






Learning to Holistically Detect Bridges From Large-Size VHR Remote Sensing Imagery

Yansheng Li , Senior Member, IEEE, Junwei Luo , Yongjun Zhang , Member, IEEE, Yihua Tan , Member, IEEE, Jin-Gang Yu , and Song Bai 

Abstract—Bridge detection in remote sensing images (RSIs) plays a crucial role in various applications, but it poses unique challenges compared to the detection of other objects. In RSIs, bridges exhibit considerable variations in terms of their spatial scales and aspect ratios. Therefore, to ensure the visibility and integrity of bridges, it is essential to perform holistic bridge detection in large-size very-high-resolution (VHR) RSIs. However, the lack of datasets with large-size VHR RSIs limits the deep learning algorithms' performance on bridge detection. Due to the limitation of GPU memory in tackling large-size images, deep learning-based object detection methods commonly adopt the cropping strategy, which inevitably results in label fragmentation and discontinuous prediction. To ameliorate the scarcity of datasets, this paper proposes a large-scale dataset named GLH-Bridge comprising 6,000 VHR RSIs sampled from diverse geographic locations across the globe. These images encompass a wide range of sizes, varying from $2,048 \times 2,048$ to $16,384 \times 16,384$ pixels, and collectively feature 59,737 bridges. These bridges span diverse backgrounds, and each of them has been manually annotated, using both an oriented bounding box (OBB) and a horizontal bounding box (HBB). Furthermore, we present an efficient network for holistic bridge detection (HBD-Net) in large-size RSIs. The HBD-Net presents a separate detector-based feature fusion (SDFF) architecture and is optimized via a shape-sensitive sample re-weighting (SSRW) strategy. The SDFF architecture performs inter-layer feature fusion (IFF) to incorporate multi-scale context in the dynamic image pyramid (DIP) of the large-size image, and the SSRW strategy is employed to ensure an equitable balance in the regression weight of bridges with various aspect ratios. Based on the proposed GLH-Bridge dataset, we establish a bridge detection benchmark including the OBB and HBB tasks, and validate the effectiveness of the proposed HBD-Net. Additionally, cross-dataset generalization experiments on two publicly available datasets illustrate the strong generalization capability of the GLH-Bridge dataset.

Index Terms—Bridge detection benchmark, deep network, large-size imagery, very -high-resolution (VHR).

I. INTRODUCTION

BRIDGES represent critical infrastructure components, serving as fundamental transportation facilities that traverse various landscapes. They hold substantial significance in the domains of civil transportation, military maneuvers, and disaster relief efforts [1]. Meanwhile, bridges exhibit rapid construction and frequent modification. For example, in 2012, the United States had about 617,000 bridges whose deterioration will increase over the next 50 years, requiring more than \$125 billion for a backlog of repairs.¹ Therefore, efficient and effective bridge detection is of paramount importance to the timely update of the navigation map and further contributes to monitoring the structural health and condition of bridges [2], [3]. Remote Sensing Images (RSIs), with their extensive geographic coverage and high revisit frequency, are well-suited as the foundational data for bridge detection. Meanwhile, considering the powerful feature representation abilities of deep networks, deep learning-based bridge detection from RSIs holds substantial promise and has become a focal point of research [4].

As illustrated in Fig. 1, detecting multi-scale bridges in RSIs is quite challenging compared to other common objects, primarily due to two main characteristics: (i) **diverse object scales**. In VHR RSIs, the lengths of bridge instances vary from a few to several thousand pixels. (ii) **extreme aspect ratios**. There are significant variances in the degree of elongation among different bridges. To ensure the detectability of small or narrow bridges, the utilization of very-high-resolution (VHR) images is crucial. At the same time, to pursue the structural integrity of large and elongated bridges in VHR images, it is essential to conduct holistic bridge detection in large-size images, which imposes strict requirements on both datasets and methods. Despite notable advancements in multi-class object detection [12], [13], [14], [15], [16] and bridge detection [4], [11], [17], there remains a deficiency in large-scale datasets and appropriate methods for holistic bridge detection in large-size VHR RSIs.

As shown in Table I, although numerous popular datasets for object detection in RSIs have been created [6], [7], [8], [18], the quantity of bridges within these datasets is limited. Furthermore, datasets explicitly created for bridge detection [4],

Manuscript received 21 November 2023; revised 12 March 2024; accepted 21 April 2024. Date of publication 29 April 2024; date of current version 5 November 2024. This work was supported by the National Natural Science Foundation of China under Grant 42371321 and Grant 42030102. Recommended for acceptance by B. Rosenhahn. (Corresponding authors: Junwei Luo; Yongjun Zhang.)

Yansheng Li, Junwei Luo, and Yongjun Zhang are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: yansheng.li@whu.edu.cn; luojunwei@whu.edu.cn; zhangyj@whu.edu.cn).

Yihua Tan is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: yhtan@hust.edu.cn).

Jin-Gang Yu is with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510641, China (e-mail: jingangyu@scut.edu.cn).

Song Bai is with ByteDance AI Lab, Beijing 100098, China (e-mail: song-bai.site@gmail.com).

The dataset and source code will be released at <https://luo-z13.github.io/GLH-Bridge-page/>.

Digital Object Identifier 10.1109/TPAMI.2024.3393024

¹[Online]. Available: <https://infrastructurereportcard.org/cat-item/bridges-infrastructure>

TABLE I
COMPARISON BETWEEN GLH-BRIDGE AND THE OTHER RELEVANT BRIDGE DETECTION DATASETS

Dataset	Number of images	Image size	GSD	Number of instances	Annotation type	Diverse backgrounds	Data source
Bridge subset of multi-class object detection dataset							
NWPU VHR-10 [5]	124	497×693~606×1,100	0.08~2	124	HBB	×	multi-source
FAIR1M [6]	581	1,000×1,000~10,000×10,000	0.3~0.8	1,008	OBB	×	GF, GoogleEarth
DOTA-v1.0 [7]	288	800×800~4,000×4,000	0.5	2,541	OBB	×	multi-source
DOTA-v2.0 [8]	382	800×800~20,000×20,000	0.5	3,043	OBB	×	multi-source
DIOR-R [9]	1,576	800×800	0.5~30	4,000	OBB	✓	Google Earth
HRRSD [10]	4,570	152×152~10,569×10,569	0.15~1.2	4,570	HBB	×	Google Earth
Dedicated bridge detection dataset							
Bridges Dataset [11]	208	4,800×2,843	0.5	322	HBB	✓	Google Earth
BridgeDetV1 [4]	4,972	668×668~1,000×1,000	2~6	8,371	OBB/HBB	×	GF, GoogleEarth
GLH-Bridge (Ours)	6,000	2,048×2,048~16,384×16,384	0.3~1	59,737	OBB/HBB	✓	Google Earth, Mapbox

Only the bridge category is selected for comparison among the multi-class object detection datasets. The comparison includes the number of images, image size, ground sampling distance (GSD), the number of instances, annotation type, backgrounds, and data source.

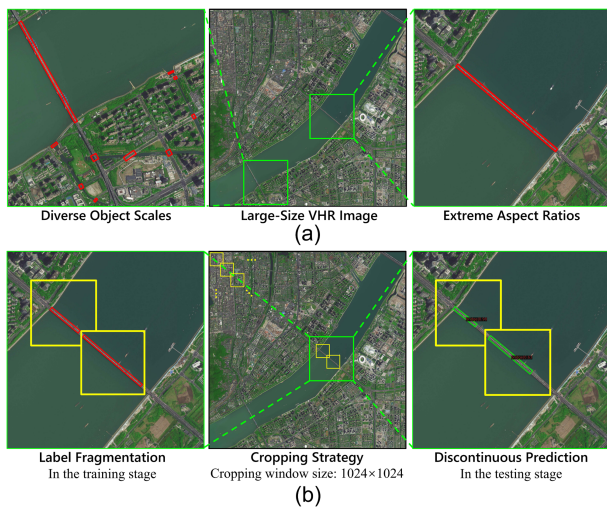


Fig. 1. The main characteristics of bridges impose strict requirements on both image resolution and size for bridge detection, as illustrated in (a). When tackling large-size images, the mainstream cropping strategy results in inaccurate labels and predictions. In (b), yellow windows denote the sliding windows (i.e., cropping windows), while red OBBs denote the labels and green OBBs show the prediction results.

[11] are often constrained by sample volumes and image sizes. Some of the existing datasets only provide horizontal bounding box (HBB) annotations instead of the accurate oriented bounding box (OBB) annotations. Therefore, training a robust and generalizable bridge detection model using the aforementioned datasets seems to be unrealistic. To address the data constraints, we construct **GLH-Bridge**, a large-scale dataset for bridge detection in large-size VHR RSIs. GLH-Bridge contains 6,000 VHR RSIs sampled globally and over 59 k manually annotated bridges. Compared with existing datasets for bridge detection, GLH-Bridge stands out by annotating multi-scale bridges in large-size VHR RSIs that encompass various background types such as *vegetation*, *dry riverbeds*, and *roads*, thereby better capturing the characteristics of bridges in real-world scenarios. In short, the GLH-Bridge exhibits comprehensive advantages and notable merits compared with existing bridge detection datasets.

To advance the research on the fundamental and practical issue, we propose a new challenging yet meaningful task: **holistic bridge detection in large-size VHR RSIs**. To address this task, the potential solutions can be categorized into four main aspects: **(i)** Given the constraints of GPU memory, mainstream deep learning-based object detection methods [15], [16], [19], [20], [21] commonly employ cropping strategies [7], [22]. However, such strategies have inherent limitations and easily cut off large bridges, as shown in Fig. 1. In addition to the cropping strategy, several object detection methods tackle the original large-size images with fixed-window downsampling strategies [23], [24], [25], resulting in a significant loss of image information; **(ii)** Methods like streaming [26] perform the forward and backward pass on smaller tiles of the large-size image, but they are unable to support deep neural network (DNN) with normalization; **(iii)** Methods like LMS [27] use memory offload to share memory across system memory (CPU DRAM) and the GPU memory. However, they introduce significant time overhead and are constrained by the maximum memory expansion rate; **(iv)** Multi-GPU tensor parallelization techniques [28], [29] have the promise to extend deep networks to support holistic processing of large-size images. However, they tend to be resource-intensive and difficult to operate in regular conditions. In summary, existing methods are ineffective under common computational resources (e.g., a single GPU with 24 GB memory) for holistic bridge detection in large-size VHR RSIs.

Considering the limitations of the aforementioned potential solutions, we propose a **holistic bridge detection network (HBD-Net)** specifically designed for bridge detection in large-size VHR RSIs. Our method presents two key merits: **(i)** The separate detector-based feature fusion (SDFF) architecture, when applied to the dynamic image pyramid (DIP), demonstrates an efficient approach for processing large-size images with minimal resource consumption. **(ii)** The shape-sensitive sample re-weighting (SSRW) strategy balances regression weights of bridges with different aspect ratios. Experimental results on GLH-Bridge demonstrate an outstanding performance of our proposed HBD-Net.

To sum up, to the best of our knowledge, this paper makes the first exploration of holistic bridge detection in large-size VHR

RSIs. The main contributions of this paper are summarized as follows:

- We propose GLH-Bridge, the first large-scale dataset for bridge detection in large-size VHR RSIs. With 59,737 bridges set against various backgrounds, this dataset offers a comprehensive representation of bridges in real-world scenarios.
- A cost-saving network for holistic bridge detection in large-size images (i.e., HBD-Net) is proposed, which can efficiently handle large-size images with the common GPU and holistically detect multi-scale bridges with the well-designed SDFP architecture and SSRW strategy.
- Using the proposed GLH-Bridge dataset, we create a benchmark for bridge detection, covering both the OBB and HBB tasks. The HBD-Net achieves superior performance compared to existing state-of-the-art algorithms. Furthermore, we conduct cross-dataset generalization experiments to demonstrate the strong generalization ability of GLH-Bridge. We hope this benchmark can contribute to the fundamental evaluation of object detection in large-size images.

The rest of this paper is organized as follows: Section II provides an overview of existing datasets and algorithms for bridge detection. Section III offers a detailed description of the proposed GLH-Bridge dataset. In Section IV, we introduce the proposed HBD-Net. Section V presents the experimental results. Finally, Section VI concludes the paper and provides insights for future work.

II. RELATED WORK

In this section, we first discuss available datasets for bridge detection. Next, we briefly review bridge detection methods and potential methods from relevant fields for object detection in large-size images.

A. Datasets for Bridge Detection in Remote Sensing Images

As shown in Table I, the current datasets utilized for bridge detection can be classified into two main categories: multi-class datasets encompassing the bridge category among others, and specialized datasets explicitly tailored for bridge detection purposes.

1) *Multi-Class Datasets for Bridge Detection*: In the literature, numerous large-scale and high-quality remote sensing object detection datasets have been proposed. For example, NWPU VHR-10 [5] is a dataset with ten categories, expanding the category of geospatial objects. DOTA [7] and DIOR [18] have raised the number of instances to a new level, reflecting the prevalence of multi-class objects in remote sensing scenes. FAIR1M [6] accomplishes a more detailed classification taxonomy of geospatial objects. Despite these datasets containing the bridge category, they have limited quantities of bridge instances. As summarized in Table I, these multi-class datasets are unable to fulfill the aforementioned three criteria of an ideal bridge detection dataset: **large volume of samples**, **large-size image**, and **VHR image**.

It has been clearly shown that the existing multi-class object detection benchmarks [8], [18] and some algorithms designed for enhancing oriented object detection [30], [31], [32] demonstrate that bridge is one of the most difficult categories to detect. For example, in DOTA-v1.0 and DOTA-v1.5 [7], the highest accuracies for the bridge category in the OBB task are 64.5% and 59.6%, respectively, which are obviously lower than the other classes.² Bridge detection, particularly in the OBB task, undoubtedly poses significant challenges. Therefore, addressing the shortage of large-scale bridge detection datasets is crucial to training high-performance bridge detection models.

2) *Specialized Datasets for Bridge Detection*: Besides multi-class object detection datasets of aerial images, researchers have developed diverse remote sensing datasets for one specific category to facilitate more adaptable and crucial single-class object detection. As shown in Table I, there exist two publicly available datasets [4], [11], which are specifically designed for bridge detection in RSIs.

Bridges Dataset [11]: Keiller et al. proposed the first dataset for bridge detection and identification in VHR RSIs, known as Bridges Dataset. This dataset comprises 500 images with a consistent size of $4,800 \times 2,843$ pixels. The image in this dataset has a spatial resolution of 0.5 m, aligning with the VHR criteria in remote sensing scenarios. It is sampled globally using ArcGIS³ and annotated bridges across different types of background terrains. However, the dataset has certain limitations. It is constrained by the relatively low number of instances and offers coarse HBB annotations for the bridges. Furthermore, the bridges are primarily located at the center of the image in this dataset, which may potentially distort the learning process for bridge detection models by prior biases.

BridgeDetV1 [4]: Guo et al. constructed a bridge detection dataset named BridgeDetV1 for detecting waterborne bridges in RSIs. The dataset consists of 5,000 images with the spatial resolution ranging from 2 ~ 6 meters and image size ranging from $668 \times 668 \sim 1,000 \times 1,000$ pixels. It encompasses a total of 8,371 bridges annotated with both HBB and OBB. Although BridgeDetV1 contains a larger number of bridges compared to previous datasets, its limited spatial resolution restricts its ability to detect small bridges. Furthermore, BridgeDetV1 only focuses on waterborne bridges, resulting in a lack of scene diversity.

As a whole, existing dedicated datasets for bridge detection are insufficient to reflect the characteristics of bridges in real-world scenarios. Therefore, it is urgent to build a comprehensive, large-scale bridge detection dataset with large-size VHR images and rich instance types.

B. Bridge Detection in Large-Size Remote Sensing Imagery

To motivate holistic bridge detection in large-size images, we discuss methods for bridge detection in RSIs and potential technologies to cope with object detection in large-size images in the following sections.

1) *Bridge Detection in Remote Sensing Imagery*: Bridge detection in RSIs is a longstanding research topic. Chaudhuri

²[Online]. Available: <https://captain-whu.github.io/DOTA/results.html>

³[Online]. Available: <https://www.arcgis.com/>

et al. [33] utilized traditional supervised classification techniques and prior knowledge to detect bridges from multi-spectral images. Sithole et al. [1] focused on bridge detection in airborne scenes by detecting the cross-sectional contours of bridges. Several traditional algorithms were also developed to detect bridges in synthetic aperture radar (SAR) images based on edge and geometric features of bridges or water bodies [34], [35], [36]. Generally, these methods mainly relied on hand-crafted features by exploiting the bridges' geometry structure and the context of the surrounding water bodies.

Recently, some deep learning-based methods for bridge detection in RSIs have been proposed. Chen et al. [37] incorporated attention modules to perform waterborne bridge detection. Guo et al. [4] introduced the prior information of water bodies and combined bridge detection with the auxiliary task of water body segmentation. Wang et al. [38] designed a module for injecting water prior information into the bridge detection task through binary segmentation maps. Some other researchers [17] used multi-feature fusion methods to perform bridge detection. However, these methods primarily concentrated on detecting bridges in small-size or low-resolution images. It is noted that the existing methods disproportionately prioritized water features for locating bridges, even though bridges span across diverse terrains. This overemphasis on water body information has caused biases in feature learning and failed to present practical scenarios. Hence, generalized bridge detection algorithms are still much underexplored.

2) *Object Detection in Large-Size Imagery*: In this section, we introduce methods designed for object detection in large-size images and methods borrowed from related fields that may have potential applications in tackling large-size images. It is worth noting that in large-size VHR images, object detection is more challenging than other tasks like semantic segmentation [39], [40], [41], [42], [43], [44] or style transfer [45], as the latter focuses on pixel-level details, while the former operates at the instance level.

In the field of object detection, cropping strategies like SAHI [22] are commonly used to handle large-size images in popular benchmarks [6], [7]. However, the use of the cropping strategy poses a significant risk of cutting off large bridges. This can lead to misalignment of the supervision signal and loss of contextual information. Moreover, some approaches have been proposed to detect objects in large-size images by downsampling the original image using a fixed size or resolution. Chen et al. [25] proposed a coupled global-local object detection network with two branches inspired by global-local networks for segmentation [39]. Deng et al. [23] utilized a global-local self-adaptive network to conduct drone-view object detection in large-size images via downsampling and self-adaptive cropping. However, as mentioned in Section I, such methods are not suitable for handling large-size images and can easily result in significant information loss.

Some potential deep learning-based technologies [46], [47] can be found in the literature to handle large-size images. Pinckaers et al. proposed streaming [26], which constructs the later activations by streaming the input image through the CNN in a tiled fashion, but it is unable to support DNN with normalization despite the fact that the normalization is a critical dependency

in modern DNNs. Le et al. [27] proposed an approach based on formal rules for graph rewriting, which is able to automatically manage GPU memory to save memory usage. However, it is restricted by the maximum memory expansion and often consumes significant computational time. Additionally, Shazeer et al. [28] proposed Mesh-TensorFlow for distributed tensor computations and data parallelism to address the memory problem (e.g., memory limitation of GPU). Nevertheless, Mesh-TensorFlow usually requires extensive computing resources, making them unfriendly for deployment on edge-computing devices. As a whole, it is not straightforward to extend the aforementioned methods to address holistic bridge detection in large-size RSIs.

Hence, it is essential to develop a cost-effective approach for bridge detection that efficiently handles large-size VHR images with common GPU hardware.

III. GLH-BRIDGE DATASET

Our goals for developing a new dataset for bridge detection are twofold: (i) to occupy the niche of large-scale datasets for bridge detection in large-size VHR RSIs. (ii) to promote a new meaningful yet challenging task: holistic bridge detection in large-size VHR RSIs. This section provides a comprehensive overview of the GLH-Bridge dataset, focusing on three key aspects: data collection, data annotation, and data analysis.

A. Data Collection

Considering the variations in imaging perspectives of RSIs and to increase data diversity, we collect images from multiple satellite sensor platforms such as Google Earth and MapBox. The GLH-Bridge dataset provides global coverage through the collection of 6,000 optical RSIs obtained from over 400 cities or regions covering Asia, Africa, South America, North America, Europe, and Oceania. The images are collected from 2019 to 2022, with the image size ranging from $2,048 \times 2,048$ pixels to $16,384 \times 16,384$ pixels, and spatial resolution varying from 0.3 m to 1.0 m. The overall distribution and some samples from the dataset are illustrated in Fig. 2.

To comprehensively obtain RSIs depicting bridges on a global scale, we employ two distinct methodologies to identify candidate areas for image acquisition. Firstly, we utilize meta-information regarding bridges sourced from the National Bridge Inventory (NBI),⁴ an extensive database curated by the Federal Highway Administration. The NBI includes comprehensive details on bridges throughout the United States, including various types such as *highway*, *railway*, *waterborne bridges*, and *tunnels*. Subsequently, we preprocess the acquired data by filtering entries based on their construction years, thereby excluding excessively antiquated bridge structures. Finally, we utilize Google Earth for image download. To ensure spatial randomness and mitigate the potential concentration of bridges within the central portions of the images, we define random spatial windows using geographic coordinates during the download process.

The other approach entails the identification of candidate geographic regions utilizing electronic maps and satellite imagery spanning a global scope, with the exclusion of the United States,

⁴[Online]. Available: <https://www.fhwa.dot.gov/bridge/nbi.cfm>

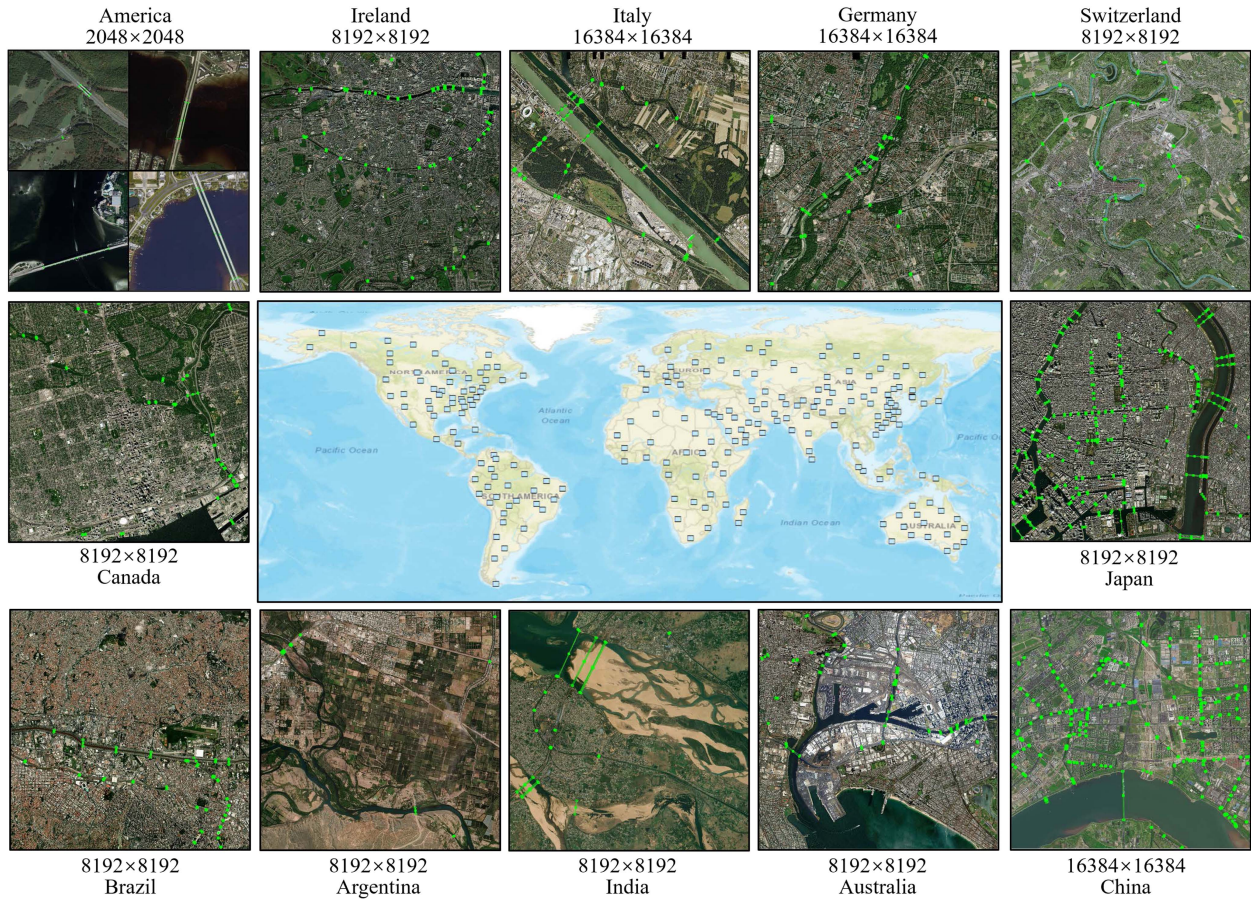


Fig. 2. The geographical distribution map of the sampled images from the proposed GLH-Bridge dataset.

to ensure a comprehensive and evenly distributed sampling area. In the selection of regions of interest, particular emphasis is placed on sampling within major urban centers. Initially, we utilize data pertaining to the locations of major cities and significant rivers in each country to compile a roster of cities exhibiting a high density of potential bridge structures. Subsequently, a random sampling strategy is applied to select areas within fixed-size regions within each city's geographic confines. Finally, RSIs are acquired from candidate geographic regions displaying diverse terrain characteristics and bridge contexts across varied regions. Notably, this methodology incorporates the collection of negative samples from regions with sparse bridge infrastructure, such as rural expanses, islands, and desert areas, to ensure uniformity in geographic distribution and dataset diversity.

With the purpose of leveraging the complementary geographic coverage via the aforementioned two image collection approaches, we partition the overall dataset randomly into training, validation, and testing sets with a ratio of 6:2:2. More specifically, the training, validation, and testing sets consist of 3613, 1194, and 1193 large-size images, respectively.

B. Data Annotation

1) *Annotation Criteria*: The geographical entity “bridge” is defined by considering both the structure and the spatial

context. In this vein, our visual interpretation process adheres to a stringent differentiation between bridges and roads. When dealing with suspended roads that cast shadows, we determine the two endpoints of one bridge based on the observation of whether they intersect distinct topographic features like valleys, rivers, or vegetation or not. This approach is crucial to ensure the exactitude of bridge labeling, with specific emphasis placed on the verification of objects that are susceptible to ambiguity, such as overpasses lacking topographical intersections or roads traversing regions between rice paddies.

The application of labeling criteria is illustrated in Fig. 3. Objects deemed to be non-bridges or bridges presenting challenges in labeling are deliberately omitted from the labeling process, such as two terminal connections shown in Fig. 3(a). Fig. 3(b) shows a road across the water with excessive curvature or an irregular shape that will not be labeled. In the process of annotating bridges, we establish the length threshold as **12 pixels** according to the size of *extremely Small* in [48], whereby bridges shorter than this threshold will not be labeled. It should be mentioned that this approach incorporates bridges with a width less than the length threshold into the dataset, thereby introducing a notable challenge in the detection of diminutive instances.

2) *Annotation Management*: The procedure of labeling GLH-Bridge encompasses a tripartite framework consisting of three stages: **pre-annotation stage**, **expert feedback and refinement stage**, and **large-scale detailed annotation stage**. In

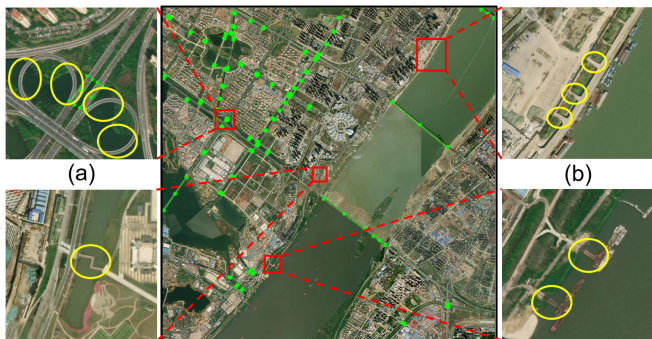


Fig. 3. Examples of labeling according to the criteria, with the yellow circle indicating scenarios that are not annotated. (a) Roads across water with excessive curvature or an irregular shape are not labeled. (b) Two terminal connections are not labeled.

light of the overhead perspective characteristic of remote sensing images, it is acknowledged that HBB is inherently limited in the ability to precisely delineate the actual positions of objects with arbitrary directions, as it contains a significant amount of irrelevant information from the background. Therefore, we use RoLabelImg⁵ to manually generate the fine OBB for bridges. Specifically, the labeled rectangular bounding box can be defined by four corner points $(x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4)$ in the clockwise order. In the initial phase of pre-annotation, we form a specialized team comprising 10 members, each possessing extensive expertise in the field of remote sensing interpretation. This team undergoes comprehensive training in fundamental annotation techniques and subsequently conducts annotation tests on a representative subset of the dataset. In the following feedback and refinement stage, experts thoroughly review and evaluate the team's initial annotations, resulting in the formulation of refined annotation criteria. Subsequently, guided by this adjustment, the team embark on the formal large-scale annotation process, accompanied by experts' random sampling inspections.

C. Dataset Analysis

In contrast to the other existing bridge detection datasets, GLH-Bridge exhibits notable advantages in terms of GSD, image size, instance quantity, and instance diversity. The GLH-Bridge dataset showcases six prominent merits.

- *Various Instance Scales*: GLH-Bridge incorporates a diverse range of bridge sizes, ranging from tiny bridges with 12 pixels to giant bridges exceeding 3000 pixels. As depicted in Fig. 4(a), large bridges show a high presence in GLH-Bridge, surpassing the quantity reported in existing datasets. This highlights the imperative of utilizing raw large-size images to preserve the integrity of bridges. Furthermore, as illustrated in Fig. 4(b) and (c), a substantial number of small bridges are showcased in GLH-Bridge. Consequently, detecting huge bridges entails processing raw large-size images, presenting a challenge in the context of conventional practice that employs small-size images for the detection of petite bridge instances.

⁵[Online]. Available: <https://github.com/cgvict/roLabelImg>

- *Extreme Aspect Ratios*: GLH-Bridge contains many giant bridges with extreme aspect ratios, as depicted in Fig. 4(a). The identification of these instances poses a formidable challenge for oriented object detection algorithms.

- *Large Image Sizes*: In the context of the GLH-Bridge, over 1,000 large-size VHR images have sizes greater than $8,000 \times 8,000$ pixels. Due to the diverse sizes of these images, conventional downsampling techniques using fixed ratios are ill-suited. The effective processing of these large-size images, while simultaneously preserving the integrity of exceptionally large bridges, presents a significant challenge for existing object detection methods.

- *Diverse Background Types*: As shown in Fig. 5, GLH-Bridge includes bridges across diverse terrains, encompassing not only *water body* but also *dry riverbeds, vegetation, valleys, deserts, urban roads, etc.* This requires object detection algorithms to possess the capability to recognize bridges across a spectrum of backgrounds. Additionally, the challenge is further exacerbated by the potential for bridges to intersect or overlap with other objects, such as roads.

- *Global Coverage*: GLH-Bridge spans the globe and includes samples from all continents. This vast and diverse region provides a wide range of bridge types and landscapes, promoting the dataset's generalizability to various scenarios.

- *Variation in Instance Density*: The distribution of bridges per image in GLH-Bridge is illustrated in Fig. 4(d). In densely populated urban areas or regions abundant in waterways and transportation, bridges are frequently densely distributed. However, rural areas or less developed regions exhibit a smaller number of bridges, with background areas occupying a significant proportion.

IV. THE PROPOSED METHOD

To holistically detect bridges from large-size VHR images, this paper presents HBD-Net, which stands as the pioneering approach expressly tailored for this objective. This section is dedicated to providing a detailed explanation of the HBD-Net.

A. Model Preview

Contemporary deep networks encounter limitations when directly processing large-size RSIs due to the constrained memory capacity of the GPU. To address this problem, we present one factorized representation (i.e., the DIP) of the original large-size image. What's more, we leverage the proposed SDFF with separate detectors to train or infer upon the DIP, and an inter-layer feature fusion (IFF) module is proposed to facilitate feature complementation between layers within the SDFF. Moreover, we enhance the performance of HBD-Net by incorporating the SSRW strategy during sample allocation, and via cross-scale-transfer distillation in SDFF. Our method is illustrated in Fig. 6.

B. HBD-Net Architecture

To effectively process the large-size image, we propose the SDFF architecture, which utilizes separate detectors to tackle the DIP and conducts feature fusion via the IFF module. We

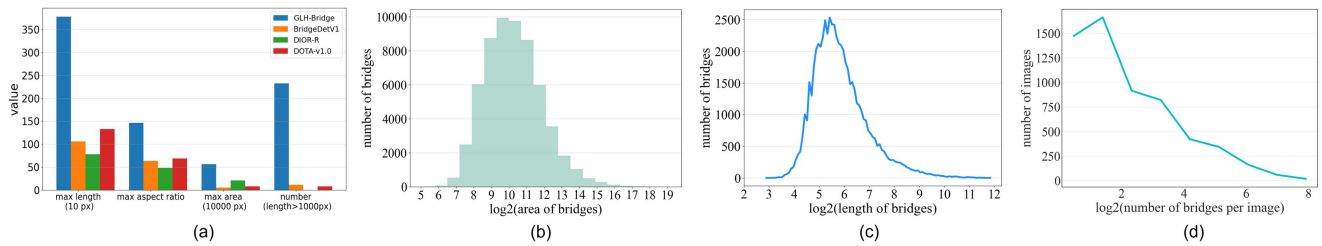


Fig. 4. Illustration of GLH-bridge's characteristics. (a) Comparison of bridges' characteristics across different datasets. (b) Distribution of bridges' areas in GLH-Bridge. (c) Distribution of bridges' length in GLH-Bridge. (d) Distribution of bridges' density in GLH-Bridge.



Fig. 5. Illustration of bridges across different backgrounds in the proposed GLH-Bridge dataset. (a) Bridges across vegetation. (b) Bridges across dry riverbeds. (c) Bridges across roads. (d) Bridges across water bodies.

will provide a detailed explanation of these components in the following sections.

1) *Separate Detectors on Dynamic Image Pyramid. DIP Construction:* When presented with a large-size VHR image with a size of $H \times W$, we progressively downsample the original large-size VHR images at a fixed ratio of σ to construct the image-level pyramid with a variable number of layers. The termination condition of the top layer (the n -th layer) of the pyramid is defined as follows:

$$\frac{H}{\sigma^{n-1}} \leq H_t \quad \text{or} \quad \frac{W}{\sigma^{n-1}} \leq W_t, \quad (1)$$

where (H_t, W_t) is the termination threshold. So we can get the DIP with n layers and the size of its top layer image is (H_n, W_n) , where $H_n = H/\sigma^{n-1}$, $W_n = W/\sigma^{n-1}$. At each layer of the DIP, we employ a fixed-sized window (the size is equal to (H_t, W_t)) to gradually extract the image patches and send them into the detector corresponding to the layer.

Separate Detectors: It is noted that retaining extremely small labels in the downsampled layers can lead to severe information loss. Additionally, the identification of tiny objects in layers with higher resolution tends to be more accurate. Against this backdrop, a set of thresholds is introduced to allocate the OBB labels to each layer of the DIP based on the OBB's length. As a result, each detector embedded within the SDFP is responsible for predicting bridges with specific scales. To enable the SDFP to possess scale sensitivity when detecting multi-scale bridges, we utilize separate object detectors at layers of the SDFP instead of a unified detector (the reason is explained in Section V-B1). Overall, one large-size VHR image is decomposed into one DIP with multiple layers, which passes through the SDFP followed by separate detectors. Hence, this factorized framework facilitates the training of HBD-Net even under constraints imposed by limited computational resources (e.g., one single GPU).

2) *Inter-Layer Feature Fusion:* Considering the variation of field-of-views in the same window at different layers of DIP, the higher layers have global information, while the lower layers contain detailed information. To effectively utilize complementary cues to feature fusion, we devise an Inter-Layer Feature Fusion (IFF) module to enable bidirectional feature sharing within the SDFP.

Similar to the basic feature extractors (e.g., Resnet [49] followed by FPN [50]), this paper recommends extracting feature pyramids from the DIP. Given the feature sets obtained by the feature pyramid network (FPN) from all image layers within the DIP, candidate feature sets are selected from the adjacent image layers. Subsequently, we perform inter-layer feature fusion on these candidate feature sets via feature alignment and fusion.

Feature Selection: We begin by identifying candidate feature sets for fusion. As shown in Fig. 7, in the case of FPN, the spatial sizes of adjacent levels in the feature pyramid always differ by $2\times$. We set P_i^j as the i -th level feature in the feature pyramid of the j -th image layer. Assuming that the j -th image layer is located in the middle of the DIP, the feature pyramids of the two neighboring layers can be represented as P^{j-1} and P^{j+1} , respectively. Given that the downsampling ratio of FPN equals the downsampling ratio σ we set for DIP, we can draw the conclusion that the following features P_{i+1}^{j-1} , P_i^j , and P_{i-1}^{j+1} have the same actual downsampling ratio. We define this candidate feature set as $P_{cand} = \{P_{i+1}^{j-1}, P_i^j, P_{i-1}^{j+1}\}$.

Feature Alignment and Fusion: After obtaining the candidate feature sets P_{cand} , we align the features within P_{cand} based on consistent spatial position and conduct fusion. The rough

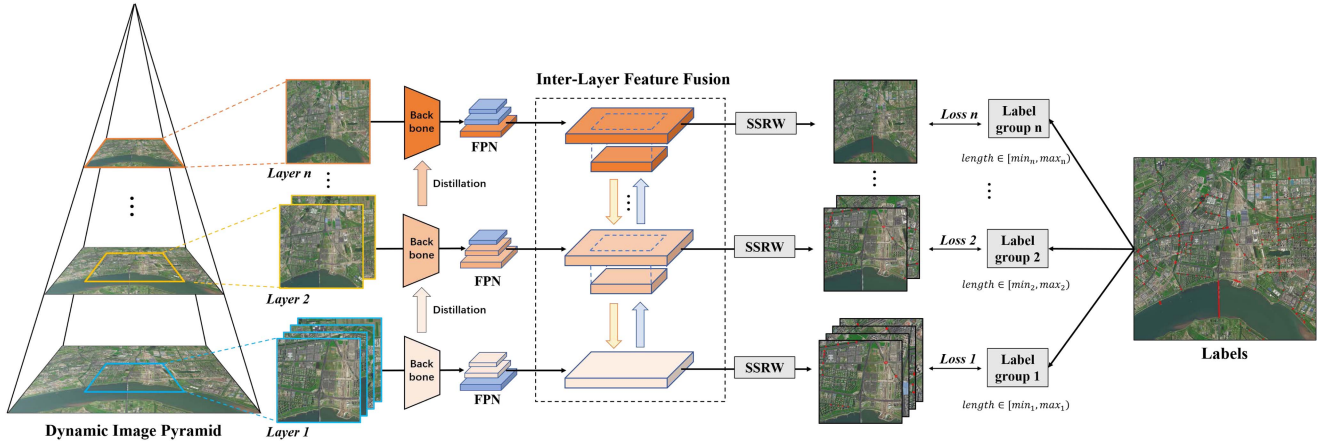


Fig. 6. The pipeline of the proposed HBD-Net. It contains the proposed SDFP architecture and SSRW strategy. The SDFP architecture consists of separate detectors and the IFF module. From the input large-size VHR image, we construct a DIP and send it to the separate detectors of the SDFP to obtain features. Then features from all detectors of the SDFP are fused via the IFF module to share both contextual and detailed texture information. The SSRW strategy is applied in the sample selection stage of object detectors to balance the regression weight. Finally, the output fusion features are fed into the object detectors' heads to obtain the results of each layer, which are used to compute the loss with corresponding ground-truth labels.

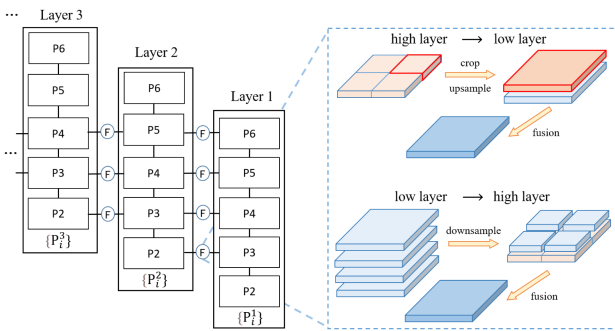


Fig. 7. Illustration of the proposed IFF module. The figure illustrates the ways of feature fusion between two adjacent layers.

representation of this process is shown in Fig. 7. For P_{i+1}^{j-1} , we begin by conducting downsampling it and then align it with the spatial consistent region on P_i^j . For P_{i-1}^{j+1} , we align it with the spatial consistent region on P_i^j by cropping. Due to the existence of sliding windows, the image features of the $(j-1)$ -th layer are extracted in batches during the training process. Consequently, these features are concatenated along the spatial dimension to fit the size of P_i^j after downsampling. Following this alignment process, we perform feature fusion as follows:

$$P_i^j = \text{act}(\text{conv}(\text{concat}(\text{align}(\{P_{i+1}^{j-1}, P_i^j, P_{i-1}^{j+1}\})))), \quad (2)$$

where act , conv , concat and align refer to activation layer (e.g., sigmoid), 1×1 convolutional layer, channel-wise concatenation operation and aforementioned alignment process, respectively. During the layer-by-layer training of the DIP, the fusion of features occurs subsequent to the training of detectors across all levels. In instances where the feature set resides within the middle image layer, it undergoes fusion with the original feature sets before fusion with those from lower and higher image layers. This fusion process remains independent of the sequence in which various layers are fused. The original feature

set is preserved without performing feature fusion. In this way, the features in each layer are fused with the features from the adjacent layers, allowing to capture of contextual information from the upper layer and detailed texture information from the lower layer.

C. HBD-Net Optimization

As bridges exhibit drastic variations in spatial scales and aspect ratios, the Intersection over Union (IoU) between the prediction and label exhibits heightened sensitivity to regressive bias, especially for boxes with larger aspect ratios. Existing methods typically rely on fixed strategies, such as employing the maximum IoU [19], [21] or distance metrics in feature maps [51], to select positive samples and assign them uniform weights. This practice is unsuitable, as it fails to account for the disparities in regression weights required for samples with distinct aspect ratios. To address this problem, we propose a shape-sensitive sample re-weighting (SSRW) strategy during the sample assignment stage. It aims to encourage the deep network to prioritize samples with extreme aspect ratios, and further balance the weighted regression losses.

As illustrated in Fig. 8, following the assignment and selection of positive and negative samples (i.e., sample points), each ground-truth box is linked to its positive samples for subsequent regression and classification predictions. For the positive samples corresponding to a ground-truth box, where w and h represent the width and height of the ground-truth box, respectively, and r denotes the normalized aspect ratio of ground-truth boxes within the mini-batch. The distance between the center point of this box and one of its corresponding positive samples is denoted as Δd . From this, the projected lengths w' and h' of Δd in the w and h directions can be computed. The relative offset factors r_w and r_h are then defined as $r_w = \frac{2w'}{w}$, $r_h = \frac{2h'}{h}$. After acquiring the relative offset factors, we use offset measurement factors Q^w and Q^h to evaluate the deviation of the selected samples. These

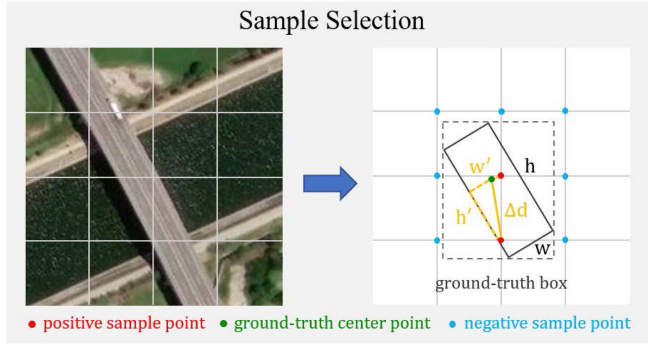


Fig. 8. Illustration of the proposed SSRW strategy. The red and blue points represent positive and negative samples selected by the object detector, respectively. For anchor-based detectors, these points correspond to the feature map locations generating anchors or proposals. For anchor-free detectors, these points indicate the grids on the feature maps. To maintain clarity and simplicity, the depiction of anchors or proposals associated with the sample points (applicable to anchor-based methods) is not depicted in this illustration.

factors can be expressed as:

$$Q^w = \ln(r_w + 1) + 1, \quad (3)$$

$$Q^h = \ln(r_h + 1) + 1. \quad (4)$$

After obtaining the offset measurement factors Q^w and Q^h , the SSRW strategy incorporates them into the regression loss weight w^{reg} to assign higher weights to more challenging samples (i.e., those with larger aspect ratios). The w^{reg} is defined as follows:

$$w^{reg} = \mu Q^w Q^h r, \quad (5)$$

where μ is the adjustment factor and set to 1.0. In this case, an increased value of Q^w and Q^h indicates a larger relative distance between the positive sample's prediction box and the ground-truth box. This suggests that the transformation of the candidate box into a high-quality regression box is more challenging. Consequently, assigning a higher w^{reg} to such positive samples enables the detector to prioritize them. Moreover, given the prevalence of small objects, it contributes to achieving an equitable balance in regression weights among bridges with varying aspect ratios, when the weight w^{reg} shifts towards objects with larger aspect ratios.

The total loss of oriented object detection and horizontal object detection is defined as follows:

$$\mathcal{L}_O = \sum_{m=1}^n \lambda^m \left(\frac{1}{N} \sum_{i \in \psi} \mathcal{L}_i^{cls} + \frac{1}{N^+} \sum_{j \in \psi_p} w_j^{reg} \mathcal{L}_j^{reg} \right), \quad (6)$$

$$\mathcal{L}_H = \sum_{m=1}^n \lambda^m \left(\frac{1}{N} \sum_{i \in \psi} \mathcal{L}_i^{cls} + \frac{1}{N^+} \sum_{j \in \psi_p} \mathcal{L}_j^{reg} \right), \quad (7)$$

where w_j^{reg} is the regression weight calculated by the proposed SSRW strategy. n is the number of layers in the DIP, and λ^m is the balanced weight corresponding to the loss of the m -th layer, which is set to 1. ψ and ψ_p represent the set of all samples and the set of positive samples, respectively. N and N^+ denote the

total number of all samples and positive samples, respectively. The classification loss \mathcal{L}_i^{cls} is focal loss [52] and the regression loss \mathcal{L}_j^{reg} is Smooth L1 loss as defined in [53].

To make full use of the supervision of multi-scale bridges and pursue the scale-sensitive detector, we train the separate detectors within the SDFD layer-by-layer. This process commences with training the bottom layer and proceeds to train each subsequent layer, culminating with the top layer. Upon the completion of training for the detector of the m -th layer ($m \in [1, n - 1]$), we use its weights to initialize the detector of the $(m + 1)$ -th layer. Meanwhile, congenetic labels from the label assign strategy are used to constrain the scale-equivalence of outputs from separate detectors, achieving cross-scale-transfer distillation and enhancing the performance of the deep network.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Benchmark

1) *Evaluation Metrics*: We establish a benchmark on the GLH-Bridge dataset for two types of object detection tasks: **OBB** detection and **HBB** detection. The Average Precision (AP) is adopted as the main evaluation metric in this study (by the IoU computation for the True Positive (TP), False Positive (FP), and False Negative (FN)). We adopt the PASCAL VOC 07 metric [62] to calculate the mean Average Precision (mAP). In the MS-COCO dataset [63], the pixel area of the ground-truth boxes is used to determine *small*, *medium*, and *large* scales to calculate the corresponding AP values, which has been widely used to assess various detection algorithms. However, it is important to note that bridges, despite varying significantly in lengths and aspect ratios, may appear to possess the same area in VHR images. Dividing bridges solely based on area may prove inadequate to accurately reflect the detection difficulty and overlook the influence of image size constraints on the detection algorithm.

In light of the aforementioned limitation, we propose new evaluation metrics based on the length of the longer side of the ground-truth boxes. Specifically, we define a set of pixel intervals as $\{(0, 50], (50, 200], (200, 800], (800, 16384]\}$ to categorize bridges based on their lengths, classifying them as *short*, *middle*, *large*, and *huge*. The corresponding APs are denoted as AP_{sh} , AP_{md} , AP_{lg} , and AP_{hg} , respectively. It is important to note that the detection of huge bridges can often be a challenging task, as they may not be effectively and completely captured within a single sliding window when employing traditional cropping strategies.

2) *Implementation Details*: The algorithms employed in our experiments are from two open-source pytorch-based algorithm libraries, MMRotate [64] and MMDetection [65]. These libraries integrate various state-of-the-art object detection algorithms, along with their corresponding backbone networks, feature extractors, and detectors. They enable the reproduction of the original accuracies of the respective algorithms within a unified algorithm framework, ensuring fairness. Hence, these two algorithm libraries were chosen for the benchmarks for our experiments.

Experiments are performed on a server with 1 Tesla V100 GPU and 16 GB memory. The backbone networks are initialized with models pre-trained on ImageNet [66]. We adopt the “2×” training schedule in MMRotate and MMDetection. The SGD optimizer is employed with a learning rate of 0.005, momentum of 0.9, and weight decay of 0.0001. When performing feature fusion among the detectors of the proposed SDF, the learning rate is set to 0.001. A linear warm-up strategy is applied for the initial 500 iterations, with a rate of 1.0/3. As for the algorithms used to establish benchmark results, the batch size is set to 4.

In the case of the HBD-Net utilized in this study, the batch size is set to 1 during training, and the learning rate is adjusted accordingly. The image processing strategies for training and testing follow the description in Section IV, the downsampling ratio σ is set to 2.0. When training the HBD-Net, we use a label filtering strategy to divide original labels into n groups to calculate loss with the outputs of n layers. To the n -th label group, the filtering threshold is represented as $[\min_n, \max_n)$, and the \min_n is set to $15 \times 2^{(n-1)}$ pixels and the \max_n is set to 1448 pixels (i.e., $1024 \times \sqrt{2}$ pixels) considering the size of the cropping window. In all experiments, random flipping was used as the only data augmentation technique.

3) *Mainstream Methods*: To assess the efficacy of the HBD-Net, we conduct a comparative evaluation against 18 advanced object detection methods. For the OBB task, we choose two-stage approaches such as Faster R-CNN-O [7], RoI Transformer [19], Oriented R-CNN [21], and ReDet [20]; one-stage approaches including FCOS-O [51], R³ Det [54], KLD [55], and Oriented RepPoints [56]; and methods for object detection in large-size images like CGL [25]. Oriented R-CNN is chosen as the baseline method for the proposed HBD-Net and CGL. For the HBB task, we choose RetinaNet [52], Faster R-CNN [57], FCOS [51], TOOD [58], Cascade R-CNN [59], ATSS [60], GuidingAnchor [61]; and methods designed for large-size images like CGL [25], GLSAN [23], and SAHI [22]. Faster R-CNN is chosen as the baseline method for CGL, GLSAN, SAHI, and HBD-Net on the HBB task. It should be noted that the SSRW strategy is not used when training the proposed HBD-Net on the HBB task.

In the case of CGL, GLSAN, and SAHI, we adopt their default strategies to process large-size images. For SAHI, we set the patch size to 1024×1024 pixels, with a 200-pixel overlap if necessary, and combine three strategies: slicing-aided hyper inference, full image inference (FI), and an overlapping patches-based cropping strategy (PO). In the case of GLSAN, we adhere to the original strategy by configuring the subregion number to 4. We incorporate the subregion image cropping component of its training data augmentation (TDA) and implement the SelfAdaptiveCrop technique during testing, employing a crop size of 1024×1024 pixels. For the other object detection methods, the original images are processed using an overlapping patches-based cropping strategy for training and testing. The cropping settings for training and testing are consistent, with a cropping window size of 1024×1024 pixels and a 200-pixel overlap.

4) *Results and Analysis*: The benchmark and experimental results for OBB and HBB tasks on GLH-Bridge are presented in Table II.

For the OBB task, the experimental results demonstrate that the HBD-Net achieves the best performance on the benchmark of GLH-Bridge, with an mAP score of 35.35%. It achieves an accuracy of 28.69% on the AP₇₅ metric, underscoring the efficacy of our approach in accurately detecting rotated bridges. Furthermore, our method achieves the best performance, 33.47% and 20.61% in the AP_{lg} and AP_{hg} metrics, respectively. This highlights the HBD-Net’s effectiveness in handling the detection of **large bridges** that may exceed the typical cropping size, particularly for instances with a length exceeding 800 pixels. Additionally, our method also shows benefits in detecting small objects, which constitute a significant portion of the dataset.

For the HBB task, we consider that the aspect ratio of the horizontal box is determined by both the orientation of bridges and their true aspect ratios. Therefore, it does not accurately reflect whether the bridges are elongated in shape. As a result, we do not incorporate the SSRW strategy for the HBB task, only utilizing the proposed SDF architecture as the employed approach. Under this setting, the HBD-Net also achieves a remarkable performance of 34.49% mAP. Furthermore, in comparison to general object detection methods, the HBD-Net showcases outstanding performance in detecting large bridges. It obtains 35.21% and 35.59% in the AP_{lg} and AP_{hg} metrics, respectively.

Additionally, for the methods designed for object detection in large-size images, although SAHI achieves fine small object detection by resizing overlapping patches, its upsampling technique provides limited benefits for VHR RSIs. CGL employs a fixed downsampling strategy, which results in information loss and suboptimal performance in AP_{hg}. GLSAN performs prediction on the downsampled original image and selects sub-blocks for detailed detection through clustering of the predicted results. However, it tends to miss scattered small bridges, and is still hard to comprehensively detect large bridges.

In conclusion, the experimental results from both the OBB and HBB tasks demonstrate the effectiveness of the HBD-Net in a general sense. It is capable of adapting to the characteristics of both horizontal and oriented bounding boxes, and the visual results are shown in Fig. 9. Additionally, our HBD-Net is independent of the specific object detection methods. Therefore, it can seamlessly accommodate a wide range of advanced one-stage or two-stage object detectors within the proposed SDF without encountering specific limitations. This observation highlights the versatility and applicability of the proposed approach in this study.

B. Component Analysis

We conduct ablation experiments on the GLH-Bridge dataset to evaluate the impact of two key components in our proposed HBD-Net (i.e., the SDF architecture and the SSRW strategy).

1) *Effectiveness of the SDF*: As shown in Table III, we explore the effectiveness of the detector utilization strategy and IFF used in the proposed SDF architecture. Our proposed SDF without cross-scale-transfer distillation and IFF demonstrates a significant enhancement in accurately detecting large bridges, with a notable 8.57% improvement in AP_{hg} metric compared to the baseline. When considering whether each layer in the SDF employs an individual detector or if all layers share a detector,

TABLE II
ACCURACY (%) OF OBB AND HBB TASKS ON GLH-BRIDGE

OBB Task	Backbone	mAP	AP ₅₀	AP ₇₅	AP _{sh}	AP _{md}	AP _{lg}	AP _{hg}
Rotated Faster R-CNN [7]	R50-FPN	31.35	67.99	22.73	30.54	35.08	18.52	5.41
RoI-Transformer [19]	R50-FPN	33.66	69.58	25.55	32.28	38.05	26.32	3.10
Rotated FCOS [51]	R50-FPN	29.28	60.14	22.74	26.98	33.78	25.02	2.13
Rotated RetinaNet [52]	R50-FPN	29.55	61.48	22.68	28.29	33.36	14.93	4.93
R ³ Det [54]	R50-FPN	31.11	68.01	22.84	29.95	34.04	23.11	4.70
KLD [55] (R ³ Det)	R50-FPN	31.92	68.47	23.67	30.88	35.15	23.32	5.83
ReDet [20]	ReR50-ReFPN	34.29	69.99	26.06	31.94	38.12	29.49	2.47
Oriented R-CNN [21]	R50-FPN	34.16	69.87	26.29	32.83	37.74	29.30	5.68
Oriented RepPoints [56]	R50-FPN	29.66	60.19	22.73	26.65	34.27	19.09	7.44
CGL [25]	R50-FPN	34.72	70.55	27.47	33.16	38.14	30.53	12.68
HBD-Net (Ours)	R50-FPN	35.35	71.69	28.69	33.38	38.93	33.47	20.61
HBB Task	Backbone	mAP	AP ₅₀	AP ₇₅	AP _{sh}	AP _{md}	AP _{lg}	AP _{hg}
Faster R-CNN [57]	R50-FPN	33.40	70.72	30.73	31.63	40.14	30.49	8.19
RetinaNet [52]	R50-FPN	30.71	67.30	27.32	28.96	37.02	24.59	3.39
FCOS [51]	R50-FPN	22.32	51.33	18.01	19.42	27.99	15.05	3.41
TOOD [58]	R50-FPN	30.43	65.01	28.52	28.04	37.05	26.41	6.41
Cascade R-CNN [59]	R50-FPN	33.71	70.84	32.10	31.84	39.50	32.09	8.01
ATSS [60]	R50-FPN	27.92	63.52	23.00	27.16	33.44	19.44	5.91
GuidingAnchor [61]	R50-FPN	33.81	71.22	31.71	31.72	39.89	31.54	7.11
GLSAN† [23]	R50-FPN	26.95	54.51	18.13	22.65	31.24	21.24	13.29
SAHI [22]	R50-FPN	34.00	71.12	30.94	32.73	41.11	31.51	18.68
CGL [25]	R50-FPN	33.93	71.25	30.47	31.91	40.28	32.62	16.12
HBD-Net* (Ours)	R50-FPN	34.49	72.45	32.68	32.29	41.54	35.21	35.59

* indicates training the HBD-Net without the proposed SSRW strategy. † indicates training the GLSAN without the Local Super-Resolution Network (LSRN).

TABLE III
ACCURACY (%) OF ABLATION STUDIES ON THE IMPACT OF DIFFERENT STRATEGIES USED IN THE PROPOSED SDFP ARCHITECTURE ON THE OBB TASK OF GLH-BRIDGE

Structure	Setting	Distillation	IFF	mAP	AP ₅₀	AP ₇₅	AP _{sh}	AP _{md}	AP _{lg}	AP _{hg}
DIP	Unified detector	×	×	34.61	69.95	26.71	32.97	38.15	30.79	12.36
	Separate detector	×	×	34.65	69.93	26.83	33.02	38.06	30.47	14.25
	Separate detector	✓	×	34.87	70.18	27.34	33.11	38.23	31.22	15.76
	Separate detector	✓	✓	35.08	70.58	28.06	33.27	38.51	32.11	17.30

“DIP” denotes that using the proposed dynamic image pyramid. “Unified detector” denotes that all layers in the SDFP use a unified object detector. “Separate detector” denotes that each layer in the SDFP uses a separate detector. “Distillation” denotes using cross-scale-transfer distillation in the training phase of the SDFP.

the former slightly outperforms the latter. When we incorporate a cross-scale-transfer distillation strategy into the process of training the SDFP, the accuracy can be improved, resulting in an additional improvement of 3.4% improvements in AP_{hg} metric. Furthermore, through the integration of the IFF module, the higher layer can benefit from the finer details provided by the lower layer, resulting in enhanced final performance in terms of the AP₇₅ and AP_{hg} metrics, which reach 28.06% and 17.30%, respectively.

2) *Effectiveness of the SSRW Strategy*: As our proposed HBD-Net utilizes respective detectors in the proposed SDFP, as shown in Table IV, we examine the effectiveness of the SSRW strategy when applied to these detectors individually. It can be observed that the proposed SSRW strategy enhances the regression accuracy of the detector when it is solely applied

to the detector of the bottom layer, it results in 1.37% and 3.30% improvements in AP_{lg} and AP_{hg} metrics, respectively, compared to the baseline. Furthermore, with the incorporation of the SDFP architecture, we extend the application of the SSRW strategy to detectors corresponding to the higher layers of the pyramid, leading to a further improvement of 2.29% in AP_{hg} metric. Given the typically larger aspect ratios of large bridges, the above experiments demonstrate the effectiveness of our proposed SSRW strategy in directing the network’s focus toward bridges with larger aspect ratios, thereby improving detection accuracy. Finally, in addition to the overall improvement in all metrics, the HBD-Net achieves a significant improvement of 2.40%, 4.17% and 14.93% in AP₇₅, AP_{lg} and AP_{hg} metrics, respectively, compared to the baseline. This study affirms the effectiveness of the proposed method in enhancing bridge

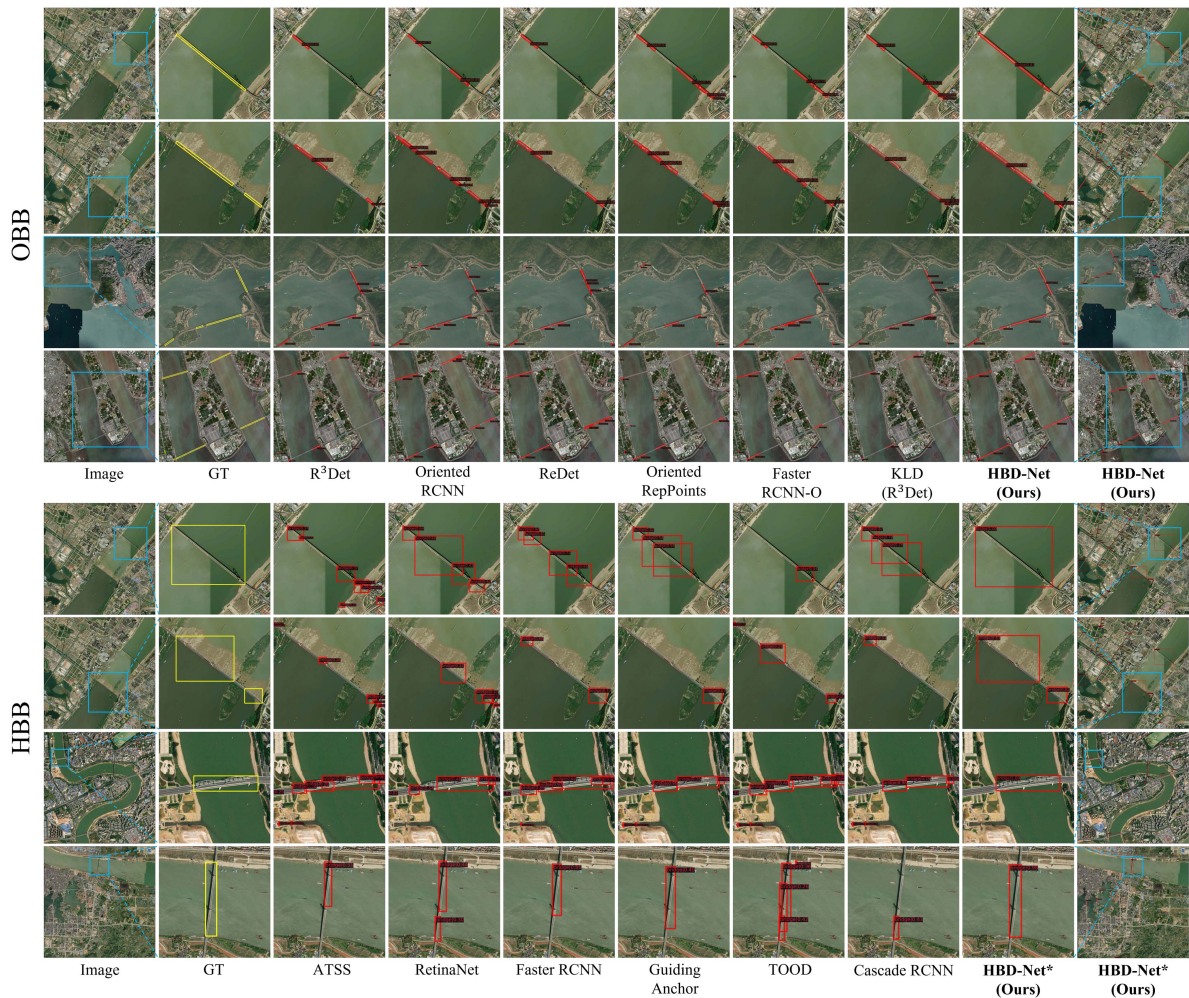


Fig. 9. The visualization results of OBB and HBB tasks on the GLH-Bridge dataset using the HBD-Net and comparison object detection methods. * indicates using the HBD-Net without the proposed SSRW strategy.

TABLE IV
ACCURACY (%) OF ABLATION STUDIES ON THE IMPACT OF THE SDFP ARCHITECTURE AND THE SSRW STRATEGY ON THE OBB TASK OF GLH-BRIDGE

SSRW ₁	SDFP	SSRW _g	mAP	AP ₅₀	AP ₇₅	AP _{sh}	AP _{md}	AP _{lg}	AP _{hg}
×	×	×	34.16	69.87	26.29	32.83	37.74	29.30	5.68
✓	×	×	35.20	70.94	27.85	33.34	38.11	30.67	8.98
×	✓	×	35.08	70.58	28.06	33.27	38.51	32.11	17.30
✓	✓	×	35.12	71.28	28.37	33.44	38.82	32.76	18.32
✓	✓	✓	35.35	71.69	28.69	33.38	38.93	33.47	20.61

SSRW₁ denotes using the SSRW strategy only on the baseline detector or the detector of the bottom layer of SDFP. SSRW_g denotes using the SSRW strategy only on the detectors of the layers except for the bottom layer of SDFP.

detection performance in large-size images, especially concerning the detection of large bridges in their entirety. It is important to note that, with the implementation of the SSRW strategy for higher-layer detectors, a decrease in the AP_{sh} metric is observed. This decrease is attributed to a decrease in the proportion of small-scale labels at the higher layers resulting from the label filtering. As a result, SSRW's role in maintaining the balance of loss between small and large objects is diminished, aligning with its intended design principles.

Moreover, to comprehensively evaluate the effectiveness of our method, we further conducted ablation experiments on the DOTA-v1.0 dataset [7]. These experiments demonstrate how our designed modules progressively enhance the performance step by step. As shown in Table V, our proposed SSRW strategy and SDFP architecture result in 0.96% and 1.26% improvement respectively in AP_{BR} metric. Our HBD-Net achieves 46.11% mAP and 56.02% AP_{BR} based on the baseline. These results highlight the capability of our proposed HBD-Net to enhance



Fig. 10. The visualized prediction results of the DIOR-R dataset by the model trained on the GLH-Bridge dataset.

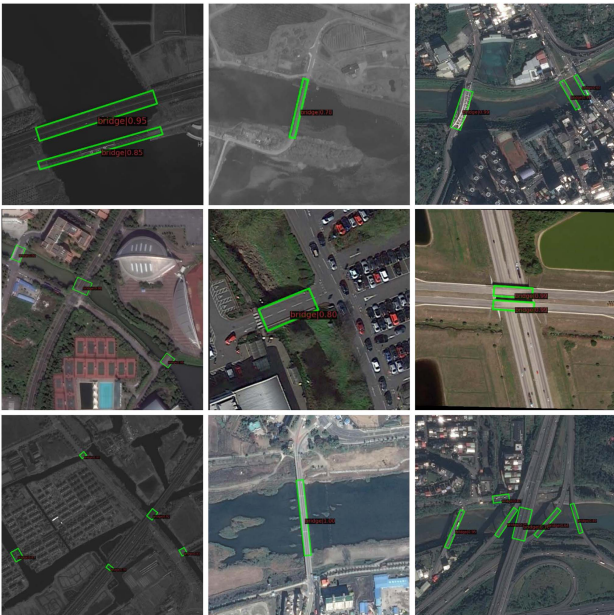


Fig. 11. The visualized prediction results of the DOTA-v1.0 dataset by the model trained on the GLH-Bridge dataset.

TABLE V
ACCURACY (%) OF ABLATION STUDIES ON THE OBB TASK ON DOTA-V1.0

SSRW	SDFP	mAP	AP ₅₀	AP ₇₅	AP _{BR}
×	×	44.92	75.80	45.45	54.52
✓	×	45.63	76.25	45.98	55.48
×	✓	45.92	76.53	46.32	55.78
✓	✓	46.11	76.95	46.76	56.02

SSRW denotes using the SSRW strategy only on the detector of the bottom layer of SDFP.

TABLE VI
MEMORY CONSUMPTION AND INFERENCE SPEED OF DIFFERENT METHODS

Method	AP ₅₀	AP _{hg}	parameter	FPS
Cropping Strategy	69.87	5.68	41.13M	0.47
TTA	70.94	12.21	41.13M	0.17
HBD-Net	71.69	18.32	56.34M	0.34
HBD-Net†	71.22	18.17	56.34M	0.66

‘TTA’ indicates multi-scale test-time augmentation, ‘parameter’ indicates the model parameters, and ‘FPS’ stands for Frames Per Second. HBD-Net† indicates employing the region selection strategy in the inference process.

TABLE VII
ACCURACY (%) ON CROSS-DATASET GENERALIZATION EXPERIMENTS

Train on \ Test on	GLH-Bridge	DOTA-v1.0	DIOR-R
	GLH-Bridge	34.16	15.78
DOTA-v1.0	49.55	45.76	18.46
DIOR-R	20.74	12.14	19.88

the performance of existing state-of-the-art object detection methods.

C. Complexity Comparison Experiments

For a more comprehensive comparison between our methodology and existing approaches for large-size images, we present the comparative results in terms of model parameters, inference speed, and accuracy across different methods. In our comparison, we opt for mainstream cropping strategies and employ multi-scale testing time augmentation (TTA), as they are specifically designed to handle large-size images or objects with significant scale variations. Additionally, we introduce a simple acceleration strategy based on our DIP structure. Specifically, in the second layer (double downsample) of the DIP, we use a predefined confidence threshold during the inference process for filtering. If the current tile lacks any objects of interest (the confidence score falls below the threshold), we skip its four corresponding tiles at the original resolution, thereby expediting the process. We refer to this acceleration strategy as the **region selection strategy**. Detailed experimental settings are as follows.

In our experimental setup, we utilize the Oriented R-CNN as the baseline, the window settings for the cropping strategy remain consistent with those in the benchmark. Additionally, the ratios for multi-scale TTA are configured as (0.5, 1.0, 1.5), following the approach adopted in MMRotate. The confidence threshold used in our region selection strategy is set to 0.3. The results are presented in Table VI. Despite our method featuring slightly more parameters, it achieves superior accuracy, particularly for large-scale bridges in AP_{hg}, as well as higher FPS compared to multi-scale TTA. Notably, our region selection strategy significantly enhances inference speed while

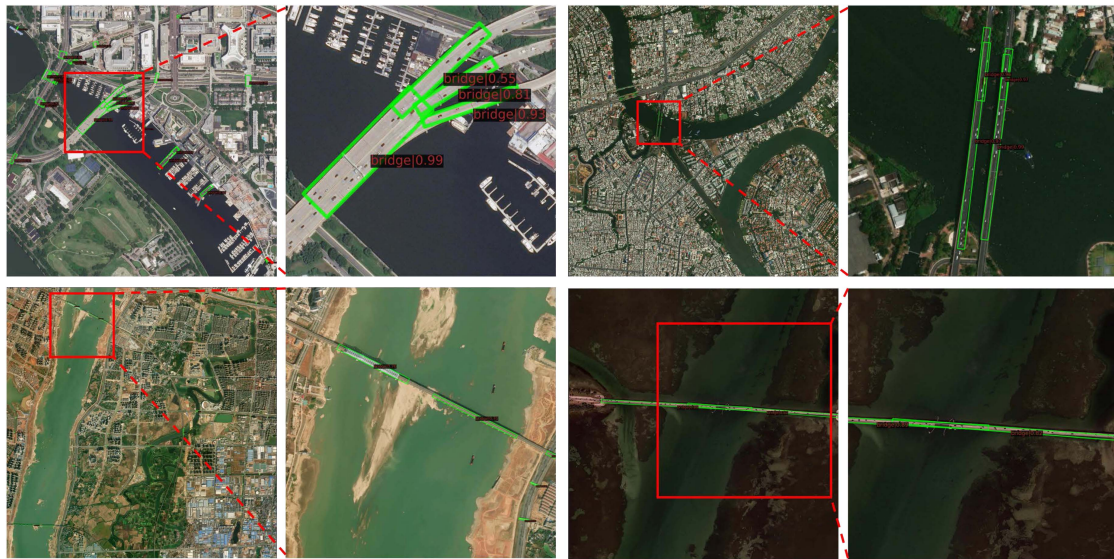


Fig. 12. Visualization of failure cases of HBD-Net. Most of these failure cases arise due to localized predictions triggered by backgrounds similar to that of bridge banks, or because the shape of the bridge is too extreme.

upholding accuracy standards. It results in a noteworthy **28.6%** reduction in inference time on the HBD-Net, speeding up the process by **40.4%** compared to the cropping strategy. All in all, our framework has room for acceleration expansion, and it holds the potential to boost processing efficiency in application scenarios.

D. Cross-Dataset Generalization Experiments

1) *Datasets*: We choose two public datasets (DOTA-v1.0 and DIOR-R) for the cross-dataset generalization experiments. These datasets are chosen based on their large-scale and diverse data characteristics, making them fundamental benchmarks in the field of remote sensing object detection. The details are as follows.

DOTA-v1.0 [7]: DOTA-v1.0 is a large-scale dataset for object detection in aerial images. Its training and validation sets contain a total of 2,541 bridges in 288 images.

DIOR-R [9]: DIOR-R is a large aerial images dataset and has various spatial resolutions, containing 4,000 bridges among 1,576 images with OBB annotation. For the DIOR-R dataset, the provided training, validation, and testing sets are utilized for the cross-dataset generalization experiments.

2) *Experimental Setting*: Cross-dataset generalization analysis is an important evaluation method for assessing the generalization performance of a dataset. We conduct cross-dataset generalization experiments using the bridge subset of the DOTA-v1.0 dataset [7] and the DIOR-R dataset [9]. For the DOTA-v1.0 dataset, we extract the bridge subset from the official training and validation sets for training purposes. The inference is performed on the official unlabeled test set using the standard format. Finally, the test results are uploaded to the official server to obtain accuracy. For the DIOR-R dataset, we select the bridge subset

within the provided training and validation sets for training and evaluate the official test set.

We employ Oriented R-CNN [21] as the algorithm for training and testing. We train models on these three datasets respectively and conduct cross-dataset evaluation. The training settings for DOTA-v1.0 and DIOR-R are kept consistent with the original papers, both with the “1×” training schedule [64]. For our constructed GLH-Bridge dataset, we utilize the training set to train our models while maintaining consistent training settings with the benchmark baseline. To ensure image size compatibility, we implement a cropping strategy with window sizes of 1024×1024 pixels for DOTA-v1.0 and 800×800 pixels for DIOR-R, along with a 200-pixel overlap. The evaluation of cross-dataset generalization experiments is conducted based on the *AP* metric.

3) *Results and Analysis*: The experimental results presented in Table VII demonstrate that GLH-Bridge has achieved outstanding zero-shot generalization performance on two mainstream benchmarks. Specifically, it has achieved a performance improvement of 3.79% on the DOTA-v1.0 dataset and 0.86% on the DIOR-R dataset. These results indicate that the ability of the GLH-Bridge dataset to provide a more comprehensive and accurate representation of bridge characteristics within the domain of the perspective of remote sensing imagery.

The visual results of the DIOR-R dataset generated by the model trained on the GLH-Bridge dataset are shown in Fig. 10. It can be observed that despite the significant variation in image resolution within the DIOR-R dataset (ranging from 0.5 m to 30 m), the model trained on GLH-Bridge exhibits the capability to identify bridges in low-resolution images. Additionally, the DIOR-R dataset contains bridges with diverse color tones and extreme aspect ratios. Despite differences in satellite sources between these images and those in the GLH-Bridge dataset, the model trained on GLH-Bridge demonstrates strong

generalization ability by successfully detecting bridges in backgrounds with high interference and images with lower resolutions.

The visualized prediction results of the DOTA-v1.0 dataset by the model trained on the GLH-Bridge dataset are shown in Fig. 11. It can be observed that despite the inclusion of panchromatic RSIs in addition to RGB images in the DOTA-v1.0 dataset, the proposed model trained on the GLH-Bridge dataset is still able to accurately identify bridges. This demonstrates that the GLH-Bridge dataset can capture the core features of bridges in RSIs, which are invariant to color. Moreover, the trained model achieves good performance in identifying small bridges in the DOTA-v1.0 dataset, which proves that the GLH-Bridge dataset has meticulous and high-quality annotations.

E. Failure Analysis

To identify potential enhancements for HBD-Net, we conduct an analysis of visualizations showcasing failure cases, as illustrated in Fig. 12. These instances of failure primarily stem from two key factors: i) The ground background beneath the bridge exhibits a pronounced contrast with the surrounding bodies of water. This contrast often leads to localized delineation of the bridge's terminus, resulting in the prediction of only a portion of the entire bridge structure. ii) The irregular or excessively complex shape of the bridge contributes to imprecise predictions.

VI. CONCLUSION

In this paper, we propose a large-scale dataset named GLH-Bridge for holistic bridge detection in large-size VHR RSIs. The proposed dataset consists of 6,000 VHR RSIs, with image sizes ranging from $2,048 \times 2,048$ to $16,384 \times 16,384$ pixels, and contains 59,737 bridges spanning diverse backgrounds with OBB and HBB annotation. The large image size, the large sample volume, and the diversity of object scale and background type make GLH-Bridge a valuable dataset, which has the premise to promote one new challenging but meaningful task: holistic bridge detection in large-size VHR RSIs. Furthermore, we present the HBD-Net, a cost-effective solution tailored for holistic bridge detection in large-size images. Based on the proposed GLH-Bridge dataset, we establish a benchmark and provide empirical validation of the effectiveness of the proposed HBD-Net. In future work, we will continue to enrich the GLH-Bridge dataset in terms of its sample volume and sub-category annotation. Furthermore, our objective encompasses the generalization of the proposed HBD-Net to cater to multi-class object detection in large-size images. We endeavor to explore methods that can concurrently enhance the accuracy of both large-scale and small-scale bridges, thus widening the applicability and effectiveness of HBD-Net across various scenarios.

ACKNOWLEDGMENT

We sincerely thank the anonymous editors and reviewers for their insightful comments and suggestions. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

REFERENCES

- [1] G. Sithole and G. Vosselman, "Bridge detection in airborne laser scanner data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 61, no. 1, pp. 33–46, 2006.
- [2] D. Hester and A. González, "A wavelet-based damage detection algorithm based on bridge acceleration response to a vehicle," *Mech. Syst. Signal Process.*, vol. 28, pp. 145–166, 2012.
- [3] D. Cantero and A. González, "Bridge damage detection using weigh-in-motion technology," *J. Bridge Eng.*, vol. 20, no. 5, 2015, Art. no. 04014078.
- [4] H. Guo, R. Zhang, Y. Wang, W. Yang, H.-C. Li, and G.-S. Xia, "Accurate bridge detection in aerial images with an auxiliary waterbody extraction task," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9651–9666, 2021.
- [5] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogrammetry Remote Sens.*, vol. 98, pp. 119–132, 2014.
- [6] X. Sun et al., "FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 184, pp. 116–130, 2022.
- [7] G.-S. Xia et al., "Dota: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
- [8] J. Ding et al., "Object detection in aerial images: A large-scale benchmark and challenges," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7778–7796, Nov. 2022.
- [9] G. Cheng et al., "Anchor-free oriented proposal generator for object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5625411.
- [10] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
- [11] K. Nogueira, C. da Silva, P. Gama, G. Machado, and J. A. Dos Santos, "A tool for bridge detection in major infrastructure works using satellite images.," in *Proc. Workshop Comput. Vis.*, 2019, pp. 72–77.
- [12] X. Yang et al., "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8232–8241.
- [13] 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.
- [14] X. Yang et al., "Detecting rotated objects as Gaussian distributions and its 3-D generalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4335–4354, Apr. 2023.
- [15] G. Nie and H. Huang, "Multi-oriented object detection in aerial images with double horizontal rectangles," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4932–4944, Apr. 2023.
- [16] X. Yang, J. Yan, W. Liao, X. Yang, J. Tang, and T. He, "SCRDet: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2384–2399, Feb. 2023.
- [17] C. Liu, J. Yang, J. Ou, and D. Fan, "Offshore bridge detection in polarimetric SAR images based on water network construction using Markov tree," *Remote Sens.*, vol. 14, no. 16, 2022, Art. no. 3888.
- [18] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [19] 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 Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.
- [20] J. Han, J. Ding, N. Xue, and G.-S. Xia, "Redet: A rotation-equivariant detector for aerial object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2786–2795.
- [21] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented R-CNN for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.
- [22] F. C. Akyon, S. O. Altinuc, and A. Temizel, "Slicing aided hyper inference and fine-tuning for small object detection," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 966–970.
- [23] S. Deng et al., "A global-local self-adaptive network for drone-view object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1556–1569, 2020.
- [24] Z.-Z. Wu, X.-F. Wang, L. Zou, L.-X. Xu, X.-L. Li, and T. Weise, "Hierarchical object detection for very high-resolution satellite images," *Appl. Soft Comput.*, vol. 113, 2021, Art. no. 107885.

- [25] X. Chen et al., "Coupled global-local object detection for large VHR aerial images," *Knowl.-Based Syst.*, vol. 260, 2023, Art. no. 110097.
- [26] H. Pinckaers, B. Van Ginneken, and G. Litjens, "Streaming convolutional neural networks for end-to-end learning with multi-megapixel images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1581–1590, Mar. 2020.
- [27] T. D. Le, H. Imai, Y. Negishi, and K. Kawachiya, "Automatic GPU memory management for large neural models in tensorflow," in *Proc. ACM SIGPLAN Int. Symp. Memory Manage.*, 2019, pp. 1–13.
- [28] N. Shazeer et al., "Mesh-tensorflow: Deep learning for supercomputers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10435–10444.
- [29] Q. Xu and Y. You, "An efficient 2D method for training super-large deep learning models," in *Proc. IEEE Int. Parallel Distrib. Process. Symp.*, 2023, pp. 222–232.
- [30] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 677–694.
- [31] W. Qian, X. Yang, S. Peng, J. Yan, and Y. Guo, "Learning modulated loss for rotated object detection," in *Proc. Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 2458–2466.
- [32] X. Yang, J. Yan, Q. Ming, W. Wang, X. Zhang, and Q. Tian, "Rethinking rotated object detection with Gaussian Wasserstein distance loss," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11830–11841.
- [33] D. Chaudhuri and A. Samal, "An automatic bridge detection technique for multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 9, pp. 2720–2727, Sep. 2008.
- [34] H. Biao, L. Ying, and J. Licheng, "Segmentation and recognition of bridges in high resolution SAR images," in *Proc. Int. Conf. Radar Proc.*, 2001, pp. 479–482.
- [35] S. Fulin, W. Zhongmou, and L. Xiaoli, "An algorithm of bridge detection in remote sensing images based on fractal," in *Proc. Int. Symp. Antennas Propag. EM Theory*, 2003, pp. 600–602.
- [36] Z. Bai, J. Yang, H. Liang, and W. Wang, "An optimal edge detector for bridge target detection in SAR images," in *Proc. Int. Conf. Commun. Circuits Syst.*, 2005, Art. no. 851.
- [37] L. Chen et al., "A new deep learning network for automatic bridge detection from SAR images based on balanced and attention mechanism," *Remote Sens.*, vol. 12, no. 3, 2020, Art. no. 441.
- [38] Z. Wang, Y. Zhang, Y. Yu, L. Zhang, J. Min, and G. Lai, "Prior-information auxiliary module: An injector to a deep learning bridge detection model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6270–6278, 2021.
- [39] W. Chen, Z. Jiang, Z. Wang, K. Cui, and X. Qian, "Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8924–8933.
- [40] L. Shan et al., "Uhrsnet: A semantic segmentation network specifically for ultra-high-resolution images," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 1460–1466.
- [41] S. Guo et al., "ISDNet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 4361–4370.
- [42] Y. Li et al., "MFVNet: A deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation," *Sci. China Inf. Sci.*, vol. 66, no. 4, pp. 1–14, 2023.
- [43] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "FarSeg: Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13715–13729, Nov. 2023.
- [44] Y. Li, B. Dang, W. Li, and Y. Zhang, "GLH-water: A large-scale dataset for global surface water detection in large-size very-high-resolution satellite imagery," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 20, pp. 22213–22221.
- [45] Z. Chen, W. Wang, E. Xie, T. Lu, and P. Luo, "Towards ultra-resolution neural style transfer via thumbnail instance normalization," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 393–400.
- [46] D. Tellez, G. Litjens, J. van der Laak, and F. Ciompi, "Neural image compression for gigapixel histopathology image analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 567–578, Feb. 2021.
- [47] H. Pinckaers, W. Bulten, J. van der Laak, and G. Litjens, "Detection of prostate cancer in whole-slide images through end-to-end training with image-level labels," *IEEE Trans. Med. Imag.*, vol. 40, no. 7, pp. 1817–1826, Jul. 2021.
- [48] G. Cheng et al., "Towards large-scale small object detection: Survey and benchmarks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13467–13488, Nov. 2023.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [50] 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.*, 2017, pp. 2117–2125.
- [51] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1922–1933, Apr. 2022.
- [52] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
- [53] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [54] 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.*, 2021, pp. 3163–3171.
- [55] X. Yang et al., "Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 18381–18394.
- [56] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1829–1838.
- [57] 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. 06, pp. 1137–1149, Jun. 2017.
- [58] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3490–3499.
- [59] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498, May 2021.
- [60] 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 Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9759–9768.
- [61] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2965–2974.
- [62] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.
- [63] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
- [64] Y. Zhou et al., "Mmrotate: A rotated object detection benchmark using pytorch," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 7331–7334.
- [65] K. Chen et al., "Mmdetection: OpenMMLab detection toolbox and benchmark," 2019, *arXiv: 1906.07155*.
- [66] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

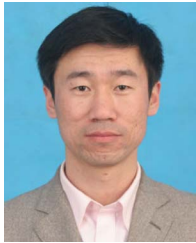


Yansheng Li (Senior Member, IEEE) received the BS degree in information and computing science from Shandong University, Weihai, China, in 2010, and the PhD degree in pattern recognition and intelligent system from the Huazhong University of Science and Technology, Wuhan, China, in 2015. He is currently a full professor with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan. From 2017 to 2018, he was a visiting assistant professor with the Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. He

has authored more than 100 peer-reviewed journal articles and conference papers. His research interests include knowledge graph, deep learning, and their applications in remote sensing Big Data mining. He was awarded the Young Surveying and Mapping Science and Technology Innovation Talent Award of the Chinese Society for Geodesy, Photogrammetry and Cartography in 2022. He received the recognition of the Best Reviewers of the IEEE TGRS in 2021 and the Best Reviewers of the IEEE GRSL in 2022. He is an associate editor of IEEE TGRS, and a Junior Editorial Member of The Innovation.



Junwei Luo received the BS degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2023. He is currently working toward the MS degree with the School of Remote Sensing and Information Engineering, Wuhan University. He has published a first-authored paper in CVPR. His research interests include remote sensing object detection and remote sensing vision-language foundation model.



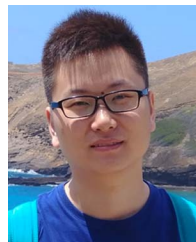
Yongjun Zhang (Member, IEEE) received the BS degree in geodesy, the MS degree in geodesy and surveying engineering, and the PhD degree in geodesy and photography from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively. He is currently the dean with the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 150 research articles and one book. His research interests include aerospace and low-altitude photogrammetry, image matching, combined block adjustment with multisource datasets, artificial intelligence-driven remote sensing image interpretation, and 3-D city reconstruction. He was a Key Member of ISPRS Workgroup II/I from 2016 to 2020. He is the PI Winner of the Second-Class National Science and Technology Progress Award in 2017. In recent years, he has also served as the session chair for above 20 international workshops or conferences. He has been frequently serving as a referee for more than 20 international journals. He is the co-editor-in-chief of *The Photogrammetric Record*.



Yihua Tan (Member, IEEE) received the PhD degree in pattern recognition and intelligent systems from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2004. From 2005 to 2006, he was a postdoctoral staff with the Department of Electronics and Information, HUST. Since 2005, he has been with the School of Artificial Intelligence and Automation, HUST, where he is currently a professor. From 2010 to 2011, he was a visiting scholar with Purdue University, West Lafayette, IN, USA, focusing on remote sensing image analysis. He has authored more than 80 papers in journals and conferences. His research interests include digital image/video processing and analysis, object detection and recognition, and machine learning.



Jin-Gang Yu received the BS degree from Xi'an Jiaotong University, Xi'an, China, in 2005, and the MS and PhD degrees from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007 and 2014, respectively. He was a postdoctoral research associate with the Department of Computer Science and Technology, University of Nebraska-Lincoln, Lincoln, NE, USA, from 2014 to 2016. He spent three years as a Research and Development Engineer with ZTE Corporation, Shenzhen, China, and Nortel Networks Corporation, Guangzhou, China, before starting the Ph.D. Program with HUST. He joined the South China University of Technology, Guangzhou, in 2016, where he is currently an associate professor. His research interests include computer vision, pattern recognition, and machine learning.



Song Bai is a computer vision lead in ByteDance/TikTok Singapore, and also holds an appointment as an adjunct assistant professor with the Department of Electrical and Computer Engineering, National University of Singapore. Before that, he was a research fellow with the University of Oxford working with Prof. Philip H.S. Torr. He serves as an associate editor of *Pattern Recognition* and a guest editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence*. He has served as Area Chair/SPC of CVPR 2023, NeurIPS 2023, and AAAI 2022. His research interests include computer vision and machine learning, and especially video understanding.