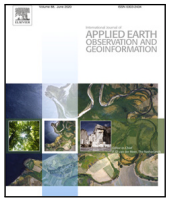
International Journal of Applied Earth Observation and Geoinformation 128 (2024) 103717

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



OBBInst: Remote sensing instance segmentation with oriented bounding box supervision

Xu Cao, Huanxin Zou *, Jun Li, Xinyi Ying, Shitian He

College of Electronic Science and Technology, National University of Defense Technology, Changsha, 410005, China

ARTICLE INFO

Keywords:
Remote sensing
Instance segmentation
Weakly supervised
Oriented bounding-box
Box-supervised

ABSTRACT

Remote sensing (RS) instance segmentation is an important but challenging task due to multi-oriented, densely arranged objects and lack of mask annotation. Compared with redundant horizontal bounding-box (HBB) and expensive pixel-level annotation, oriented bounding box (OBB) annotations can provide compact object depicts with lower annotation costs. Therefore, we propose the first weakly supervised remote sensing instance segmentation method with OBB supervision (namely OBBInst) to reduce the annotation burden and make full use of existing abundant OBB annotations. Based on BoxInst (a high-performance instance segmentation method with box annotations), OBBInst has customized a framework for OBB annotation to unify the incompatibility between existing HBB-based and OBB-based methods. In addition, we propose an oriented projection method with a corresponding loss function to achieve more precise target depicts of OBB annotation. Moreover, we propose an edge similarity loss to incorporate Canny edge prior into deep learning framework for more precise edge identification of densely arranged objects. We have conducted extensive experiments on iSAID and HRSC datasets, and the experimental results demonstrate that OBBInst can achieve the state-of-the-art performance as compared to existing box-supervised methods. In addition, OBBInst dramatically narrows the performance gap between weakly and fully supervised instance segmentation (23.9% vs. 35.6% in iSAID dataset and 79.5% vs. 84.9% in HRSC dataset).

1. Introduction

Object detection and segmentation in remote sensing images (RSIs) has been a significant yet challenging task in remote sensing (RS) interpretation systems, and has various important applications, including environmental monitoring (Ali et al., 2022; Dai et al., 2023), geological disaster detection (Xie et al., 2022; Teng et al., 2021) and land use & development (Bhagavathy and Manjunath, 2006; Shi et al., 2020). Unlike the extensive explorations of horizontal bounding-box (HBB) annotations for natural images in the field of computer vision, HBB is seldom used in RSIs due to their unique characteristics of multi-direction, non-overlap, and dense arrangement. As shown in Fig. 1, HBB annotation results in a large number of redundant information (*e.g.*, as shown in Fig. 1(a), the green area occupies 88.6% of the HBB annotation Liu et al., 2016) and ambiguous semantics (*e.g.*, as shown in Fig. 1(c), the HBB annotation of ship A contains much information of ship B). Compared with HBB annotation, oriented bounding box (OBB) annotation introduces an additional angle parameter to perform a more precise and compact object description, as shown in Fig. 1(e) and (f).

Based on OBB annotation, a large number of methods have emerged (Zhang et al., 2021a; Chen et al., 2021a; Liu et al., 2021; Zhang et al., 2018; Xu et al., 2020) to pursue the box-based detection performance peak. However, per-pixel mask can further provide finer pixel-level location (Gong et al., 2021; Zhang et al., 2021b) to facilitate various downstream tasks, including terrain classification (Julius Fusic et al., 2022; Hu et al., 2023), change detection (Su et al., 2022; Venugopal, 2020), and urban planning (Guo et al., 2018; Mao et al., 2022). Therefore, numerous RS semantic segmentation methods (Diakogiannis et al., 2020; Li et al., 2021b,a) have been proposed to perform per-pixel mask inference. However, the inherent instance-agnostic problem of semantic segmentation results in inferior performance on densely arranged objects (Liang et al., 2022; He et al., 2022). As shown in Fig. 2, semantic segmentation methods can only segment a cluster of objects with a single mask and corresponding class prediction, which provides confusing signals for subsequent direction determination (Li et al., 2021c; Yue et al., 2022) and posture depiction (Chanlongrat et al., 2022).

Recently, several works (Gong et al., 2021; Zhang et al., 2021b; Chen et al., 2021b; Jian et al., 2019) have been proposed to perform instance segmentation in RSIs. Jian et al. (2019) combined the

* Corresponding author.
E-mail addresses: cx2020@nudt.edu.cn (X. Cao), zouhuanxin@nudt.edu.cn (H. Zou), junli@nudt.edu.cn (J. Li), yingxinyi18@nudt.edu.cn (X. Ying), heshitian19@nudt.edu.cn (S. He).

https://doi.org/10.1016/j.jag.2024.103717
Received 28 November 2023; Received in revised form 14 January 2024; Accepted 13 February 2024
Available online 22 February 2024
1569-8432/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

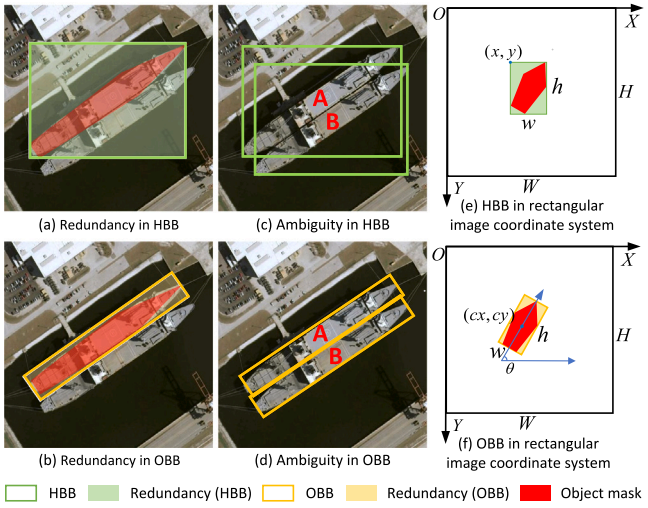


Fig. 1. Differences between HBB and OBB annotation. (a), (b) compare the redundancy (*i.e.*, green and yellow areas) between HBB and OBB. (c), (d) illustrate the ambiguous semantics of HBB. (e), (f) show different representations of HBB and OBB in the rectangular image coordinate system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

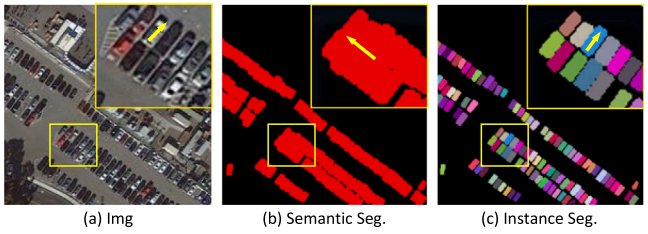


Fig. 2. Differences between semantic segmentation and instance segmentation. (a), (b), (c) show the image and corresponding semantic segmentation and instance segmentation results. We show the zoom-in regions of densely arranged objects for better visualization, and the yellow arrow represents the direction generated by Li et al. (2021c). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

modified U-Net (Ronneberger et al., 2015) with multi-task learning (MTL) to perform instance segmentation of buildings in RSIs. Gong et al. (2021) incorporated the boundary information in Mask R-CNN (He et al., 2017) to achieve more precise instance segmentation. Based on the fully convolutional network (FCN) (Long et al., 2015), Zhang et al. (2021b) utilized semantic attention with extra supervision to strengthen the feature representation capability for performance improvement. However, there are only a few attempts, and its potential remains locked, unlike the extensive explorations (Ronneberger et al., 2015; Long et al., 2015; Badrinarayanan et al., 2017; Zhao et al., 2017; Chen et al., 2017) for natural images. This is mainly due to the potential reasons, including lack of large-scale, accurately annotated datasets and RS objects featured by their usually small size, arbitrary orientation, and locally dense arrangement. Moreover, most existing methods are fully supervised, which usually requires large-scale object mask annotations for training. As shown in Fig. 2(c), this is extremely time-consuming and labor-intensive (Bearman et al., 2016; Everingham et al., 2009; Khoreva et al., 2017).

Therefore, a natural question arises: Can we develop a new framework for RS instance segmentation with box annotations? In fact, to substantially reduce the annotation cost for segmentation tasks, weakly supervised segmentation methods with HBB annotation (Arun et al., 2020; Dai et al., 2015; Kulharia et al., 2020; Papandreou et al., 2015; Rajchl et al., 2016; Song et al., 2019; Tian et al., 2021) have been studied in the field of computer vision. Although these weakly

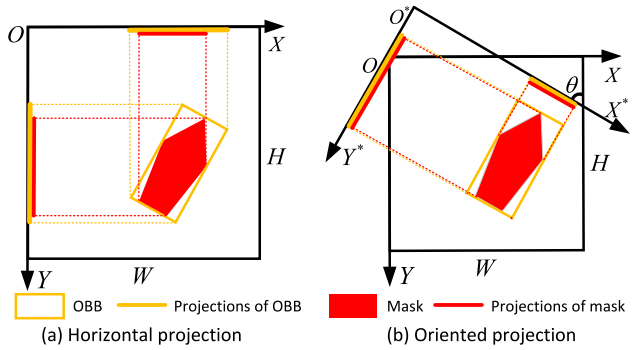


Fig. 3. Illustrations of (a) horizontal projection in rectangular image coordinate system and (b) oriented projection in rotated image coordinate system. The red lines represent the projections of mask on the coordinate system, while the yellow ones represent those of OBB. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

supervised methods achieve promising results, they are not compatible with RS instance segmentation methods due to the following reasons: (1) RS objects are usually labeled with OBB, while HBB-based methods (Dai et al., 2015; Kulharia et al., 2020) cannot be directly applied for RS instance segmentation methods. (2) Horizontal projection (Tian et al., 2021) produces large coordinate offsets due to the unsuitable coordinate system, and thus results in severe accuracy loss (see details in Section 4.2.1). As shown in Fig. 3(a), the horizontal projection of OBB differs from that of mask for a large gap (over 30%).

In this work, we intend to conduct the first study of weakly supervised remote sensing instance segmentation with OBB supervision, and propose a framework tailored for OBB annotation based on BoxInst (Tian et al., 2021) (namely **OBBInst**) to unify the incompatibility between existing HBB-based (Khoreva et al., 2017; Hsu et al., 2019; Tian et al., 2021) and OBB-based methods. OBBInst consists of OBB regression branch and mask regression branch to perform OBB prediction and instance segmentation, respectively. In addition, inspired by horizontal projection loss (Tian et al., 2021), we propose an oriented projection loss to eliminate the projection differences between OBB and mask by rotated image coordinate system, as shown in Fig. 3(b). Furthermore, since edge information is beneficial for mask prediction (He et al., 2017; Gong et al., 2021), we propose an edge similarity loss to further incorporate the Canny (1986) edge supervision for substantial performance improvement. Specifically, the edge similarity loss calculates the structural similarity (SSIM) between the predicted and GT edges generated by Canny to fully use the model-based edge prior in a data-driven manner. Several approaches (Rodriguez-Serrano et al., 2016; Gong et al., 2021; Cheng et al., 2023) have demonstrated that data-driven approaches can improve the efficiency of data usage and adaptability to complex data.

The main contributions can be summarized as follows:

1. We propose the first weakly supervised RS instance segmentation method with OBB supervision (namely **OBBInst**), which can largely reduce the annotation burden.

2. OBBInst breakthroughs the gap between HBB-based and OBB-based box-supervised methods, and can perform more precise object depicts by an oriented projection loss. In addition, OBBInst introduces an edge similarity loss to incorporate the model-based edge prior in a data-driven manner for further performance improvement.

3. Extensive experiments on iSAID (Waqas Zamir et al., 2019) and HRSC 2016 (Liu et al., 2017) datasets have demonstrated that OBBInst can surpass the existing HBB-based weakly supervised methods for a large margin with minor parameter and FLOPS increases. In addition, with low OBB annotation cost, OBBInst can achieve over 67.1% and 93.6% mask AP of their fully supervised performance in iSAID and HRSC datasets, respectively.

2. Related works

2.1. Box-supervised segmentation

In the field of computer vision, a few works attempted to obtain semantic masks using box annotations. As the pioneering work, Dai et al. (2015) proposed BoxSup to iteratively train a convolutional network to refine the estimated masks from region proposals generated by MCG (Pont-Tuset et al., 2016). Supervised by pseudo labels generated by GrabCut (Rother et al., 2004), Kulharia et al. (2020) proposed Box2Seg to predict per-class attention maps for false alarm elimination. Song et al. (2019) introduced a box-driven class-wise masking model (BCM) to predict pixel labels by calculating the mean filling rates between predictions and segment proposals generated by CRF (Arun et al., 2020) as prior cues. Note that, the aforementioned methods employ pseudo labels generated by unsupervised segmentation methods as supervision and develop elaborate methods for mask refinement. Therefore, they cannot work without mask annotation, and drastically decrease the training efficiency.

Since semantic segmentation cannot provide distinct results of densely arranged objects, box-supervised instance segmentation has raised more and more attention recently. The earliest framework SDI (Khoreva et al., 2017) utilized region proposals generated by MCG (Pont-Tuset et al., 2016) to refine the estimated masks by an iterative training strategy. Hsu et al. (2019) combined Mask R-CNN (He et al., 2017) with multiple instance learning (MIL), and sampled positive and negative bags based on region of interest (ROI) annotations. In conclusion, the aforementioned methods all require region segmentation proposals generated by unsupervised segmentation methods. To relieve the supervision burden, modified from CondInst (Tian et al., 2020a), BoxInst (Tian et al., 2021) introduced a projection loss and pairwise affinity loss to achieve instance segmentation without any auxiliaries. Note that, BoxInst is designed for HBB annotation, which cannot be directly applied to RS instance segmentation with OBB annotation. In addition, as shown in Fig. 3(a), projection loss (Tian et al., 2021) cannot provide precise object depicts, resulting in a severe performance drop. Our OBBInst can well address the aforementioned problems to unify the incompatibility between HBB-based and OBB-based methods and provide more precise object depicts.

2.2. Remote sensing segmentation

FCN (Long et al., 2015) has been widely used for semantic segmentation in natural images. Following this idea, various methods, including SegNet (Badrinarayanan et al., 2017), U-Net (Ronneberger et al., 2015), PSPNet (Zhao et al., 2017), and DeepLab (Chen et al., 2017) have been developed for substantial performance improvements. Based on the current powerful paradigms on generic semantic segmentation, RS semantic segmentation methods (Diakogiannis et al., 2020; Li et al., 2021b,a; Julius Fusic et al., 2022; Guo et al., 2018; Venugopal, 2020) have emerged in recent years. ResUNet-a (Diakogiannis et al., 2020) combined residual connections, atrous convolutions, and pyramid scene parsing pooling with MTL to perform RSIs segmentation. MAResU-Net (Li et al., 2021a) incorporated multi-scale features of U-Net, and designed a multi-scale skip connection with symmetric convolution for more distinct feature representation. MANet (Li et al., 2021b) employed multiple efficient attention modules to exploit contextual dependencies while alleviating the computational burden of attention modules. Cheng et al. (2023) adopted the Canny operator to extract the edge information of images as the auxiliary modality fusion for road segmentation in RSIs. Schuegraf et al. (2022) used morphology and watershed algorithms in traditional methods to post-process the deep network semantic output to generate instance-level segmentation. Qiu et al. (2024) proposed an efficient generative adversarial transformer (GATrans) to achieve high-precision remote sensing semantic segmentation while maintaining an extremely efficient size. RSSGLT (Satyawant et al., 2024) captured the global and local features by leveraging the benefits of the transformer and convolution mechanisms for remote sensing image segmentation.

Compared with the extensive explorations of semantic segmentation in RSIs, RS instance segmentation is rarely discussed. However, due to the ambiguous semantics of densely arranged objects shown in Fig. 2, instance segmentation raises more and more attention in RSIs. Jian et al. (2019) combined the modified U-Net (Ronneberger et al., 2015) with MTL to perform instance segmentation of buildings in RSIs. Gong et al. (2021) incorporated the boundary information in Mask R-CNN (He et al., 2017) by a penalty map for more discriminative edges of different objects. Based on FCN (Long et al., 2015), Zhang et al. (2021b) utilized semantic attention with extra supervision to strengthen the feature representation capability and reduce the background noise. The above methods are all fully supervised, which need extra supervision. Instead, our OBBInst can combine box supervision with model-based edge prior to perform instance segmentation in RSIs without per-pixel mask supervision.

3. Method

In this section, we introduce our OBBInst in detail. Specifically, Section 3.1 introduces the overall framework of our method. Sections 3.2 and 3.3 introduce the OBB and mask regression branch to perform OBB and mask predictions under OBB supervision.

3.1. Overall framework

As shown in Fig. 4, OBBInst consists of a backbone (He et al., 2016) for feature extraction, a feature pyramid network (FPN) (Papandreou et al., 2015) for multi-scale feature fusion, and two sub-branches (*i.e.*, OBB and mask regression branches) to generate the OBB and mask outputs. The input image is first sent to the ResNet (He et al., 2016) for basic feature extraction, which is then sent to FPN (Papandreou et al., 2015) for multi-level feature fusion. On the one hand, the fused multi-level feature (*i.e.*, P3–P7 features) are sent to five shared OBB heads to regress class, OBB and center-ness map (Tian et al., 2020b) predictions. On the other hand, the P3 feature is processed by convolutions to obtain mask feature F_{mask}, which is first concatenated by the relative coordinates (*i.e.*, $\tilde{F}_{mask}$) and then sent to the mask head for mask prediction. Note that, without GT mask, we utilize oriented projection loss to minimize the discrepancy between the projections of the mask predictions and GT OBBs. This essentially ensures that the tightest OBB covering the predicted mask matches the GT OBB. In addition, we utilize pairwise affinity loss and edge similarity loss for classification guidance and edge enhancement.

3.2. OBB regression branch

As shown in Fig. 5(a), OBB regression branch employs convolutional blocks to generate class prediction c, OBB prediction (l, t, r, b, θ) and center-ness score c_{tn} of each regression point. Among them, c represents the classification score, and (l, t, r, b) represents the distance from the regression point to the left, top, right, and bottom of OBB, respectively. θ represents the angle prediction of OBB. Note that, our method employs a pixel-by-pixel regression strategy to reduce missing detection. However, this approach unavoidably results in a large number of low-quality and center-offset boxes due to the imbalance between regression points and GT objects. Inspired by FCOS (Tian et al., 2020b), we employ a center-ness branch to suppress low-quality boxes by assigning a center-aware value c_{tn}, which can exhibit higher response values of pixels that are close to the center of OBB. c_{tn} can be formulated as:

$$c_{tn} = \sqrt{\frac{\min(l, r)}{\max(l, r)} \times \frac{\min(t, b)}{\max(t, b)}}. \tag{1}$$

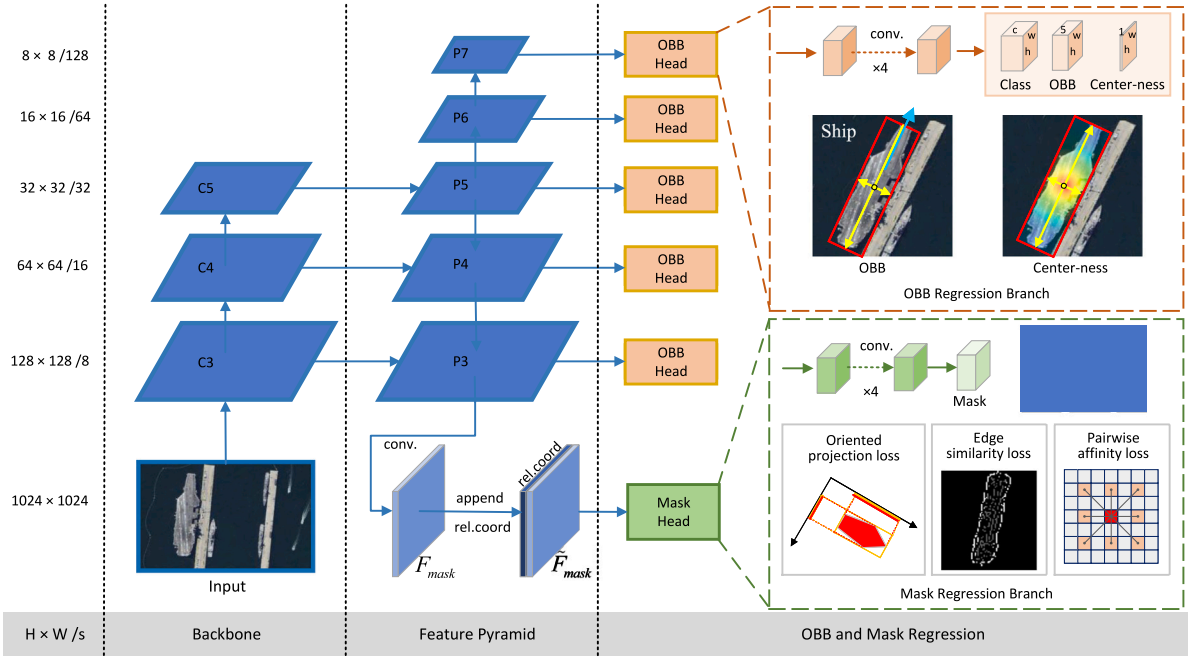


Fig. 4. The overall framework of OBBInst, which consists of three main parts: backbone, feature pyramid, and OBB and mask regression. C3–C5 and P3–P7 represent multi-level feature maps. The OBB regression branch consists of five shared OBB heads to perform class, OBB and center-ness prediction. P3 feature is first processed by a convolution layer and then concatenated by the relative coordinates, which are sent to the mask regression branch to generate the output masks. Note that, without GT masks, we utilize orienting projection loss, edge similarity loss, and pairwise affinity loss to supervise the network training.

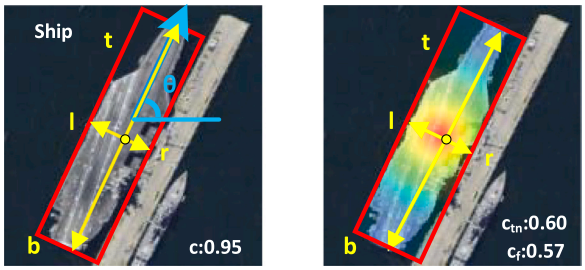


Fig. 5. An illustration of (a) OBB, class and (b) Center-ness map predictions. c represents the classification score, and (l, t, r, b) represents the distance from the regression point to the left, top, right, and bottom of OBB. θ represents the angle prediction of OBB. $c_{tn} \in [0, 1]$ is the center-ness score, which represents the distance from the center of the object. Red represents higher value, and blue represents lower value. Center-ness map are used to suppress the low-quality boxes in the inference stage. Note that, the final classification score c_f is the multiplication between c and c_{tn}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

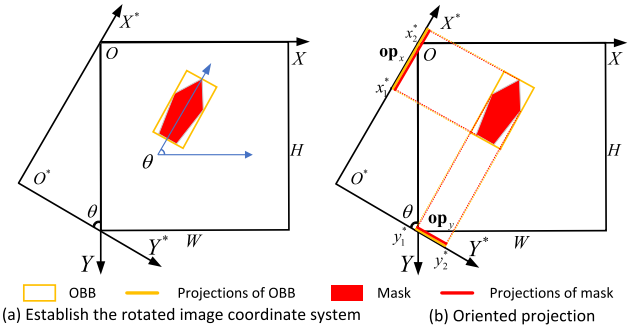


Fig. 6. An illustration of two-step oriented projection. (a) shows that the rotated image coordinate system $X^*O^*Y^*$ is generated by rotating the rectangle image coordinate system XOY via θ degree in clockwise manner, and (b) shows the oriented projection of OBB $(x_1^*x_2^*, y_1^*y_2^*)$ on the rotated image coordinate system along the X^*-axis and Y^*-axis.

As shown in Fig. 5(b), red represents higher value and blue represents lower value. c_{tn} decreases from 1 to 0 as the predicted location is far away from the center of the object. c_{tn} is used to adjust the confidence of each regression point to suppress the low-quality boxes, and the final confidence c_f is generated by the multiplication of classification score c and center-ness score c_{tn}.

We design an OBB regression loss (*i.e.*, L_{OBBreg}) to supervise the network training, including classification loss (*i.e.*, L_{cls}), OBB regression loss (*i.e.*, L_{reg}) and center-ness loss (*i.e.*, L_{center}), which can be formulated as:

$$L_{OBBreg} = L_{cls} + L_{reg} + L_{center}, \tag{2}$$

where L_{cls} represents focal loss (Lin et al., 2017) to address class imbalance and can be formulated as:

$$L_{cls} = -\alpha_t(1 - p_t)^{\gamma} \log(p_t), \tag{3}$$

$$p_t = \begin{cases} p, & \text{if y = 1} \\ 1 - p, & \text{otherwise} \end{cases}, \tag{4}$$

$$\alpha_t = \begin{cases} \alpha, & \text{if y = 1} \\ 1 - \alpha, & \text{otherwise} \end{cases}, \tag{5}$$

where $y \in \{0, 1\}$ represents the label and p represents the prediction probability. γ is the focal parameter to adjust the rate of weight reduction and is set to 2. α is the positive–negative sample ratio, and is set to 0.25.

L_{reg} represents KFIoU loss (Yang et al., 2022) to measure the difference between the predicted and GT OBB for boundary continuity improvement, and can be formulated as:

$$L_{reg} = 1 - \frac{\upsilon_{B_3}(Cov)}{\upsilon_{B_1}(Cov_1) + \upsilon_{B_2}(Cov_2) + \upsilon_{B_3}(Cov)}, \tag{6}$$

where υ_B represents the union region of target boxes and predicted boxes. Cov represents the covariance of two Gaussian distributions.

L_{center} represents cross entropy loss (Tian et al., 2020b) for regression point refinement and low-quality OBB suppression, and can be

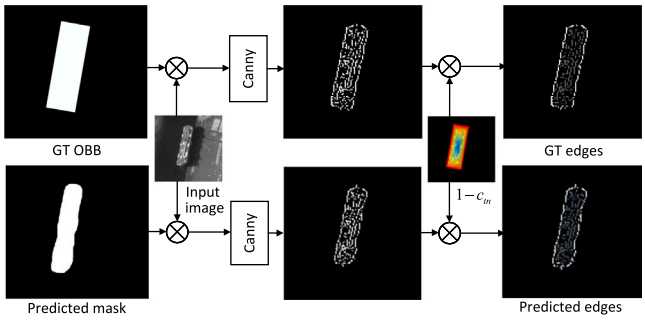


Fig. 7. An illustration of edge similarity loss. ⊗ represents the Hadamard production. Inverse center-ness scores (*i.e.*, $1-c_{tn}$) are used to suppress the inner edges of the GT OBBs, and we employ SSIM between the predicted and GT edges as the loss function for edge supervision.

formulated as:

$$L_{center} = -\sum_{i=1}^{C} x_i \log y_i, \tag{7}$$

where C represents the total number of the GT label y. x_i and y_i represent the label and score of i_{th} predicted OBB.

3.3. Mask regression branch

As shown in Fig. 4, P3 feature is first sent to convolutions to generate feature F_{mask}. Then F_{mask} is concatenated with its relative coordinates and then sent to the mask head (Tian et al., 2020a) to predict the instance mask. To supervise the network training, we design a mask regression loss L_{Seg}, which can be formulated as:

$$L_{Seg} = L_{OOBB} + L_{pairwise} + L_{ES}, \tag{8}$$

where L_{OOBB} represents oriented projection loss that minimizes the discrepancy of the oriented projections between the predicted and the GT mask. $L_{pairwise}$ represents pairwise affinity loss that encourages predicted and GT masks to have the same pairwise label similarity in proximal pixels. L_{ES} represents edge similarity loss that incorporates the model-based edge prior in a data-driven manner to refine the output segmentation results. The overall loss function is a summation of the mask and the OBB regression losses, and the details of L_{OOBB} and L_{ES} are presented in Sections 3.3.1 and 3.3.2.

3.3.1. Oriented projection loss

As shown in Fig. 6(a), to generate the rotated image coordinate system $X^*O^*Y^*$, we turn the rectangle image coordinate system XOY by θ degrees in a counterclockwise manner. The process can be formulated as:

$$\begin{cases} x^* = x\cos\theta - y\sin\theta, \\ y^* = x\sin\theta + y\cos\theta, \end{cases} \tag{9}$$

where (x, y) represents a point in XOY and (x^*, y^*) represents the corresponding point in $X^*O^*Y^*$. Let $\mathbf{m}_{OBB} \in \mathbb{R}^{H\times W}$ be the OBB mask in rotated image coordinate system $X^*O^*Y^*$. As shown in Fig. 6(b), the oriented projection in X^*-axis and in Y^*-axis can be formulated as:

$$\begin{cases} \mathrm{OProj}_x(\mathbf{m}_{OBB}) = \max_y(\mathbf{m}_{OBB}) = \mathbf{op}_x, \\ \mathrm{OProj}_y(\mathbf{m}_{OBB}) = \max_x(\mathbf{m}_{OBB}) = \mathbf{op}_y, \end{cases} \tag{10}$$

where $\mathrm{OProj}_x : \mathbb{R}^{H\times W} \to \mathbb{R}^W$ and $\mathrm{OProj}_y : \mathbb{R}^{H\times W} \to \mathbb{R}^H$ represents oriented projection along X^* and Y^* axis. $\max_x$ and $\max_y$ are the max operations along with X^*-axis and Y^*-axis, and $\mathbf{op}_x \in \mathbb{R}^W$, $\mathbf{op}_y \in \mathbb{R}^H$ denotes the 1-D segmentation mask on X^* and Y^* axis, respectively.

Finally, we calculate the oriented projection loss (*i.e.*, L_{OOBB}) to supervise the network training, which can be formulated as:

$$\begin{aligned} L_{OOBB} &= L(\mathrm{OProj}_x(\mathbf{m}^g), \mathrm{OProj}_x(\mathbf{m}^p)) \\ &\quad + L(\mathrm{OProj}_y(\mathbf{m}^g), \mathrm{OProj}_y(\mathbf{m}^p)) \\ &= L(\mathbf{op}_x^g, \mathbf{op}_x^p) + L(\mathbf{op}_y^g, \mathbf{op}_y^p) \quad , \end{aligned} \tag{11}$$

where $\mathbf{m}^g$, $\mathbf{m}^p$ represent the GT and predicted masks, respectively. $L(\cdot,\cdot)$ represents the Dice loss (Tian et al., 2020a), and can be formulated as:

$$L(X, Y) = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|}, \tag{12}$$

where X and Y represent the oriented projection of the GT and predicted masks, respectively.

3.3.2. Edge similarity loss

Since edge information is beneficial to mask prediction (He et al., 2017; Gong et al., 2021), we incorporate the Canny edge (Canny, 1986) as an additional supervision due to its high response to edges and extremely low computational cost. As shown in Fig. 7, to generate the GT edges, we first utilize Canny algorithm (Canny, 1986) to segment the edges, and then eliminate the edges beyond the GT bboxes. Then the edge image is multiplied with the inverse center-ness score (*i.e.*, $\hat{c}_{tn} = 1 - c_{tn}$) to suppress the inner edges of the GT OBBs. The predicted edge is generated in the same way by replacing GT OBB with mask prediction. We employ SSIM to measure the difference between the predicted and the GT edges and introduce an edge similarity loss (*i.e.*, L_{ES}), which can be formulated as:

$$\begin{aligned} L_{ES} = \mathrm{SSIM}(&\mathrm{Canny}(I \otimes G) \times (1 - c_{tn}), \\ &\mathrm{Canny}(I \otimes P) \times (1 - c_{tn})), \end{aligned} \tag{13}$$

where $I \in \mathbb{R}^{H\times W}$ represents the grayscale of the input image and $G, P \in \mathbb{R}^{H\times W}$ represent the GT OBB and the mask prediction. c_{tn} is the center-ness score map. $\mathrm{Canny}(\cdot)$ represents the canny operation and $\mathrm{SSIM}(\cdot,\cdot)$ represents the structural similarity measure.

3.3.3. Pairwise affinity loss

We employ pairwise affinity loss (Tian et al., 2021) to encourage the predicted and GT masks to exhibit the same pairwise label similarity on neighborhood pixels. Specifically, we first build an undirected graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_m\}$ is the set of the pixels in the image, and $E = \{e_1, e_2, \ldots, e_n\}$ is the set of the edges. Note that, each pixel in V is connected with its $K \times K - 1$ dilated neighborhood pixels and K is set to 3. Then we define $y_{e_i} \in \{0, 1\}$ be the label for edge e_i, where $y_{e_i} = 1$ represents that two pixels linked by edge e_i have the same GT label, and $y_{e_i} = 0$ is vice versa. Let pixels (a, b) and (c, d) be the two endpoints of edge e_i. The network prediction $\tilde{m}_{a,b} \in \{0, 1\}$ can be viewed as the probability of pixel (a, b) being foreground. Then the probability of $y_{e_i} = 1$ is defined as:

$$P(y_{e_i} = 1) = \tilde{m}_{a,b} \cdot \tilde{m}_{c,d} + (1 - \tilde{m}_{a,b}) \cdot (1 - \tilde{m}_{c,d}). \tag{14}$$

We employ binary cross entropy (BCE) loss for optimization, which can be formulated as:

$$\begin{aligned} L_{Pairwise} = -\frac{1}{N} \sum_{e\in E_{in}} &(y_e \log P(y_e = 1) \\ &+ (1 - y_e) \log P(y_e = 0)), \end{aligned} \tag{15}$$

where E_{in} is the set of the edges containing at least one pixel in the box. N is the number of edges in E_{in}.

4. Experiments

In this section, we first introduce the experiment settings, and then conduct ablation studies to validate our method. Finally, we compare our OBBInst with several segmentation methods under full supervision and box supervision to demonstrate the superiority of our method.

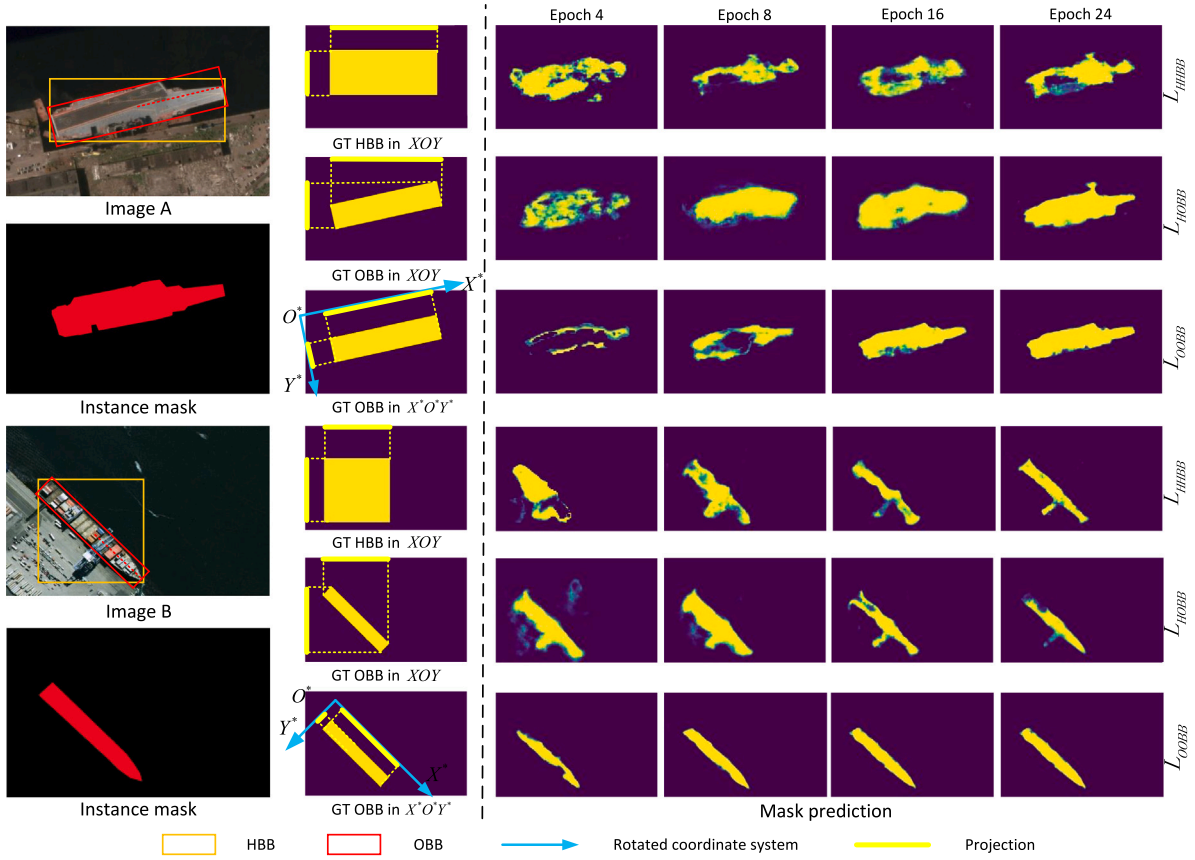


Fig. 8. Visualizations of mask prediction under different projection methods in the 4_{th}, 8_{th}, 16_{th}, and 24_{th} epochs during training. $X^*O^*Y^*$ and XOY represent the rotated and rectangle image coordinate system, respectively. L_{HHBB}, L_{HOBB} and L_{OOBB} represent the corresponding losses of horizontal projection with HBB annotation, horizontal projection with OBB annotation, and oriented projection with OBB annotation, respectively.

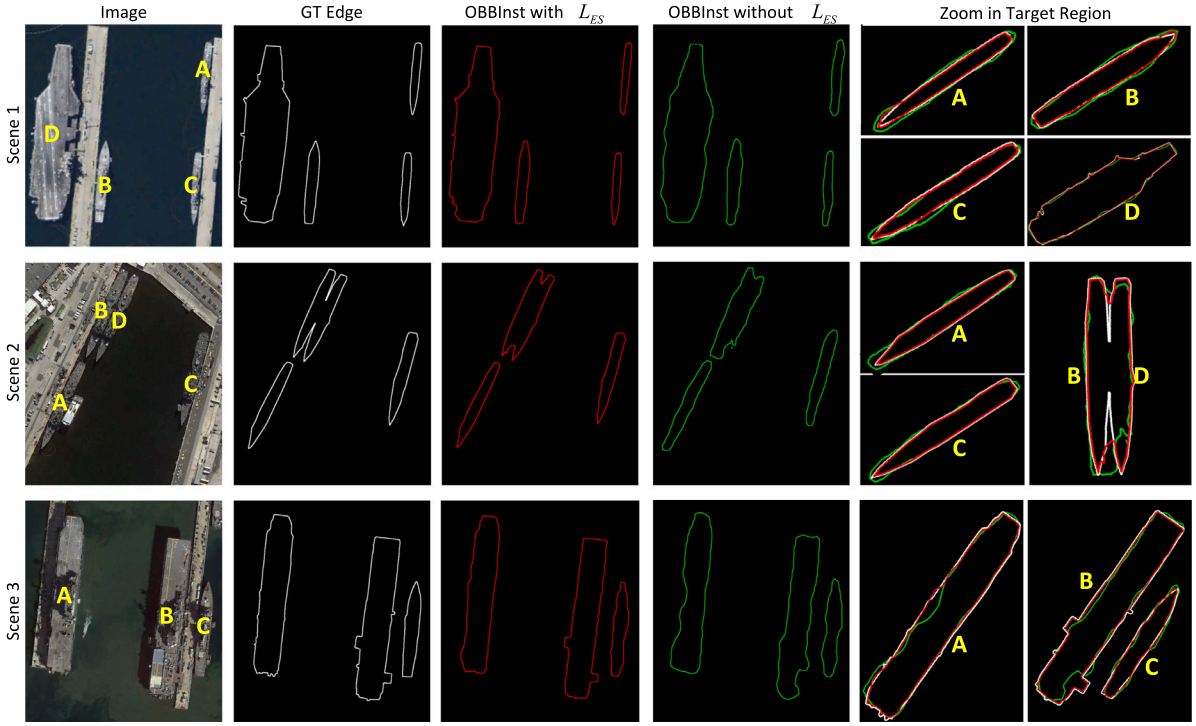


Fig. 9. Visualizations of edge predictions of OBBInst trained with and without edge similarity loss. We superimpose GT edges with two edge predictions, and display the patch of each object separately. Note that, we rotate the image patch at a certain angle for better visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Ablation studies of different projection methods for instance segmentation. "H-HBB" represents horizontal projection with HBB annotation. "H-OBB" represents horizontal projection with OBB annotation. "O-OBB" represents oriented projection with OBB annotation. Best results are shown in **boldface**.

Projection	Evaluation metrics (%)					
	AP^m	AP^m_{50}	AP^m_{75}	Prec.	Rec.	F1
H-HBB	72.32	77.64	73.57	72.06	**79.57**	75.53
H-OBB	68.92	72.33	69.41	62.32	75.41	68.24
O-OBB	**75.90**	**82.45**	**76.51**	**90.55**	76.20	**82.76**

Table 2
Ablation results of different loss functions for object detection. L_{HHBB} represents horizontal projection loss with HBB annotation, and L_{OOBB} represents the oriented projection loss with OBB annotation. L_{ES} represents the edge similarity loss. Best results are shown in **boldface**.

L_{HHBB}	L_{OOBB}	L_{ES}	AP^b (%)	AP^b_{50} (%)	AP^b_{75} (%)
✓			70.82	86.86	71.21
✓		✓	71.44	87.78	71.95
	✓		71.90	88.06	72.34
	✓	✓	**72.08**	**88.11**	**72.63**

4.1. Experimental settings

In this subsection, we sequentially introduce the datasets, the evaluation metrics and the implementation details.

4.1.1. Datasets

The iSAID (Waqas Zamir et al., 2019) dataset contains 2806 images, 655,451 object instances across 15 categories. As the first large-scale RS instance segmentation dataset, iSAID fully reflects the common features and scale distribution differences in RSIs. Therefore, performance evaluation on iSAID for RS instance segmentation algorithms is highly reliable.

The HRSC 2016 (Li et al., 2021a) dataset (*i.e.*, a specific version that contains OBB and mask annotation) presents several challenges for the box-supervised RSIs segmentation on the ship. First, ships near the shore are densely arranged, and thus HBB annotation presents a high overlap rate, and cannot provide proper supervision for network training. Second, the complex backgrounds of RSIs (*e.g.*, nearshore textures) and high similarity among different ship result in a high false alarms rate. In addition, since HRSC 2016 dataset includes both OBB and HBB annotations for ship objects, it is suitable to verify the effectiveness of the proposed method.

4.1.2. Evaluation metrics

For performance evaluation of pixel-wise instance segmentation, we use pixel-level precision (Prec.) and recall (Rec.) to evaluate the localization and classification accuracy of the predicted masks. F1-score and mask average precision (AP^m) are also used for comprehensive evaluation. For performance evaluation of box-wise object detection, we use box-level average precision (AP^b) based on the Intersection of Union (IoU) and Rotated IoU (RIoU) for HBB and OBB, respectively. The evaluation metrics can be formulated as follows:

$$\begin{aligned} Prec. &= (TP+TN)/(TP+TN+FP+FN), \\ Rec. &= TP/(TP+FN), \\ \mathrm{F1} &= 2\times Rec.\times Prec./(Prec.+Rec.), \\ \mathrm{AP}^m &= \int_0^1 p^m(r^m)dr, \mathrm{AP}^b = \int_0^1 p^b(r^b)dr, \end{aligned} \tag{16}$$

where TP, TN, FP and FN represent the numbers of true positive, true negative, false positive, and false negative samples, respectively. p^m, r^m and p^b, r^b represent the pixel-level and box-level precision and recall, respectively. Moreover, we introduce AP_S, AP_M and AP_L for performance evaluation on objects with different sizes (*i.e.*, small, medium and large). AP_{50} and AP_{75} are calculated under IoU = 0.5 and IoU = 0.75, respectively.

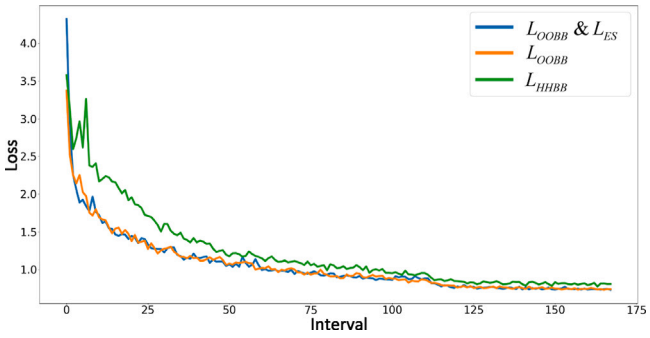


Fig. 10. Loss curves under different loss functions during the training process. L_{HHBB} represents horizontal projection loss with HBB annotation, and L_{OOBB} represents the oriented projection loss with OBB annotation. L_{ES} represents the edge similarity loss.

4.1.3. Implementation details

We evaluated our method on a PC with an Nvidia GTX-3090Ti GPU (24 GB), Intel Core i7 CPU and Ubuntu18.04. All models are implemented by MMDetection (Chen et al., 2019) code library. We use the standard partition of training and validation sets in both datasets. We cropped the original images into 1024 × 1024 patches with 128 overlapped pixels, and used the cropped patches for training and inference. In the training phase, random flip was used for data augmentation.

If not specified, we employed ResNet50 (He et al., 2016) as the default backbone, which was initialized with ImageNet (Deng et al., 2009) pre-trained weights. All models were trained for 24 epochs with a batch size of 2. The initial learning rate was set to 0.01 and reduced by a factor of 10 at epochs 16 and 22, respectively.

4.2. Ablation studies

In this subsection, we compare our oriented projection with other projection methods to validate its effectiveness. Then we investigate the influence of different loss functions on mask prediction, OBB prediction and network convergence.

4.2.1. Different projection methods

To evaluate the effectiveness of the proposed oriented projection, we employ BoxInst (with ResNet-50 as the backbone) as the baseline model, and design three variants (*i.e.*, H-HBB, H-OBB, O-OBB) to compare three different projection methods. Among them, H-HBB represents horizontal projection with HBB annotation (*i.e.,* projection method in BoxInst (Tian et al., 2021)). H-OBB represents horizontal projection with OBB annotation. O-OBB represents oriented projection with OBB annotation.

Quantitative segmentation results are listed in Table 1. It can be observed that O-OBB can achieve the highest AP^m (75.90%) and F1-score (82.76%), 3.58% and 7.23% higher than H-HBB. This is because, oriented projection with OBB annotation can provide more precise object depicts and more reasonable projection guidance for box-to-mask learning, and thus results in superior performance. It is worth noticing that, H-OBB achieves the lowest AP^m and F1-score, lower than H-HBB for 3.40% and 7.39%. As shown in Fig. 3(a), even if OBB provides more precise target depicts, inaccurate mask projection can introduce many false alarms by coordinate offset, and thus greatly degrade the segmentation performance.

As shown in Fig. 8, we visualize the mask predictions of different projection methods in the 4th, 8th, 16th, and 24th training epochs. Note that, corresponding projection loss is used to supervise the network training. It can be observed that H-HBB shows inferior segmentation performance and slow network convergence. Due to more precise target depicts, H-OBB introduces more visual-pleasing mask

Table 3
Ablation results of different loss functions for instance segmentation. L_{HHBB} represents horizontal projection loss with HBB annotation, and L_{OOBB} represents the oriented projection loss with OBB annotation. L_{ES} represents the edge similarity loss. Best results are shown in **boldface**.

L_{HHBB}	L_{OOBB}	L_{ES}	AP^m (%)	AP^m_{50} (%)	AP^m_{75} (%)	Prec. (%)	Rec. (%)	F1 (%)	Params (MB)	GFlops	FPS
✓			72.32	77.64	73.57	72.06	**79.57**	75.53	33.85	122.58	5.3
✓		✓	73.19	80.92	74.80	80.74	78.20	78.46	33.85	122.60	5.3
	✓		75.90	82.45	76.51	90.55	76.20	82.76	36.21	135.32	5.3
	✓	✓	**76.12**	**83.50**	**76.84**	**90.83**	77.45	**83.60**	36.21	135.34	5.3

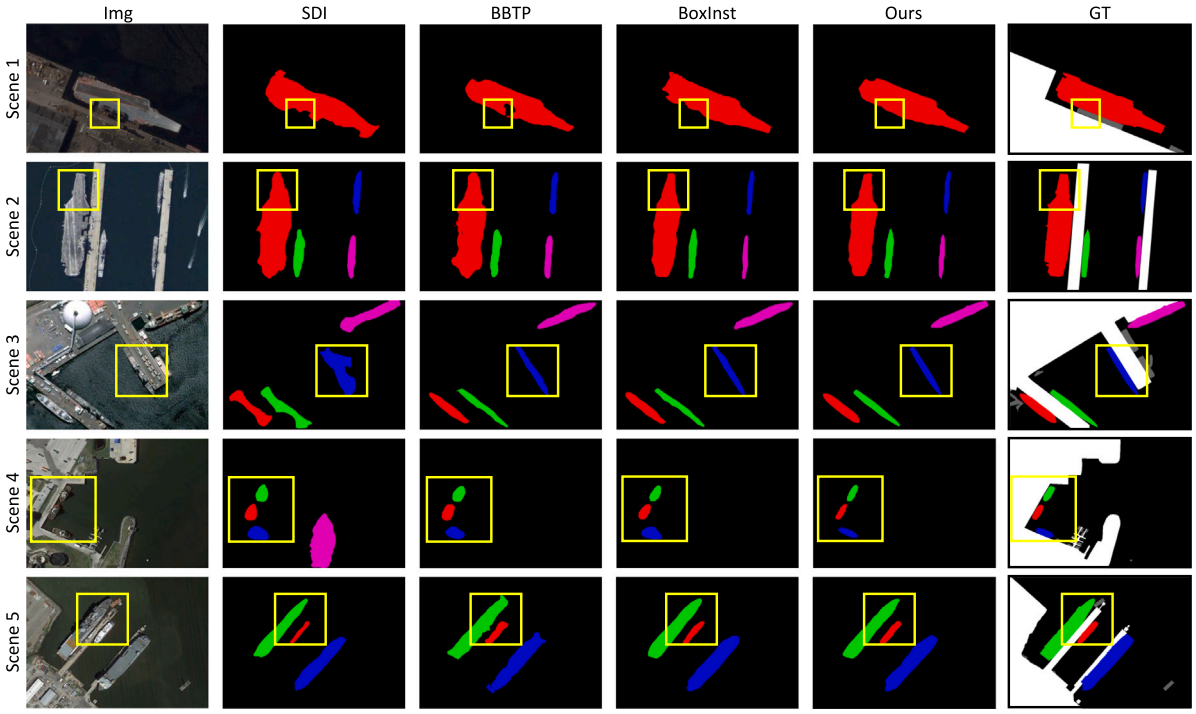


Fig. 11. Qualitative results achieved by different methods on HRSC 2016 dataset. White and gray regions in GT masks represent the land and other unknown objects. We highlight the significant region by yellow boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Quantitative results achieved by fully supervised (Full Sup.) methods and box-supervised (Box Sup.) methods on iSAID and HRSC 2016 datasets. "#Sched." represents the training strategy, and "$n\times$" represents $n\times 12$ training epochs. "Params." represents the parameters of the algorithms. Best results are shown in **boldface**.

	Methods	Backbone	#Sched.	Params.	iSAID						HRSC 2016					
					AP^m	AP^m_{50}	AP^m_{75}	AP^m_S	AP^m_M	AP^m_L	AP^m	AP^m_{50}	AP^m_{75}	AP^m_S	AP^m_M	AP^m_L
Full Sup.	Mask R-CNN (He et al., 2017)	ResNet 50-FPN	3×	44.4M	24.7	50.3	22.5	13.4	30.5	38.3	79.6	84.3	79.0	61.2	79.0	81.6
	Mask R-CNN[a] (He et al., 2017)	ResNet 50-FPN	3×	44.4M	33.2	55.9	35.4	**35.3**	**45.3**	21.8	83.4	85.0	80.2	62.7	82.5	83.0
	PANet (Liu et al., 2018)	ResNet 50-FPN	2×	66.1M	34.1	56.3	36.2	19.2	42.4	**46.8**	84.6	85.1	80.7	63.2	84.1	**83.4**
	SS-MRCNN (Zhang et al., 2021b)	ResNet 101-FPN	2×	72.3M	**35.6**	**57.4**	**38.3**	25.1	45.2	37.7	**84.9**	**85.5**	**82.2**	**63.5**	**85.7**	82.7
	PolarMask (Xie et al., 2020)	ResNet 101-FPN	2×	38.5M	21.1	45.6	22.3	16.0	26.8	36.1	75.3	81.2	74.8	58.2	72.0	79.0
Box Sup.	SDI (Khoreva et al., 2017)	VGG-16	2×	22.4M	12.9	18.3	13.5	8.6	15.3	17.5	45.5	51.4	46.9	28.1	42.2	45.2
	BBTP (Hsu et al., 2019)	ResNet 101-FPN	1×	42.6M	16.5	22.6	16.8	12.1	20.9	31.5	58.2	61.1	57.0	40.9	60.7	61.5
	BoxInst (Tian et al., 2021)	ResNet 50-FPN	3×	33.9M	20.5	40.6	20.5	13.6	23.6	34.0	72.3	77.6	73.6	53.3	70.2	73.0
	BoxInst (Tian et al., 2021)	ResNet 101-FPN	3×	52.9M	21.6	43.9	22.3	17.9	27.4	36.8	74.7	78.1	73.3	54.1	70.3	74.3
	OBBInst	ResNet 50-FPN	2×	36.2M	21.2	45.3	22.1	16.5	26.0	36.6	76.1	83.5	76.8	58.3	74.1	79.6
	OBBInst	ResNet 101-FPN	2×	55.2M	**23.9**	**45.5**	**24.9**	**20.2**	**27.9**	**36.9**	**79.5**	**85.5**	**78.9**	**62.4**	**76.8**	**80.2**

[a] Indicates that Mask R-CNN trained with edge label.

prediction and accelerates network convergence. However, the coordinate offset of incorrect mask projection leads to noisy predictions. Compared with H-HBB and H-OBB, our oriented projection with OBB annotation under rotated image coordinate system $X^*O^*Y^*$ guarantees faster convergence speed and higher prediction accuracy.

4.2.2. *Different loss functions for instance segmentation*

We conduct ablation experiments to investigate the influence of different loss functions for instance segmentation, and the results are shown in Table 3. Compared with horizontal projection loss L_{HHBB} with HBB annotation in BoxInst (Tian et al., 2021), oriented projection loss L_{OOBB} with OBB annotation can provide more precise clues of target mask, and thus introduces significant performance improvement with a reasonable recall drop. To validate the effectiveness and generalization of edge similarity loss L_{ES}, we add L_{ES} to L_{HHBB} and L_{OOBB}, and retrain the network from scratch on HBB and OBB annotations, respectively. Note that, pairwise affinity loss is used as the default setting for fair comparison. It can be observed that edge similarity loss introduces substantial performance improvements on both losses, which demonstrates that edge supervision improves the accuracy of

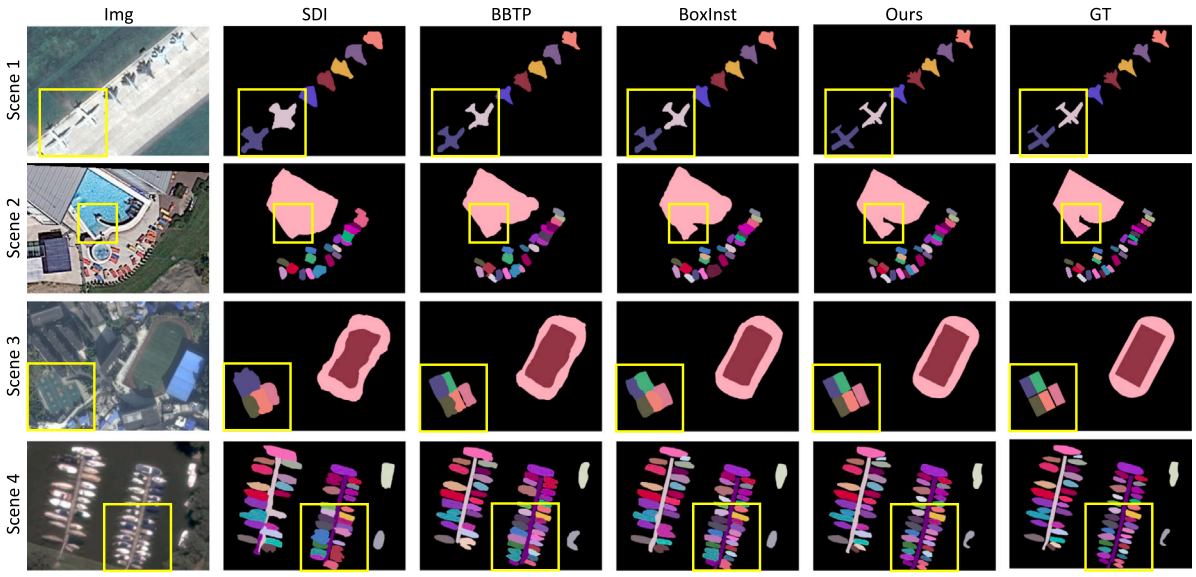


Fig. 12. Qualitative results achieved by different methods on iSAID dataset. We highlight the significant region by yellow boxes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pixel-wise classification. In addition, oriented projection loss introduces an additional angle prediction branch to supervise OBB, and thus results in 2.4 M and 12.8 G increases in terms of parameters and FLOPs. However, the increases introduced by edge similarity loss are minor that can be ignored. Despite of parameters and FLOPs increases, inference speed (FPS) is not affected to maintain the same as BoxInst (Tian et al., 2021).

We visualize the edge of the predictions of OBBInst trained with and without L_{ES} in Fig. 9. For better visualization, we first rotate the image at a certain angle, and then visualize the zoom-in region of each target with overlapped edges (*i.e.,* GT edges and two edge predictions). It can be observed that green edges (*i.e.*, OBBInst trained without L_{ES}) usually exceed the ground truth (GT) edges, while the red edges (*i.e.*, OBBInst trained with L_{ES}) are closer to GT edges. Specifically, for object D in image 1, red edges can better preserve the small bumps on the left of the object. In addition, objects B and D in scene 2 are densely arranged, and green edges identify B and D as one object while red edges tend to distinguish B from D. It is demonstrated that additional edge supervision can benefit edge identification, and thus improve the performance of instance segmentation.

4.2.3. Different loss functions for object detection

We conduct ablation experiments to investigate the influence of different loss functions for object detection, and the results are shown in Table 2. Similar to the conclusions in Section 3.3.2, compared with L_{HHBB}, L_{OOBB} introduces more precise target depicts, and thus results in great performance gain (1.12% in AP^b). L_{ES} provides additional edge supervision to show consistent performance improvements both on L_{HHBB} and L_{OOBB}. It is worth noting that even if L_{OOBB} and L_{ES} are designed to promote the mask prediction branch, they also benefit object detection by providing more sophisticated target clues.

4.2.4. Different loss functions for network convergence

To make quantitative analyses of network convergence under different loss functions, we display the loss curves of networks trained by L_{HHBB}, L_{OOBB}, L_{OOBB} & L_{ES} in Fig. 10. It can be observed that the loss curve of the network trained by L_{OOBB} & L_{ES} declines the fastest and the lowest, which fully demonstrates the superiority of our method. In addition, the loss curve of the network trained by L_{OOBB} declines faster and lower than the network trained by L_{HHBB}. Moreover, at the 10th iteration stage, the network trained by L_{HHBB} has a loss fluctuation, while the network trained by L_{OOBB} convergent steadily, which demonstrates that oriented projection loss can stabilize the training process. Furthermore, visualization in Fig. 8, shows that L_{OOBB} can supervise the network to predict accurate masks at a very early stage.

4.3. Comparisons with state-of-the-art methods

We compare our proposed OBBInst with current state-of-the-art fully supervised segmentation methods Mask R-CNN (He et al., 2017), PANet (Liu et al., 2018), SS-MRCNN (Zhang et al., 2021b), PolarMask (Xie et al., 2020) and box-supervised instance segmentation methods SDI (Khoreva et al., 2017), BBTP (Hsu et al., 2019), BoxInst (Tian et al., 2021) on the iSAID (Waqas Zamir et al., 2019) and the HRSC 2016 (Li et al., 2021a) datasets. Note that, we use HBB annotation to train networks that are not compatible with OBB annotations.

Quantitative results are shown in Table 4. It can be observed that our method can achieve 23.9% mask AP in the iSAID dataset, 79.5% mask AP in the HRSC 2016 dataset, which outperforms the previous state-of-the-art method (Tian et al., 2021) over 2.3% and 4.8% with the same backbone and training settings. In addition, OBBInst also presents competitive performance as compared with current state-of-the-art fully supervised instance segmentation methods. Without mask annotations to supervise the network training, OBBInst can even perform superior than some recent fully supervised methods such as PolarMask (Xie et al., 2020) (*e.g.*, 21.1% vs. 23.9% of AP^m in the iSAID dataset) and Mask R-CNN (He et al., 2017) (*e.g.*, 79.6% v.s. 79.5% of AP^m in the HRSC 2016 dataset). The above experimental results demonstrate that OBBInst dramatically narrows the performance gap between the fully supervised and box-supervised instance segmentation algorithms, and reveals the great potential of box-supervised in RS instance segmentation at the first time.

Qualitative results achieved by different methods on HRSC 2016 and iSAID datasets are shown in Figs. 11 and 12. It can be observed from Fig. 11 that OBBInst exhibits higher overlapped rate and lower false alarm rate with the GT mask. In addition, OBBInst shows high robustness to objects with arbitrary orientations due to more precise target depicts by oriented projection method and corresponding loss function. It can be observed from Fig. 12 that, SDI and BBTP misidentify multiple densely arranged small objects as a single instance in scenes 2 and 4. BoxInst shows improved performance, but also fails in some instances due to imprecise depicts and semantic ambiguity of HBB annotations. Compared with them, our OBBInst can well address the

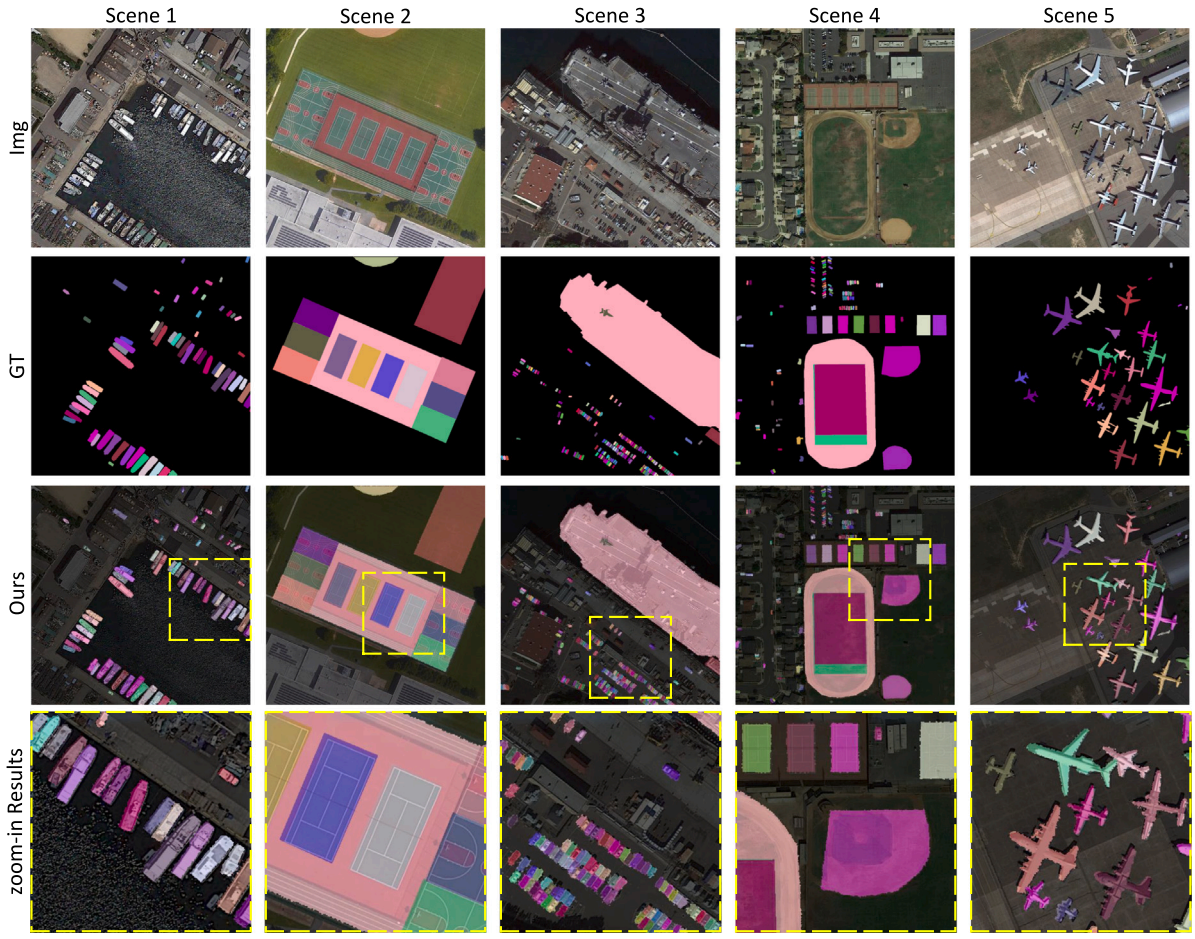


Fig. 13. Qualitative results achieved by OBBInst on iSAID dataset. We zoom in the regions with densely arranged objects and complex contours in the last row.

multiple densely arranged objects and overcome the drawbacks of unsuitable HBB annotation and projection.

Qualitative results achieved by OBBInst on iSAID dataset are shown in Fig. 13. It is observed that OBBInst performs better on objects with more regular contours (*e.g.*, tennis courts in the second column). For objects with complex contours (*e.g.*, aircrafts in the fifth column), OBBInst can basically segment the object contours with slight performance degradation. For densely arranged objects (*e.g.*, ships in the first column and cars in the third column), OBBInst shows high robustness to achieve precise instance segmentation, which fully demonstrates the effectiveness and superiority of our method.

5. Conclusion

In this work, we propose the first work to achieve RSIs instance segmentation using OBB supervision. In our method, we propose an OBBInst framework together with oriented projection loss to perform precise instance segmentation. In addition, an edge similarity loss is introduced to incorporate the Canny edge supervision in a data-driven manner. Extensive experiments have demonstrated the effectiveness and superiority of OBBInst. In addition, OBBInst can even surpass some existing fully supervised methods, which demonstrates the great potential of box-supervised methods in remote sensing instance segmentation.

CRediT authorship contribution statement

Xu Cao: Methodology, Formal analysis, Data curation, Conceptualization. **Huanxin Zou:** Resources, Project administration, Funding acquisition. **Jun Li:** Supervision, Resources, Project administration. **Xinyi Ying:** Formal analysis. **Shitian He:** Validation.

Declaration of competing interest

All authors disclosed no relevant relationships.

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62071474. All the authors listed have approved the manuscript that is enclosed.

References

Ali, S.M., Krishna, A.V., Kuttippurath, J., Gupta, A., Tirkey, A., Raman, M., Sahay, A., 2022. Improvement in estimation of phytoplankton size class in Arabian sea using remote sensing seasurements. IEEE Trans. Geosci. Remote Sens. 60, 1–12.

Arun, A., Jawahar, C., Kumar, M.P., 2020. Weakly supervised instance segmentation by learning annotation consistent instances. In: European Conference on Computer Vision. pp. 254–270.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L., 2016. What's the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision. pp. 549–565.

Bhagavathy, S., Manjunath, B.S., 2006. Modeling and detection of geospatial objects using texture motifs. IEEE Trans. Geosci. Remote Sens. 44 (12), 3706–3715.

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. (6), 679–698.

Chanlongrat, W., Apichanapong, T., Sinngam, P., Chaisangmongkon, W., 2022. A semi-automated system for person re-identification adaptation to cross-outfit and cross-posture scenarios. Appl. Intell. 1–20.

Chen, X., Ma, L., Du, Q., 2021a. Oriented object detection by searching corner points in remote sensing imagery. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chen, Z., Shang, Y., Python, A., Cai, Y., Yin, J., 2021b. DB-BlendMask: Decomposed attention and balanced blendmask for instance segmentation of high-resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–15.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D., 2019. MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155.

Cheng, B., Tian, M., Jiang, S., Liu, W., Pang, Y., 2023. Multi-task learning and multimodal fusion for road segmentation. IEEE Access 18947–18959.

Dai, J., He, K., Sun, J., 2015. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: IEEE International Conference on Computer Vision. pp. 1635–1643.

Dai, X., Xia, M., Weng, L., Hu, K., Lin, H., Qian, M., 2023. Multi-scale location attention network for building and water segmentation of remote sensing image. IEEE Trans. Geosci. Remote Sens..

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162, 94–114.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2009. The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. 88, 303–308.

Gong, Y., Zhang, F., Jia, X., Mao, Z., Huang, X., Li, D., 2021. Instance segmentation in very high resolution remote sensing imagery based on hard-to-segment instance learning and boundary shape analysis. Remote Sens. 14 (1), 23.

Guo, Z., Shengoku, H., Wu, G., Chen, Q., Yuan, W., Shi, X., Shao, X., Xu, Y., Shibasaki, R., 2018. Semantic segmentation for urban planning maps based on U-net. In: IEEE International Geoscience and Remote Sensing Symposium. pp. 6187–6190.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: IEEE International Conference on Computer Vision. pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

He, X., Zhou, Y., Zhao, J., Zhang, D., Yao, R., Xue, Y., 2022. Swin transformer embedding unet for remote sensing image semantic segmentation. IEEE Trans. Geosci. Remote Sens. 60, 1–15. http://dx.doi.org/10.1109/TGRS.2022.3144165.

Hsu, C.-C., Hsu, K.-J., Tsai, C.-C., Lin, Y.-Y., Chuang, Y.-Y., 2019. Weakly supervised instance segmentation using the bounding box tightness prior. Adv. Neural Inf. Process. Syst. 32.

Hu, H., Feng, C., Cui, X., Zhang, K., Bu, X., Yang, F., 2023. A sample enhancement method based on simple linear iterative clustering superpixel segmentation applied to multibeam seabed classification. IEEE Trans. Geosci. Remote Sens. 61, 1–15.

Jian, H., Qin, Q., Xu, W., Sui, J., 2019. Instance segmentation of buildings from high-resolution remote sensing images with multitask learning. J. Peking Univ. 55 (6), 1067–1077.

Julius Fusic, S., Hariharan, K., Sitharthan, R., Karthikeyan, S., 2022. Scene terrain classification for autonomous vehicle navigation based on semantic segmentation method. Trans. Inst. Meas. Control 44 (13), 2574–2587.

Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B., 2017. Simple does it: Weakly supervised instance and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 876–885.

Kulharia, V., Chandra, S., Agrawal, A., Torr, P., Tyagi, A., 2020. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In: European Conference on Computer Vision. pp. 290–308.

Li, Y., Wang, Z., Wang, J., Wang, P., 2021c. SDCDet: Robust remote sensing object detection based on instance segmentation direction correction. In: International Conference on Pattern Recognition and Artificial Intelligence. pp. 385–389.

Li, R., Zheng, S., Duan, C., Su, J., Zhang, C., 2021a. Multistage attention resu-net for semantic segmentation of fine-resolution remote sensing images. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M., 2021b. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Liang, J., Li, H., Xu, F., Chen, J., Zhou, M., Yin, L., Zhai, Z., Chai, X., 2022. A fast deployable instance elimination segmentation algorithm based on watershed transform for dense cereal grain images. Agriculture 12 (9), 1486.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: IEEE International Conference on Computer Vision. pp. 2980–2988.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8759–8768.

Liu, Z., Wang, H., Weng, L., Yang, Y., 2016. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. IEEE Geosci. Remote Sens. Lett. 13 (8), 1074–1078.

Liu, Z., Yuan, L., Weng, L., Yang, Y., 2017. A high resolution optical satellite image dataset for ship recognition and some new baselines. In: International Conference on Pattern Recognition Applications and Methods. pp. 324–331.

Liu, S., Zhang, L., Lu, H., He, Y., 2021. Center-boundary dual attention for oriented object detection in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Mao, L., Zheng, Z., Meng, X., Zhou, Y., Zhao, P., Yang, Z., Long, Y., 2022. Large-scale automatic identification of urban vacant land using semantic segmentation of high-resolution remote sensing images. Landsc. Urban Plan. 222, 104384.

Papandreou, G., Chen, L.-C., Murphy, K.P., Yuille, A.L., 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: IEEE International Conference on Computer Vision. pp. 1742–1750.

Pont-Tuset, J., Arbelaez, P., Barron, J.T., Marques, F., Malik, J., 2016. Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (1), 128–140.

Qiu, L., Yu, D., Zhang, X., Zhang, C., 2024. Efficient remote-sensing segmentation with generative adversarial transformer. IEEE Geosci. Remote Sens. Lett. 1–5.

Rajchl, M., Lee, M.C., Oktay, O., Kamnitsas, K., Passerat-Palmbach, J., Bai, W., Damodaram, M., Rutherford, M.A., Hajnal, J.V., Kainz, B., et al., 2016. Deepcut: Object segmentation from bounding box annotations using convolutional neural networks. IEEE Trans. Med. Imaging 36 (2), 674–683.

Rodriguez-Serrano, J.A., Larlus, D., Dai, Z., 2016. Data-driven detection of prominent objects. IEEE Trans. Pattern Anal. Mach. Intell. 1969–1982.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241.

Rother, C., Kolmogorov, V., Blake, A., 2004. "GrabCut" interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23 (3), 309–314.

Satyawant, K., Abhishek, K., Dong-Gyu, L., 2024. RSSGLT: Remote sensing image segmentation network based on global–local transformer. IEEE Geosci. Remote Sens. Lett. 1–5.

Schuegraf, P., Schnell, J., Henry, C., Bittner, K., 2022. Building section instance segmentation with combined classical and deep learning methods. ISPRS Ann. Photogram. Remote Sens. Spat. Inf. Sci. 407–414.

Shi, S., Zhong, Y., Zhao, J., Lv, P., Liu, Y., Zhang, L., 2020. Land-use/land-cover change detection based on class-prior object-oriented conditional random field framework for high spatial resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Song, C., Huang, Y., Ouyang, W., Wang, L., 2019. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3136–3145.

Su, L., Xie, Q., Zhao, F., Cao, X., 2022. Change detection for multispectral images using modified semantic segmentation network. J. Appl. Remote Sens. 16 (1), 014518–014518.

Teng, Z., Duan, Y., Liu, Y., Zhang, B., Fan, J., 2021. Global to local: Clip-LSTM-based object detection from remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Tian, Z., Shen, C., Chen, H., 2020a. Conditional convolutions for instance segmentation. In: European Conference on Computer Vision. pp. 282–298.

Tian, Z., Shen, C., Chen, H., He, T., 2020b. Fcos: A simple and strong anchor-free object detector. IEEE Trans. Pattern Anal. Mach. Intell. 44 (4), 1922–1933.

Tian, Z., Shen, C., Wang, X., Chen, H., 2021. Boxinst: High-performance instance segmentation with box annotations. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5443–5452.

Venugopal, N., 2020. Automatic semantic segmentation with deeplab dilated learning network for change detection in remote sensing images. Neural Process. Lett. 51, 2355–2377.

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., Bai, X., 2019. Isaid: A large-scale dataset for instance segmentation in aerial images. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 28–37.

Xie, Y., Feng, D., Chen, H., Liu, Z., Mao, W., Zhu, J., Hu, Y., Baik, S.W., 2022. Damaged building detection from post-earthquake remote sensing imagery considering heterogeneity characteristics. IEEE Trans. Geosci. Remote Sens. 60, 1–17.

Xie, E., Sun, P., Song, X., Wang, W., Liu, X., Liang, D., Shen, C., Luo, P., 2020. Polarmask: Single shot instance segmentation with polar representation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 12193–12202.

Xu, Y., Fu, M., Wang, Q., Wang, Y., Chen, K., Xia, G.-S., Bai, X., 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. IEEE Trans. Pattern Anal. Mach. Intell. 43 (4), 1452–1459. http://dx.doi.org/10.1109/TPAMI.2020.2974745.

Yang, X., Zhou, Y., Zhang, G., Yang, J., Wang, W., Yan, J., Zhang, X., Tian, Q., 2022. The kfiou loss for rotated object detection. arXiv preprint arXiv:2201.12558.

Yue, M., Fu, G., Wu, M., Zhao, Y., Zhang, S., 2022. Vehicle motion segmentation via combining neural networks and geometric methods. Robot. Auton. Syst. 155, 104166.

Zhang, Z., Guo, W., Zhu, S., Yu, W., 2018. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. IEEE Geosci. Remote Sens. Lett. 15 (11), 1745–1749. http://dx.doi.org/10.1109/LGRS.2018.28569.

Zhang, C., Xiong, B., Li, X., Kuang, G., 2021a. Aspect-ratio-guided detection for oriented objects in remote sensing images. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

Zhang, T., Zhang, X., Zhu, P., Tang, X., Li, C., Jiao, L., Zhou, H., 2021b. Semantic attention and scale complementary network for instance segmentation in remote sensing images. IEEE Trans. Cybern. 52 (10), 10999–11013.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890.