Mix2}$, $PPAM_2$ reduces the parameter count by 32.31% while increasing mAP50 by 0.13%, mAP75 by 0.01%, and mAP50:95 by 0.05%.

*2) Dual Pooling Attention Module:* Channel attention is often complemented by spatial attention to achieve better performance [37], [38], [39]. When focusing on spatial attention, model performance may be enhanced through the utilization of larger convolutional kernels, drawing inspiration from [38]. Therefore, when designing our spatial attention module, the impact of larger convolutions is also taken into consideration by us. Prior to that, pertinent ablation experiments were performed on various spatial attention mechanisms illustrated in Fig. 8, employing $3\times3$ convolutions to identify the architecture yielding superior performance. The experimental results are presented in Table IV. It is suggested by the findings that improved performance can be achieved by combining both max-pooling and average-pooling. Specifically, the addition operation is outperformed by the concatenation operation in terms of information fusion. Furthermore, an alternative spatial attention method named gather-excite attention module (GEAM) [41] is introduced, and its superior performance compared to using a single pooling spatial attention module is demonstrated. Moreover, GEAM is extended to $GEAM_{mix}$, incorporating both max-pooling and average-pooling. In comparison to GEAM, mAP50 is enhanced by 0.3%, mAP75 is enhanced by 0.19%, and mAP50:95 is enhanced by 0.1% by $GEAM_{mix}$, illustrating the beneficial impact of pooling on computing attention weights. Lastly, drawing inspiration from our proposed dual pooling attention module (DPAM), a simple add operation is employed to append additional feature maps to $S_{concat}$ for attention calculation. Surprisingly, improved localization performance is shown by our proposed DPAM with the addition of a simple feature map, resulting in a 0.17% improvement in mAP75.

Based on various evaluation metrics, we conducted ablation experiments using DPAM, an extension of S*concat* with supplementary feature maps, involving different convolutional kernel sizes: $3\times3$, $5\times5$, $7\times7$, and $9\times9$. The results indicate that larger convolutions, particularly the $7\times7$ kernel, generally result in improved performance, exhibiting the best performance. Compared to the $3\times3$ convolution, the $7\times7$ convolution improves mAP50 by 0.56%, mAP75 by 0.20%, and mAP50:95 by 0.09%. Therefore, we choose the $7\times7$ kernel size for DPAM, serving as the spatial attention module for our model.

*3) Pyramid Dual Pooling Attention Path Aggregation Network:* Three distinct combination strategies for integrating two types of attention were investigated by us: serial channel-spatial, serial spatial-channel, and parallel. A summary of the comparative results for the three distinct implementations is presented in Table V. It has been verified by our experiments that the optimal performance is yielded by the parallel combination of diverse attentions. Regarding information transmission, irrelevant and redundant information is primarily suppressed by attention, with a weight between 0 and 1 assigned to each channel/pixel, favoring values closer to 1 for the attended objects and closer to 0 for

TABLE IV
THE COMPARISON OF DIFFERENT SPATIAL ATTENTION METHODS

| Different Spatial Attention | Backbone | Convolution Kernel Size | mAP50(%) | mAP75(%) | mAP50:95(%) | Parameters(M) | FLOPs(G) |
|---|---|---|---|---|---|---|---|
| baseline | R-50-PAN | - | 84.08 | 75.84 | 59.50 | 42.89872 | 96.01633 |
| $S_{avg}$ | R-50-PAN | 3×3 | 84.20 | 75.88 | 59.25 | 42.89876 | 96.01664 |
| $S_{max}$ | R-50-PAN | 3×3 | 84.29 | 75.89 | 59.46 | 42.89876 | 96.01664 |
| $S_{add}$ | R-50-PAN | 3×3 | 84.02 | 75.95 | 59.54 | 42.89876 | 96.01664 |
| $S_{concat}$ | R-50-PAN | 3×3 | 84.41 | 76.20 | 59.91 | 42.89876 | 96.01694 |
| GEAM | R-50-PAN | 3×3 | 84.06 | 76.10 | 59.54 | 43.63728 | 96.25984 |
| $GEAM_{mix}$ | R-50-PAN | 3×3 | 84.36 | 76.29 | 59.64 | 43.63735 | 96.26045 |
| Proposed DPAM | R-50-PAN | 3×3 | 84.42 | 76.37 | 59.94 | 42.89883 | 96.01725 |
| Proposed DPAM | R-50-PAN | 5×5 | 84.43 | 76.43 | 59.95 | 42.89902 | 96.01889 |
| Proposed DPAM | R-50-PAN | 7×7 | **84.98** | **76.57** | **60.03** | 42.89931 | 96.02133 |
| proposed DPAM | R-50-PAN | 9×9 | 84.48 | 76.43 | 59.97 | 42.89969 | 96.02459 |

TABLE V
THE COMPARISON OF DIFFERENT ATTENTION COMBINATION STRATEGIES

| Backbone | Combination Strategy | | | mAP50(%) | mAP75(%) | mAP50:95(%) |
|---|---|---|---|---|---|---|
| | *Serial(C+S)* | *Serial(S+C)* | *Parrallel* | | | |
| R-50-PAN | ✓ | | | 84.66 | 76.71 | 60.26 |
| R-50-PAN | | ✓ | | 84.68 | 76.89 | 60.55 |
| R-50-PAN | | | ✓ | **85.26** | **77.07** | **60.73** |

TABLE VI
THE COMPARISON OF DIFFERENT HYBRID ATTENTION METHODS

| Methods | Backbone | mAP50(%) | mAP75(%) | mAP50:95(%) | Parameters(M) | FLOPs(G) | FPS |
|---|---|---|---|---|---|---|---|
| baseline | R-50-PAN | 84.04 | 75.84 | 59.50 | 42.89872 | 96.01633 | 42.387 |
| CBAM* [38] | R-50-PAN | 83.98 | 75.92 | 59.54 | 43.63859 | 96.06644 | 34.212 |
| BAM* [39] | R-50-PAN | 85.07 | 77.03 | 60.43 | 44.41068 | 97.23978 | 41.490 |
| scSE* [37] | R-50-PAN | 84.50 | 76.73 | 60.38 | 49.46051 | 104.97716 | 40.516 |
| PDPA-PAN (ours) | R-50-PAN | 85.26 | 77.07 | 60.73 | 53.35179 | 96.08094 | 36.018 |
| PDPA-PAN*(ours) | R-50-PAN | **85.43** | **77.18** | **60.84** | 53.84594 | 96.12124 | 33.937 |

TABLE VII
THE COMPARISON OF DIFFERENT BACKBONE

| Methods | Backbone | mAP50(%) | mAP75(%) | mAP50:95(%) | Parameters(M) | FLOPs(G) | FPS |
|---|---|---|---|---|---|---|---|
| PDPA-PAN* | R-18-PAN | 81.92 | 72.06 | 56.30 | 31.04326 | 74.95499 | 42.170 |
| PDPA-PAN* | R-50-PAN | **85.43** | **77.18** | **60.84** | 53.84594 | 96.12124 | 33.937 |
| PDPA-PAN* | R-101-PAN | 84.90 | 76.66 | 60.08 | 72.83807 | 126.55043 | 28.839 |
| PDPA-PAN* | Swin-T-PAN | 84.42 | 75.57 | 59.71 | 48.48875 | 97.05491 | 30.124 |

the background. When a serial strategy is employed, certain features may be activated in the channel dimension but suppressed in the spatial dimension, resulting in the further suppression of the additional information due to the nature of the Sigmoid function. Other hybrid attention mechanisms, such as BAM [39], CBAM [38], and scSE [37], have been additionally tested, as indicated in Table VI. Among them, parallel strategies are adopted by both BAM and scSE, and their performance surpasses that of the serial strategy used in CBAM. Interestingly, in the serial strategy, results contrary to those in [38] were obtained. It is postulated that the performance of attention mechanisms will differ based on their positions within the model and the adopted combination strategies.

*4) Discussion:* Visualization examples of Grad-CAM [58] are depicted in Fig. 9. This underscores the ability of PDPA-PAN to concentrate on information pertinent to the foreground object, consequently enhancing performance in diverse challenging scenarios. Individual attention methods struggle to effectively concentrate on foreground objects. For instance, the proposed channel attention method PPAM exclusively addresses the contours of the foreground, whereas

TABLE VIII
DETECTION RESULTS OF DIFFERENT METHODS FOR THE LH-UAV-VEHICLE DATASET

| Methods | Backbone | Input Size | mAP50(%) | mAP75(%) | mAP50:95(%) | Parameters(M) | FLOPs(G) | FPS |
|---|---|---|---|---|---|---|---|---|
| $R^3$Det [49] | R-50-PAN | 640×640 | 75.53 | 68.02 | 52.96 | 48.29373 | 176.75693 | 28.678 |
| RsDet [54] | R-50-PAN | 640×640 | 75.85 | 65.32 | 52.29 | 37.36763 | 83.57702 | 52.364 |
| ATSS [55] | R-50-PAN | 640×640 | 81.55 | 73.68 | 57.74 | 37.20167 | 82.16221 | 56.951 |
| $S^2$ANet [51] | R-50-PAN | 640×640 | 82.80 | 74.11 | 57.25 | 39.7294 | 77.94153 | 46.880 |
| Faster RCNN(OBB) [48] | R-50-PAN | 640×640 | 76.14 | 61.54 | 50.77 | 42.89718 | 95.96375 | 43.456 |
| RoITransformer [50] | R-50-PAN | 640×640 | 83.56 | 75.81 | 59.15 | 56.81848 | 109.88503 | 35.021 |
| Gliding [56] | R-50-PAN | 640×640 | 83.05 | 66.14 | 53.28 | 42.90128 | 95.96785 | 45.619 |
| Oriented RCNN [52] (baseline) | R-50-PAN | 640×640 | 84.08 | 75.84 | 59.50 | 42.89872 | 96.01633 | 42.387 |
| PDPA-PAN* (ours) | R-50-PAN | 640×640 | **85.43** | **77.18** | **60.84** | 53.84594 | 96.12124 | 33.937 |

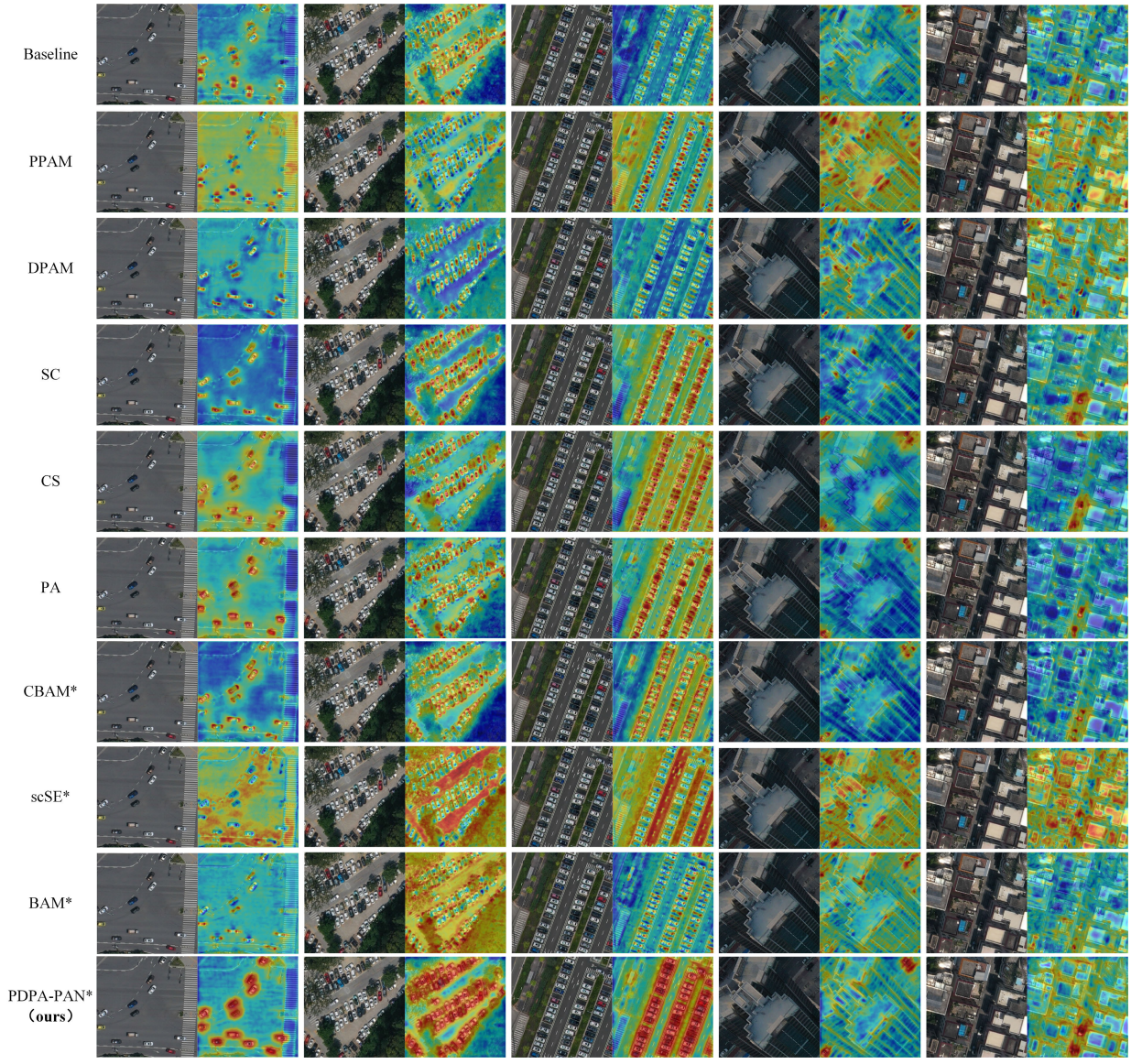


Fig. 9. Grad-CAM visualization is performed on the Oriented RCNN detector, employing PDPA-PAN along with several other hybrid attention methods. The PDPA-PAN proposed in our study has the capability to model contextual information that emphasizes the foreground, resulting in enhanced performance across diverse challenging scenarios.

the proposed spatial attention method DPAM singularly emphasizes the center of the foreground. Both are crucial in object detection tasks, where contour features aid in localization and central features contribute to classification. Furthermore, we have visually compared various hybrid attention combination methods along with related hybrid attention

(a) Detection results from baseline (b) Detection results from PDPA-PAN

Fig. 10. The qualitative analysis of LH-UAV-Vehicle was performed on PDPA-PAN. (a) Detection results from baseline. (b) Detection results from our proposed PDPA-PAN. The error detection are represented by the area that was selected by the yellow solid box in the figure.

methods. In Fig. 9, SC denotes the use of spatial attention DPAM followed by channel attention PPAM, CS denotes the use of channel attention PPAM followed by spatial attention DPAM, and PA denotes the parallel combination of two different attentions. PDPA-PAN* represents our proposed method, while CBAM*, scSE*, and BAM* denote alternative hybrid attention methods subjected to ablation experiments replacing our proposed PDPAM. Hybrid attention enables a concentration on pertinent information of the detection object, resulting from the synergy between suitable attention modules and combination methods. Additionally, we compared various hybrid attention methods, some of which may not effectively concentrate on foreground objects. Moreover, vsiualization results demonstrated that combining our proposed PPAM and DPAM in parallel yields improved highlighting of foreground features and suppression of background interference. Additionally, we explored various backbone networks for PDPA-PAN, such as ResNet18 [59], ResNet101 [59] and SwinTransformer-Tiny [60]. The specific results are shown in Table VII. Employing ResNet50 as the primary network led to optimal performance. However, backbone networks with either higher or lower computational complexities exhibited slightly inferior performance, potentially attributed to issues of model overfitting or underfitting. While larger backbone networks are typically anticipated to improve performance, excessively intricate architectures might capture subtle features and noise from the training set, compromising their ability to generalize to the test set. Conversely, smaller backbone networks, due to their simplicity, struggle to capture the essential data representations, making them unable to adapt to the intricacies of the data. In conclusion, the choice of backbone should align with the task's complexity. In our specific task, ResNet50 proves to be well-suited, effectively avoiding both overfitting and underfitting.

### C. Overall Performance

An mAP50 of 85.43% on the LH-UAV-Vehicle dataset is achieved by our designed PDPA-PAN framework, which integrates attention, deep semantic information, and spatial details. Additionally, a comparison is performed between excellent single-stage detection methods (including both anchor-based and anchor-free methods) and two-stage high-precision detection methods, utilizing evaluation metrics such as mAP75 and mAP50:95. The specific results are presented in Table VIII. Moreover, superior performance is achieved by embedding our proposed PDPAM on both sides of PAN. Additionally, similar hybrid attention methods, including BAM, CBAM, and scSE, are incorporated in a manner consistent with Table VI. The best performance is achieved when our proposed PDPA-PAN is combined with the top-performing two-stage detector, Oriented RCNN [52], which is noteworthy. All methods are trained under identical conditions, ensuring the reliability of the experimental results. The hybrid attention methods, namely BAM*, CBAM*, and scSE*, are incorporated in a similar manner for the best baseline. The addition of attention to both the input and output features of PAN is indicated by the * symbol, and it should be noted. Our method outperforms the baseline, BAM*, CBAM*, and scSE* by 1.35%, 0.36%, 0.93%, and 1.45% in mAP50, and by 1.34%, 0.15%, 0.45%, and 1.26% in mAP75, as well as by 1.34%, 0.41%, 0.46%, and 1.3% in mAP50:95.

The detection results on the LH-UAV-Vehicle dataset using our proposed method are displayed in Fig. 10. It is worth noting that all vehicles from an aerial perspective are covered by our LH-UAV-Vehicle dataset, encompassing various complex backgrounds and densely scenarios.

## VI. Conclusion and Future Work

The LH-UAV-Vehicle dataset for vehicle detection using UAVs flying at 250-400 meters was created in this paper, addressing the lack of vehicle detection data from UAVs at high-altitude. The pyramid dual pooling attention path aggregation network (PDPA-PAN) architecture was introduced by us to enhance vehicle localization. This architecture combines attention and path aggregation network (PAN), considering the high-altitude scenes that result in increased information density of vehicle objects and a wider field of view with more complex backgrounds. The spatial attention

and channel attention mechanisms have been extended in this study by introducing two distinct modules: pyramid pooling attention module (PPAM) and dual pooling attention module (DPAM). Additionally, a pyramid dual pooling attention module (PDPAM) was designed to combine these two different attention mechanisms. Moreover, PAN based on feature pyramid network (FPN) extension was employed to fuse deep semantic and spatial information from feature maps at different levels, resulting in more refined features. The effectiveness of the proposed framework (PDPA-PAN) and its internal modules has been validated through extensive experiments. State-of-the-art performance on the LH-UAV-Vehicle dataset was achieved by our proposed framework. While our PDPA-PAN is indeed deemed lightweight and efficient in terms of FLOPs, its influence on inference speed and parameter count is noteworthy. Future research will emphasize hardware optimization and the refinement of a more lightweight design for the proposed framework.

In this work, the main focus was on improving the accuracy of vehicle detection at high-altitude. As LH-UAV-Vehicle is collected from the real world, vehicle detection performance can be impacted by persisting long-tail problems. Furthermore, precise vehicle localization is essential for urbanization surveying. However, the method of erasing vehicle objects without disrupting the original image distribution has not yet been addressed by this paper. Moving forward, addressing the long-tail distribution problem in the data will be prioritized, and more effective methods to enhance robustness to tail data will be investigated. Secondly, the task of erasing vehicle objects is equally significant. Solving the challenge of enhancing the accuracy of urbanization survey tasks by mitigating the influence of vehicles is yet to be addressed. Moreover, video modal data can be utilized in diverse detection tasks, including dynamic traffic flow analysis and vehicle tracking. In our forthcoming endeavors, we will also explore the amalgamation of video modal data and its application in dynamic traffic flow analysis, a task of significance in intelligent city management. Furthermore, The meticulous partitioning of diverse scenarios significantly contributes to augmenting the comprehensiveness and universality of the dataset. This process aids in evaluating the model's performance in specific scenarios with greater precision. We intend to enhance the LH-UAV-Vehicle dataset in our future research, developing a vehicle detection dataset with fine-grained scenario distinctions. Ultimately, it is genuinely believed by us that valuable assistance will be offered to researchers in this field by the LH-UAV-Vehicle dataset, fostering the advancement of UAV-based vehicle detection in intelligent urban management.

## REFERENCES


[1] K. Guan, H. Yi, D. He, B. Ai, and Z. Zhong, "Towards 6G: Paradigm of realistic terahertz channel modeling," *China Commun.*, vol. 18, no. 5, pp. 1–18, May 2021.

[2] H. Jiang, M. Mukherjee, J. Zhou, and J. Lloret, "Channel modeling and characteristics for 6G wireless communications," *IEEE Netw.*, vol. 35, no. 1, pp. 296–303, Jan./Feb. 2021.

[3] T. Ye, W. Qin, Z. Zhao, X. Gao, X. Deng, and Y. Ouyang, "Real-time object detection network in UAV-vision based on CNN and transformer," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.

[4] S. Wang, F. Jiang, B. Zhang, R. Ma, and Q. Hao, "Development of UAV-based target tracking and recognition systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3409–3422, Aug. 2020.

[5] P. Zhu, J. Zheng, D. Du, L. Wen, Y. Sun, and Q. Hu, "Multi-drone-based single object tracking with agent sharing network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4058–4070, Oct. 2021.

[6] L. Sun et al., "Aerial-PASS: Panoramic annular scene segmentation in drone videos," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Bonn, Germany, Aug. 2021, pp. 1–6.

[7] R. Makrigiorgis, C. Kyrkou, and P. Kolios, "How high can you detect? Improved accuracy and efficiency at varying altitudes for aerial vehicle detection," in *Proc. Int. Conf. Unmanned Aircr. Syst. (ICUAS)*, Warsaw, Poland, Jun. 2023, pp. 167–174.

[8] P. Chen, Y. Dang, R. Liang, W. Zhu, and X. He, "Real-time object tracking on a drone with multi-inertial sensing data," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 131–139, Jan. 2018.

[9] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[10] H. Fang, S. Guo, X. Wang, S. Liu, C. Lin, and P. Du, "Automatic urban scene-level binary change detection based on a novel sample selection approach and advanced triplet neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5601518.

[11] J. Yu, H. Gao, J. Sun, D. Zhou, and Z. Ju, "Spatial cognition-driven deep learning for car detection in unmanned aerial vehicle imagery," *IEEE Trans. Cogn. Devel. Syst.*, vol. 14, no. 4, pp. 1574–1583, Dec. 2022.

[12] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1758–1770, Jun. 2020.

[13] X. Li, X. Li, Z. Li, X. Xiong, M. O. Khyam, and C. Sun, "Robust vehicle detection in high-resolution aerial images with imbalanced data," *IEEE Trans. Artif. Intell.*, vol. 2, no. 3, pp. 238–250, Jun. 2021.

[14] Z. Zhang, Y. Liu, T. Liu, Z. Lin, and S. Wang, "DAGN: A real-time UAV remote sensing image vehicle detection framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1884–1888, Nov. 2020.

[15] A. Bouguettaya, H. Zarzour, A. Kechida, and A. M. Taberkit, "Vehicle detection from UAV imagery with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6047–6067, Nov. 2022.

[16] Z. Wang, J. Zhan, C. Duan, X. Guan, P. Lu, and K. Yang, "A review of vehicle detection techniques for intelligent vehicles," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 3811–3831, Aug. 2023.

[17] J. Shen, W. Zhou, N. Liu, H. Sun, D. Li, and Y. Zhang, "An anchor-free lightweight deep convolutional network for vehicle detection in aerial images," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24330–24342, Dec. 2022.

[18] H. Xu, M. Guo, N. Nedjah, J. Zhang, and P. Li, "Vehicle and pedestrian detection algorithm based on lightweight YOLOv3-promote and semi-precision acceleration," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 19760–19771, Oct. 2022.

[19] R. Liu, Z. Yuan, and T. Liu, "Learning TBox with a cascaded anchor-free network for vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 321–332, Jan. 2022.

[20] M. Hassaballah, M. A. Kenk, K. Muhammad, and S. Minaee, "Vehicle detection and tracking in adverse weather using a deep learning framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4230–4242, Jul. 2021.

[21] T. Ye, W. Qin, Y. Li, S. Wang, J. Zhang, and Z. Zhao, "Dense and small object detection in UAV-vision based on a global-local feature enhanced network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.

[22] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Y. Li, "Soft-Weighted-Average ensemble vehicle detection method based on single-stage and two-stage deep learning models," *IEEE Trans. Intell. Vehicles*, vol. 6, no. 1, pp. 100–109, Mar. 2021.

[23] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 375–391.

[24] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 445–461.

[25] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 549–565.

[26] D. Du et al., "VisDrone-DET2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 213–226.

[27] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, Sep. 2015.

[28] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, Oct. 2022.

[29] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf.*, Amsterdam, The Netherlands, Oct. 2016, pp. 21–37.

[30] 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.

[31] 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.

[32] Y. Wang et al., "An energy-efficient classification system for peach ripeness using YOLOv4 and flexible piezoelectric sensor," *Comput. Electron. Agricult.*, vol. 210, Jul. 2023, Art. no. 107909.

[33] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 7464–7475.

[34] K. Yang, J. Zhang, S. Reiß, X. Hu, and R. Stiefelhagen, "Capturing omni-range context for omnidirectional segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1376–1386.

[35] Q. Song, K. Mei, and R. Huang, "AttaNet: Attention-augmented network for fast and accurate scene parsing," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 2567–2575.

[36] J. Wang et al., "Superpixel segmentation with squeeze-and-excitation networks," *Signal, Image Video Process.*, vol. 16, no. 5, pp. 1161–1168, Jul. 2022.

[37] C. Wang, A. Bochkovskiy, and H. Liao, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Med. Image Comput. Comput. Assist. Intervent.*, Granada, Spain, 2018, pp. 421–429.

[38] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 3–19.

[39] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Newcastle, U.K., 2018, p. 147.

[40] J. Guo et al., "SPANet: Spatial pyramid attention network for enhanced image recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[41] J. Hu, L. Shen, and S. Albanie, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. NeurIPS*, Montréal, QC, Canada, 2018, pp. 9423–9433.

[42] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 510–519.

[43] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 16748–16759.

[44] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, and L. Yuan, "DaViT: Dual attention vision transformers," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel, 2022, pp. 74–92.

[45] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016.

[46] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 3974–3983.

[47] 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)*, Venice, Italy, Oct. 2017, pp. 2999–3007.

[48] 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.

[49] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, Venice, Italy, 2021, pp. 3163–3171.

[50] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 2849–2858.

[51] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602511.

[52] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Motreal, QC, Canada, Oct. 2021, pp. 3500–3509.

[53] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, "An image is worth $16\times16$ words: Transformers for image recognition," in *Proc. Int. Conf. Learn. Represent.*, Virtual Event, Austria, May 2021.

[54] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," in *Proc. AAAI Conf. Arti. Intell.*, May 2021, vol. 35, no. 3, pp. 2458–2466.

[55] 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)*, Seattle, WA, USA, Jun. 2020, pp. 9756–9765.

[56] Y. Xu et al., "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 4, pp. 1452–1459, Apr. 2021.

[57] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[58] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[60] Z. Liu, Y. Lin, and Y. Cao, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002.

**Zilu Ying** received the B.S., M.S., and Ph.D. degrees in electrical information engineering from Beihang University, Beijing, China, in 1985, 1988, and 2009, respectively. He is currently a Full Professor with Wuyi University, Jiangmen, China. He is also the Executive Director of Guangdong Society of Image and Graphics. His research interests include biometric extraction and pattern recognition. He is a member of the Signal Processing Branch, Chinese Institute of Electronics.

**Jianhong Zhou** received the B.S. degree from the Wuyi University of Communication Engineering in 2020. He is currently pursuing the master's degree with the Department of Intelligence Manufacturing, Wuyi University. His research interests include computer vision, representational contrastive learning, image processing, and object detection.

**Yikui Zhai** (Senior Member, IEEE) received the bachelor's degree in optical electronics information and communication engineering and the master's degree in signal and information processing from Shantou University, Shantou, China, in 2004 and 2007, respectively, and the Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013. Since October 2007, he has been with the Department of Intelligence Manufacturing, Wuyi University, Jiangmen, China, where he is currently a Professor. Since 2016, he has also been a Visiting Scholar with the Department of Computer Science, University of Milan, Milan, Italy. His research interests include image processing, deep learning, and pattern recognition.

**Hao Quan** received the bachelor's and master's degrees in computer science from the Università degli Studi di Milano, in 2014 and 2017, respectively, and the Ph.D. degree in computer science and engineering from Politecnico di Milano, Milan, Italy, in 2022. Since July 2023, he has been a Post-Doctoral Researcher with Politecnico di Milano. His current research interests include artificial intelligence, robotic technology, and machine learning.

**Wenba Li** (Student Member, IEEE) received the B.S. degree from Guangdong University of Technology, Guangzhou, China, in 2020. He is currently pursuing the master's degree with the Department of Intelligence Manufacturing, Wuyi University, Jiangmen, China. His research interests include image classification, change detection, and pattern recognition.

**Angelo Genovese** (Senior Member, IEEE) received the Ph.D. degree in computer engineering and automation from the Politecnico di Milano, Milano, Italy.

He is currently a Full Professor with the Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 papers in international journals, proceedings of international conferences, books, and book chapters. His main research interests include pattern recognition, machine learning, machine perception, robotics, computer vision, signal and image processing, three-dimensional reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.

Dr. Genovese is an Associate Editor of the *Journal of Ambient Intelligence and Humanized Computing* (Springer).

**Vincenzo Piuri** (Fellow, IEEE) received the M.S. and Ph.D. degrees in computer engineering from the Politecnico di Milano, Milan, Italy, in 1984 and 1988, respectively.

He was the Department Chair with the University of Milan, Milan, from 2007 to 2012, where he has been a Full Professor since 2000. He was an Associate Professor with the Politecnico di Milano from 1992 to 2000, a Visiting Professor with The University of Texas at Austin, Austin, TX, USA, from 1996 to 1999, and a Visiting Researcher with George Mason University, Fairfax, VA, USA, from 2012 to 2016. He founded a startup company, Sensure srl, Bergamo, Italy, in the area of intelligent systems for industrial applications (leading it from 2007 to 2010). He was active in industrial research projects with several companies.

Dr. Piuri is an ACM Fellow.

**Fabio Scotti** (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milano, Milan, Italy, in 2003.

He is an Associate Editor of IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS and *Soft Computing* (Springer). He has been an Associate Editor of IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY and a Guest Co-Editor of IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT.