Article

Meta-Knowledge Guided Weakly Supervised Instance Segmentation for Optical and SAR Image Interpretation

Man Chen [1,2], Yao Zhang [1], Enping Chen [2], Yahao Hu [1], Yifei Xie [1] and Zhisong Pan [1,*]

[1] College of Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China; 19205060770@stu.csust.edu.cn (M.C.); zhangyao@aeu.edu.cn (Y.Z.); huyahao@aeu.edu.cn (Y.H.); yifeixie@aeu.edu.cn (Y.X.)

[2] School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410114, China; 21205050959@stu.csust.edu.cn

* Correspondence: panzhisong@aeu.edu.cn

Abstract: The interpretation of optical and synthetic aperture radar (SAR) images in remote sensing is general for many tasks, such as environmental monitoring, marine management, and resource planning. Instance segmentation of optical and SAR images, which can simultaneously provide instance-level localization and pixel-level classification of objects of interest, is a crucial and challenging task in image interpretation. Considering that most current methods for instance segmentation of optical and SAR images rely on expensive pixel-level annotation, we develop a weakly supervised instance segmentation (WSIS) method to balance the visual processing requirements with the annotation cost. First, we decompose the prior knowledge of the mask-aware task in WSIS into three meta-knowledge components: fundamental knowledge, apparent knowledge, and detailed knowledge inspired by human visual perception habits of "whole to part" and "coarse to detailed." Then, a meta-knowledge-guided weakly supervised instance segmentation network (MGWI-Net) is proposed. In this network, the weakly supervised mask (WSM) head can instantiate both fundamental knowledge and apparent knowledge to perform mask awareness without any annotations at the pixel level. The network also includes a mask information awareness assist (MIAA) head, which can implicitly guide the network to learn detailed information about edges through the boundary-sensitive feature of the fully connected conditional random field (CRF), facilitating the instantiation of detailed knowledge. The experimental results show that the MGWI-Net can efficiently generate instance masks for optical and SAR images and achieve the approximate instance segmentation results of the fully supervised method with about one-eighth of the annotation production time. The model parameters and processing speed of our network are also competitive. This study can provide inexpensive and convenient technical support for applying and promoting instance segmentation methods for optical and SAR images.

Keywords: remote sensing; weakly supervised instance segmentation; meta-knowledge; optical images; synthetic aperture radar images

check for updates

Citation: Chen, M.; Zhang, Y.; Chen, E.; Hu, Y.; Xie, Y.; Pan, Z. Meta-Knowledge Guided Weakly Supervised Instance Segmentation for Optical and SAR Image Interpretation. *Remote Sens.* **2023**, *15*, 2357. https://doi.org/10.3390/rs15092357

Academic Editor: Dusan Gleich

Received: 26 February 2023
Revised: 23 April 2023
Accepted: 27 April 2023
Published: 29 April 2023

1. Introduction

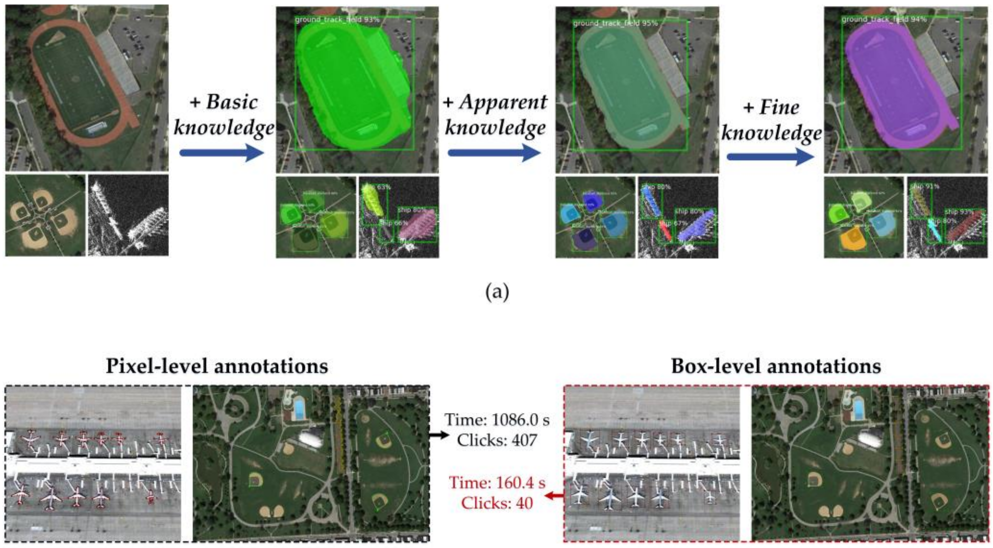
The acquisition methods of optical and synthetic aperture radar (SAR) images in remote sensing have been gradually diversified and improved in quality with the development of Earth observation technology. The interpretation of optical and SAR images can help relevant departments to effectively obtain valuable information, which can be beneficial for many tasks, including environmental monitoring [1,2], marine management [3,4], resource planning [5], and natural disaster damage analysis [6]. Therefore, research on optical and SAR image interpretation is of great practical importance.

Due to the natural advantage of deep convolutional neural networks (DCNN) in image processing, researchers have developed many DCNN methods for remote sensing image interpretation. According to their level of sophistication, these methods can be categorized as image-level [7–12], object-level [13–16], and pixel-level [17–27]. Image-level methods usually directly determine the categories of images but cannot perform specific analysis of the objects of interest. Object-level methods extract objects of interest in remote sensing images by training the network with bounding box annotations so that they can provide richer information than image-level methods. Pixel-level methods can be divided into semantic segmentation [17–20] and instance segmentation [21–27]. Semantic segmentation methods can segment the objects of interest in the image pixel by pixel but neglect to distinguish different instances within the same category. In contrast, instance segmentation methods can combine the advantages of object-level and semantic segmentation methods and simultaneously perform instance-level localization and pixel-level classification of objects of interest to interpret richer information. Thus, we focus on instance segmentation methods for optical and SAR images in remote sensing, aiming at pixel-level segmentation of objects of interest and distinguishing different instances within the same class.

Nowadays, many instance segmentation methods of optical and SAR images in remote sensing have emerged. In particular, HQ-ISNet [21] preserves the high-resolution features with the HR feature pyramid network (HRFPN) and optimizes the information flow between the head networks with the instance segmentation network version 2 (ISNetV2). CPISNet [22] proposes a new adaptive feature extraction network (AFEN) and improves the performance of the network with a cascaded architecture and an elaborated RoI extractor (ERoIE). Following the development of the attention mechanism, Zhang et al. [23] introduced a semantic attention module (SAM) into the optical image segmentation task to improve the network's response to objects and extend the mask to facilitate the integration of multi-scale information. Furthermore, SAR-CNN [24] applied the global attention module (GAM), semantic attention module (SAM), and anchor attention module (AAM) to the feature extraction, feature fusion, and object location of the SAR image instance segmentation task, respectively, making the network adaptable to complex backgrounds. Inspired by the advanced anchor-free detector, Shi et al. [25] proposed an anchor-free network for remote sensing image segmentation that can improve the initialization proposals and preserve the spatial detail information through the box refinement module (BRM) and saliency supplement module (SSM). GCBANet [26] can establish global dependencies through the global context information modeling module (GCIMM) and implement cross-scale box prediction through the boundary awareness box prediction module (BABPM). ESIP [27] extracts hierarchical feature maps in images through Swin Transformer and promotes information interaction between head networks by combining context information flow (CIF), which can improve the segmentation accuracy of optical and SAR images. Although the above methods have made some breakthroughs in the instance segmentation of optical and SAR images, they all rely on pixel-level annotations to train the network, which can come with expensive labor costs and pose an inconvenience to the practical application and extension of the methods.

To reduce the annotation cost of instance segmentation of optical and SAR images in remote sensing, we further focus on weakly supervised instance segmentation (WSIS) methods. According to the type of supervision information, existing WSIS methods can be divided into the image-level supervision paradigm [28–34] and the box-level supervision paradigm [35–40]. Although the image-level supervision paradigm can save extensive annotation costs, the image-level labels provide too little supervision information, limiting the accuracy of segmentation. Therefore, we do not adopt this paradigm for interpreting optical and SAR images. In contrast, box-level supervision paradigm methods can provide more supervision information with an acceptable amount of annotation. According to [41], the average time for each bounding box annotation is 10.2 s, while each pixel-level annotation takes 79 s, which is about eight times slower. In this study, we consider introducing the box-supervised paradigm into instance segmentation of optical and SAR images in remote sensing to balance the visual processing requirements and the annotation cost.

For the box-supervised paradigm instance segmentation task, some approaches [35–37] have attempted to mine prior knowledge from existing information and construct supervision information to guide network training. We suggest deconstructing prior knowledge in WSIS through the meta-knowledge theory and human visual perception habits of “whole to part” and “coarse to detailed”. Based on this, we decompose the prior knowledge of the mask-aware task in WSIS into three meta-knowledge components: fundamental knowledge, apparent knowledge, and detailed knowledge. Among these, fundamental knowledge mainly includes the positions and sizes of the objects, which can guide our model to locate and coarsely perceive objects and lay the foundation for refined mask awareness. Apparent knowledge corresponds to intuitive information, such as colors, which can further distinguish objects from the background. Detailed knowledge is the detailed information about the edges that enable perfect and detailed mask awareness and the generation of a high-quality instance mask. We also instantiate these meta-knowledge components to construct comprehensive supervisory information for our model, which can help our model perform mask awareness accurately and significantly reduce annotation costs (as shown in Figure 1).



(a)

(b)

Figure 1. (**a**) Effect of meta-knowledge components on guiding instance segmentation. The original objects are shown on the far left, followed by the results of instance segmentation with progressive addition of fundamental, apparent, and detailed knowledge. As can be seen, the segmentation effect gradually improves by adding the meta-knowledge components. (**b**) Comparison of annotation cost for pixel-level and box-level annotations. We quantify the annotation cost in terms of production time and workload (number of clicks). On the left and right are pixel-level and box-level annotations of images, respectively. Our experiments show that the annotation time of box-level annotations is about one-eighth of that of pixel-level annotations, as detailed in Section 4.5.

Specifically, we propose a meta-knowledge guided weakly supervised instance segmentation network (MGWI-Net) to implement instance segmentation of remote sensing images and reduce the reliance on expensive pixel-level annotations. This network includes a weakly supervised mask (WSM) head that can instantiate fundamental knowledge and apparent knowledge through the projection and color similarity loss functions. The projection loss function can guide our network to learn basic information and then realize the coarse perception of objects in remote sensing images. The color similarity loss function focuses on intuitive information and guides our network to distinguish objects from the background. In addition, the network contains a mask information awareness assisted (MIAA) head, which consists of an additional semantic branch and an auxiliary loss function. This head can implicitly guide the network to learn detailed information through the boundary-sensitive feature of the fully connected conditional random field (CRF) [42], which can instantiate detailed knowledge. Ultimately, these three meta-knowledge components can help our model accurately perceive mask information through simple box-level annotations. This study can lay the foundation for applying and promoting instance segmentation methods of optical and SAR images in remote sensing.

The contributions of this paper are summarized as follows:

- We thoroughly weigh the visual processing requirements and annotation costs, and then introduce the instance segmentation of the box-level supervised paradigm into the interpretation of optical and SAR images in remote sensing;
- Through the meta-knowledge theory and human visual perception habits, we decompose the prior knowledge of the mask-aware task in WSIS into three meta-knowledge components: fundamental knowledge, apparent knowledge, and detailed knowledge, which can provide a unified representation of the mask-aware task;
- By instantiating these meta-knowledge components, we propose the MGWI-Net. The WSM head in this network can instantiate both fundamental and epistemic knowledge to perform mask awareness without any annotations at the pixel level. The MIAA head can implicitly guide the network to learn detailed information through the boundary-sensitive feature of the fully connected CRF, enabling the instantiation of detailed knowledge;
- The experimental results show in the NWPU VHR-10 instance segmentation dataset and the SSDD dataset that the proposed three meta-knowledge components can guide the MGWI-Net to accurately segment the instance masks and ultimately achieve the approximate instance segmentation results of the fully supervised approach with about one-eighth of the annotation time.

2. Related Work

2.1. Instance Segmentation

Currently, mainstream instance segmentation methods can be divided into two main types: top-down [43–47] and bottom-up [48–52]. Top-down methods follow the detect-then-segment formulation. The most representative is mask R-CNN [43], which adds a mask head for faster R-CNN [53] via a fully convolutional network (FCN) [54] and then performs mask prediction. Thanks to the pioneering work of the mask R-CNN, some methods, such as the PANet [44], cascade mask R-CNN [45], HTC [46], and SCNet [47], have implemented various extensions to achieve performance improvements. Although top-down methods are intuitive in structure, the two-stage process slows down the segmentation speed.

Bottom-up methods aim to group the pixels of each instance directly and predict the corresponding semantic class, so that they are superior in segmentation speed. Specifically, YOLACT [48] generates a set of prototype masks, predicts the mask coefficients for each instance, and then combines the prototype masks with the mask coefficients to obtain the final instance masks. PolarMask [49] models the contours in a polar coordinate system and decomposes the instance segmentation task into an instance center classification problem and a dense distance regression problem. Unlike most methods that use a mask head with fixed weights, CondInst [50] introduces dynamic instance-aware mask heads that allow the network to adapt masks to instances, which is flexible and convenient. SOLO [51] can assign pixel categories based on the location and size of instances, transforming the instance segmentation problem into a pixel-wise classification problem. SOLOv2 [52] extends SOLO with mask kernel prediction, feature learning, and matrix non-maximum suppression (matrix NMS) to achieve better segmentation results. Compared to top-down instance segmentation methods, such methods can better balance accuracy and speed. However, both types rely on expensive pixel-level annotations, which limits their application and dissemination. Therefore, we focus on WSIS methods to reduce the annotation cost of instance segmentation of optical and SAR images.

2.2. Weakly Supervised Instance Segmentation

As fully supervised instance segmentation methods require expensive pixel-level annotations, WSIS has received increasing attention from researchers. Existing WSIS methods can be divided into two paradigms according to the annotation type: the image-level supervision paradigm [28–34] and box-level supervision paradigm [35–40]. Image-level supervision paradigm methods mainly rely on segmentation proposals [28,29], target ranking [30], or pseudo-annotation generation [31–34] to mine the instance masks from images. Considering that the peaks in the class response map have a strong correlation with object instances, PRM [30] guides neural network learning with the class peak response so that it has the capability of instance segmentation. Label-PENet [33] includes a multi-label classification module, an object detection module, an instance refinement module, and an instance segmentation module, and then implements the mining and optimization of semantic information by cascading these modules to obtain pseudo-labels with high accuracy. LACI [34] mainly contains a conditional network for generating pseudo-annotations, allowing pseudo-label consistency between objects of the same category. Although this paradigm can significantly reduce annotation costs, image-level labels provide too little supervisory information and limited segmentation accuracy.

In contrast, the box-supervision paradigm can provide more supervised information with an acceptable amount of annotation, making it easier to balance the visual processing requirements and annotation costs. BBTP [35] considers the instance segmentation task of the box-supervised paradigm as a multi-instance learning process. It first treats the rows or columns inside the box as positive bags and those outside the box as negative bags, and then sets the condition that the maximum network output inside the positive bags converges to 1 and the maximum network output inside the negative bags converges to 0 to achieve instance segmentation. Hao et al. [37] proposed a contour prior that can guide the network to accurately distinguish objects from backgrounds by maximizing the inner product of the gradients between the mask proposal and the corresponding image region. DiscoBox [38] combines WSIS, object detection, and semantic relevance mining to exploit multi-level structured knowledge fully and then introduces self-supervised learning constructed from the consistency constraint between the degraded and original models so that the model's ability to perceive masks improves. BoxCaseg [39] defines the WSIS problem of the box-supervised paradigm as a category-independent object segmentation problem based on bounding boxes and integrates box-supervised information with saliency detection to provide good segmentation results.

3. Methodology

In this section, we first provide a detailed description of the meta-knowledge in WSIS. Then, we introduce the proposed MGWI-Net, focusing on this network's schemes of meta-knowledge instantiation.

3.1. Meta-Knowledge

The term meta-knowledge originates from meta-learning, the goal of which is to train a model that can adapt to a new task or scenario with few data or training iterations [55]. In general, meta-knowledge can be the basis for the design of critical components of the model, such as initial parameters, network architecture, supervisory information, or optimization strategies [56,57]. In WSIS, the lack of pixel-level annotations may lead to limited supervisory information and create difficulties for the mask-aware task. Some methods [35–37] have been proposed to attempt to mine prior knowledge from existing information and construct supervised information via prior knowledge to guide the network for training. As shown in Figure 2, we combine the meta-knowledge theory and human visual perception habits to decompose the prior knowledge of the mask-aware task in WSIS into three meta-knowledge components. This can provide a unified representation of prior knowledge in line with human visual perception habits. Furthermore, we instantiate these meta-knowledge components to construct efficient supervised information for the model, allowing adaptation from simple box-level annotation to complex instance segmentation.

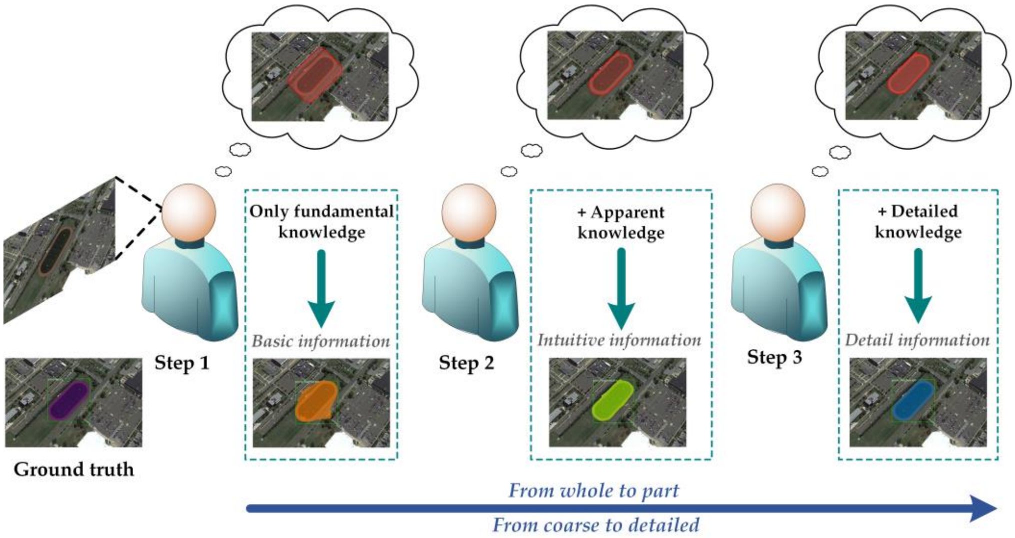


Figure 2. Human visual perception habits and meta-knowledge. We summarize human visual perception habits as "whole to part" and "coarse to detailed" processes and decompose the prior knowledge in WSIS into three meta-knowledge components: fundamental, apparent, and detailed knowledge. These components can guide the network to perceive basic, intuitive, and detailed information, respectively, and eventually adapt from simple box-supervised annotation to complex instance segmentation.

As shown in Figure 2, human visual perception habits can be summarized as a "whole to part" and "coarse to detailed" process. Accordingly, we decompose prior knowledge (K) in WSIS into three meta-knowledge components: fundamental knowledge (K_f), apparent knowledge (K_a), and detailed knowledge (K_d), which can be described via

$$K = \left\{K_f, K_a, K_d\right\} \tag{1}$$

In the following, we will provide a detailed description of the following three components:

- Fundamental knowledge, K_f: In visual perception, people tend first to acquire basic information, such as the position and size of an object, which is indispensable for refined observation later. Therefore, we define the fundamental knowledge of mask-aware task in WSIS as the position and size of the object, aiming to help the model locate and roughly perceive the object and lay the foundation for subsequently refined mask awareness;
- Apparent knowledge, K_a: According to visual perception habits, people naturally focus on objects' intuitive information (such as color) after acquiring basic information and using this intuitive information to distinguish objects from the background. Therefore, we define the apparent knowledge of the mask-aware task as the color and instantiate this knowledge as a specific loss function in the MGWI-Net to implement the preliminary extraction of masks. See Section 3.2.2 for details;
- Detailed knowledge, K_d: As shown in Step 3 of Figure 2, after the successful use of the fundamental information and the intuitive information, it is necessary to pay attention to the detailed information, such as the edge of the object, and then achieve perfect and detailed mask awareness. Based on this, we define the detailed knowledge as the edge and instantiate this knowledge with an additional task branch and a new loss function in the MGWI-Net, so that this model can accurately perceive the mask.

It is worth noting that conventional work typically integrates meta-knowledge into the network training process to accomplish model adaptation to unknown tasks or scenarios through a bilevel or multi-level optimization procedure [58,59]. In contrast to these complex optimization strategies, we directly instantiate meta-knowledge as supervised information to guide model learning explicitly. Although this approach increases the complexity of the supervised information, it allows the model to adapt from simple box-level annotation to complex instance segmentation without the bilevel or multi-level optimization procedure.

3.2. Meta-Knowledge-Guided Weakly Supervised Instance Segmentation Network

3.2.1. Overview

Figure 3 illustrates the overall framework of the MGWI-Net. The network belongs to an efficient bottom-up method and mainly consists of a feature extraction network and some head networks, the core components of which are the WSM head and the MAA head in the head networks. The WSM head instantiates fundamental and apparent knowledge through the projection and color similarity loss functions, respectively. The projection loss function enables the network to learn basic information, such as the positions and sizes of the objects without pixel-level annotations, achieving coarse awareness of objects in remote sensing images. The color similarity loss function focuses on intuitive information and guides the network to distinguish the objects from the background. The MIAA head provides an additional semantic branch and awareness-assisted loss function to instantiate detailed knowledge, which can implicitly guide the network to learn detailed information about edges by incorporating the boundary-sensitive feature of the fully connected CRF. The instantiation of the above meta-knowledge enables the network to have accurate mask-aware capabilities without pixel-level annotations.

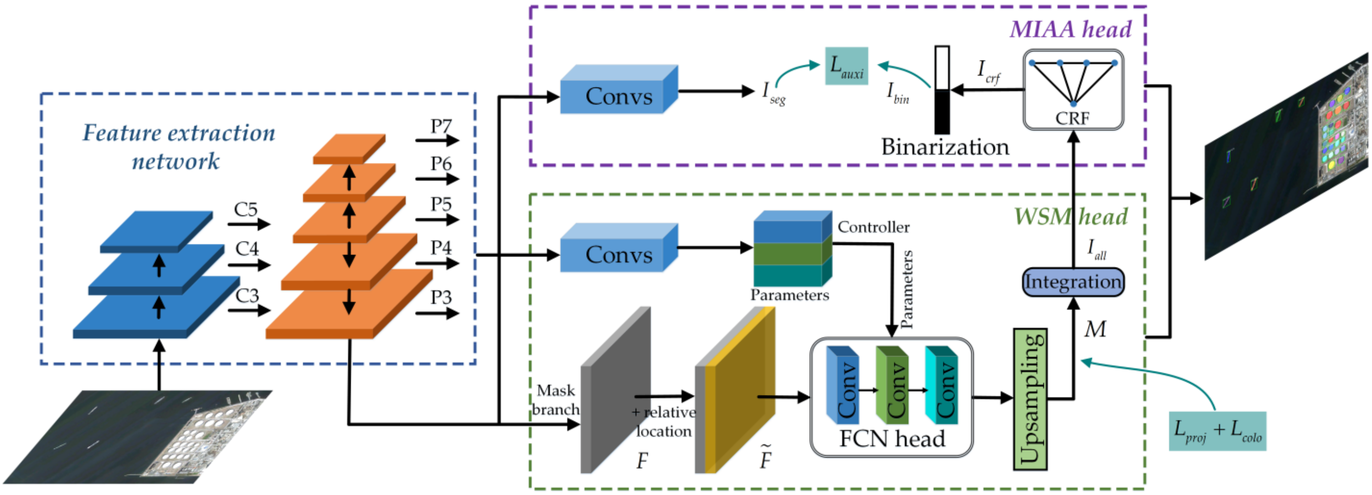


Figure 3. The overall framework of MGWI-Net. We regard the mask-aware task in WSIS as a task adaptation problem and scientifically decompose a priori knowledge utilizing the meta-knowledge theory and human visual perception habits, resulting in three meta-knowledge components: fundamental, apparent and detailed knowledge. The WSM head and the MIAA head in the MGWI-Net can instantiate these three meta-knowledge components, allowing adaptation from simple box-level annotation to complex instance segmentation. The WSM head instantiates fundamental and apparent knowledge to make the network achieve mask-aware capabilities without pixel-level annotations. The MIAA head can implicitly guide the network to learn detailed information, such as information about edges, in conjunction with the boundary-sensitive feature of the fully connected CRF, which instantiates detailed knowledge. The MGWI-Net head also includes a classification head, center-ness head, and box head, which are not shown in the figure for simplicity.

3.2.2. Weakly Supervised Mask Head

The WSM head contains two main loss functions: the projection loss function and the color similarity loss function, which could instantiate fundamental knowledge (K_f) and apparent knowledge (K_a) for providing supervisory information to the network.

Projection loss function: As described in Section 3.1, the preliminary work in the mask-aware task is to determine the positions and sizes of the objects according to human visual perception habits. Fortunately, the position of an object's mask can be obtained directly from its predicted bounding box since its bounding box usually surrounds the corresponding mask. On the other hand, the projection of an object's mask in one direction (horizontal or vertical) should be the same as the projection of this object's bounding box in the same direction [36]. Therefore, the projection loss function can be designed based on this feature so that the minimum bounding rectangle of the predicted mask matches the corresponding bounding box to determine the sizes in the basic information.

As shown in Figure 4, in the whole image, the supervisory information generated by the annotation box is denoted as $B \in \mathbb{R}^{W \times H}$, and the mask prediction result is $M \in \mathbb{R}^{W \times H}$, where W is the image's width, and H is the height. Then, we can extract the projections of the annotation box in the horizontal and vertical directions from the label annotation information and express them as B_{hor} and B_{ver}. The projections of the mask prediction result from M in the horizontal and vertical directions are represented as M_{hor} and M_{ver}, which can be expressed as follows:

$$M_{hor} = \max_{h} M(w, h), \tag{2}$$

$$M_{ver} = \max_{w} M(w, h), \tag{3}$$

where $w = 1, 2, \cdots, W$ and $h = 1, 2, \cdots, H$.

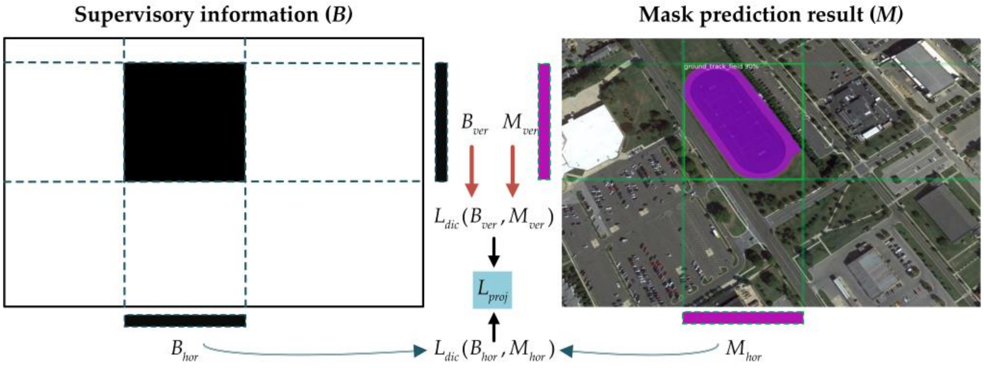

Supervisory information (B)
Mask prediction result (M)
B_{ver}
M_{ver}
$L_{dic}(B_{ver}, M_{ver})$
L_{proj}
B_{hor}
$L_{dic}(B_{hor}, M_{hor})$
M_{hor}
ground_track_field 70%

Figure 4. Schematic diagram of the projection loss function. The projection loss function consists of constraints in horizontal and vertical directions that can make the minimum bounding rectangle of the predicted mask match the corresponding annotation box. In this way, the MGWI-Net can learn the sizes in the basic information without pixel-level annotations and achieve coarse awareness of the object through this loss function.

Next, the constraint in the horizontal direction can be constructed by minimizing the difference between the annotation box projection (B_{hor}) and the prediction mask projection (M_{hor}). Similarly, the constraint in the vertical direction can be built with B_{ver} and M_{ver}. These two directional constraints can form the projection loss function (L_{proj}) together, as shown in Equation (4). We use the dice loss (L_{dic}) [60] to reduce the effect of sample imbalance in remote sensing images.

$$\begin{aligned} L_{proj} &= L_{dic}(B_{hor}, M_{hor}) + L_{dic}(B_{ver}, M_{ver}) \\ &= 2 - \frac{2|B_{hor} \cap M_{hor}|}{|B_{hor}|^2 + |M_{hor}|^2 + \varepsilon} - \frac{2|B_{ver} \cap M_{ver}|}{|B_{ver}|^2 + |M_{ver}|^2 + \varepsilon} \end{aligned} \tag{4}$$

where ε is the hyperparameter and takes the value of 1×10^{-5} for the stability of the denominator.

During training, the projection loss function (L_{proj}) applies to all instances in the image. By minimizing the above projection loss function (L_{proj}), the minimum bounding rectangle of the predicted mask can match the corresponding bounding box. Thus, the network could learn fundamental knowledge without pixel-level annotations for coarse mask awareness. However, instantiating only fundamental knowledge may lead to multiple mask prediction results with identical minimum bounding rectangles, so it is necessary to instantiate apparent knowledge and detailed knowledge to form a complete network guidance system.

Color similarity loss function: The projection loss function can guide the model to learn some basic information and to realize coarse mask awareness. However, the mask-aware task needs to further distinguish objects from the background. In general, for most pixel locations in optical and SAR images, if the color similarity between two-pixel locations in the vicinity of the location is high, there is a high probability that they belong to the same object (or background) [40]. The color similarity loss function can be designed based on the abovementioned color relationship to instantiate apparent knowledge (K_a) and guide the network to further separate the objects from the background.

First, for each pixel position in the image (I), a neighborhood consisting of eight different pixel positions is defined, as shown in Figure 5. Then, neighborhood partitioning is extended to all pixel positions in the image (I) to obtain eight neighborhood images (I^i_{nei}), where $i = 1, 2, \cdots, 8$ and the color similarity for I and I^i_{nei} is given by

$$S^i(w,h) = \exp(-\frac{1}{\sigma}\left|I(w,h) - I^i_{nei}(w,h)\right|) \tag{5}$$

where σ is the hyperparameter. In order to make the two-pixel positions with significant differences in the image have high color similarity and effectively reduce the possibility of misclassifying two-pixel positions belonging to the same object as different objects, we set the hyperparameter σ to a conservative value of 2. Then, the similarity threshold (θ) is set. If the color similarity between I and I^i_{nei} at (w,h) is over the θ, they are considered to belong to the same category; otherwise, no judgment is made. Therefore, the probability ($P^i_I(w,h)$) that the image (I) and one of its corresponding neighborhood images (I^i_{nei}) have the same category at pixel location (w,h) is 1 if $S^i(w,h) \geq \theta$.

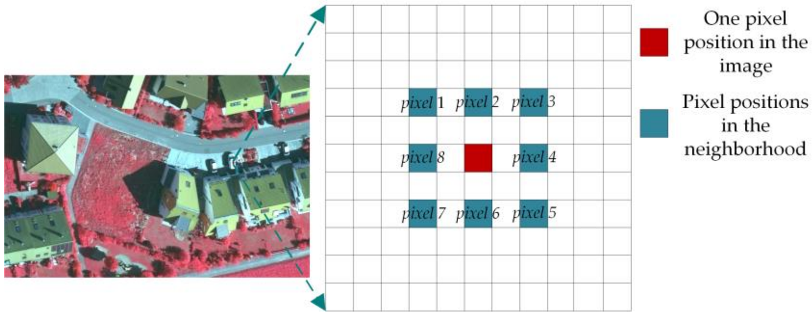

pixel 1
pixel 2
pixel 3
pixel 8
pixel 4
pixel 7
pixel 6
pixel 5
🟥 One pixel position in the image
🟦 Pixel positions in the neighborhood

Figure 5. Schematic representation of the neighborhood of a pixel location. For every pixel position in a remote sensing image, we delineate a neighborhood of 8 different pixel positions and calculate the color similarity from this.

Similarly, for the mask prediction result (M) of the image (I), we also construct the neighborhood of eight different pixel positions and then obtain eight neighborhood images (M^i_{nei}). Next, the probability ($P^i_M(w,h)$) that the mask prediction result (M) and one of its neighborhood images (M^i_{nei}) have the same class at the pixel position (w,h) can be defined as follows:

$$\begin{aligned} P^i_M(w,h) = {} & M(w,h) \cdot M^i_{nei}(w,h) \\ & +(1 - M(w,h)) \cdot (1 - M^i_{nei}(w,h)) \end{aligned} \tag{6}$$

Finally, the color similarity loss function can be formulated by minimizing the difference between $P^i_I(w,h)$ and $P^i_M(w,h)$, which guides the network to perceive intuitive information about colors and instantiate apparent knowledge. Additionally, considering the natural tendency of humans to focus on pixel positions near objects during the stage of distinguishing the objects from the background in visual perception, we similarly restrict the calculation of the color similarity loss function within the boxes. This constraint prevents the calculation process from overly relying on the pixel positions of the background, reflecting the gradual relationship between the position information in fundamental knowledge and the instantiation process of apparent knowledge. Specifically, the color similarity loss function (L_{colo}) can be defined as follows:

$$L_{colo} = -\frac{1}{N}\sum\nolimits_{i}^{8}\sum\nolimits_{(w,h)}^{B_{inte}} P_{I}^{i}(w,h)\log(P_{M}^{i}(w,h)) \tag{7}$$

where N is the total number of pixels in the annotation box with high visual similarity. B_{inte} denotes the pixel position in the annotation box, which can reduce the dependence of the gradient descent on the pixel positions of the background.

Network: Regarding the network structure, the WSM head inherits from the dynamic instance-aware mask head in CondInst [50], which consists of a mask branch, a controller, and a FCN head. The mask branch can process the feature maps from the feature extraction network and generate the input to the FCN head after the fusion of relative position information. The FCN head corresponds to the instances in the image one by one through network parameters dynamically generated by the controller so that the network focuses only on its corresponding instance region during segmentation, ignoring the influence of other instances.

In conjunction with the above structure, the mask prediction steps of the WSM head can be briefly summarized. First, the output of P_3 of the FPN is processed through the mask branch to generate a feature map (F). Then, the F is combined with the coordinate map, generating a feature map ($\widetilde{F}$) containing relative position information. Finally, mask predictions are performed separately by the FPN head with different network parameters and combined with upsampling to improve the prediction resolution.

In the training phase, the supervised information of the WSM head is formed by the projection loss function (L_{proj}) and the color similarity loss function (L_{colo}) together. They can instantiate fundamental and apparent knowledge and guide the network to perceive the mask information without pixel-level annotations. In the next section, we further instantiate detailed knowledge and guide the network to accurately instance segmentation through detailed information.

3.2.3. Mask Information Awareness Assist Head

The two loss functions in the WSM head can instantiate fundamental knowledge and apparent knowledge and guide the network to perform primary mask awareness. However, the lack of fine-grained mask annotation may lead to the network's poor ability to perceive detailed information about edges, which is not conducive to achieving high-quality segmentation. Therefore, we introduce detailed knowledge to make the network pay more attention to the detailed information of the object so that the network can realize perfect and detailed mask awareness. Specifically, we design an MIAA head that provides an additional semantic branch and an awareness-assisted loss function that can implicitly guide the network to learn detailed information about edges by incorporating the boundary-sensitive feature of the fully connected CRF [42]. This head can instantiate detailed knowledge and help the MGWI-Net perceive the mask accurately.

As the WSM head can be applied to the input feature map (i.e., P3 from FPN) the same number of times as the number of instances in the image to predict instance masks separately for each instance in the image, we should first integrate the mask prediction results (M) from the WSM head to obtain the overall mask prediction image (I_{all}) of the image (I), as shown in Figure 3. Then, the overall mask prediction image (I_{all}) is treated as a unary term, and the color and pixel location differences with the bilateral kernel in the image (I) are constructed to obtain a pairwise term. Therefore, the unsupervised fully connected CRF is established, which uses the mean-field approximation for optimization. Next, I_{all} is processed by the unsupervised fully connected CRF to produce the image

(I_{crf}), the instance masks of which contain detailed information about edges. Further, we binarize I_{crf} to generate the final binarized image (I_{bin}). Finally, the binarized image (I_{bin}) is directly taken as the ground truth. The awareness-assisted loss function (L_{auxi}) can be built though the semantic branch's output (I_{seg}) and the binarized image (I_{bin}), which can realize the mask information's backpropagation and promote the instantiation of detailed knowledge. Considering the sample imbalance factor in remote sensing images, we construct the awareness-assisted loss function (L_{auxi}) based on the focal loss (L_{foc}) [61]. Thus, L_{auxi} can be written as

$$L_{auxi} = L_{foc}(I_{seg}, I_{bin}) \tag{8}$$

During network training, the above MIAA head designed with the boundary-sensitive feature of the fully connected CRF is first able to implicitly add detailed information about edges to the output of the WSM head to enrich the mask prediction results. We then monitor the semantic branch through the awareness-assisted loss function (L_{auxi}), which enables the backpropagation of fine-grained pixel-level mask information. It is worth noting that the I_{bin} in the MIAA head is obtained by further processing the output masks from the WSM head, which is strongly influenced by the instantiation components of fundamental and apparent knowledge. Therefore, the instantiation component of detailed knowledge exhibits a gradual relationship with both fundamental and apparent knowledge, similarly to the human visual perception process in which the perception of object details requires a coarse perception foundation. Ultimately, even without fine-grained pixel-level annotations, the network can accurately perceive detailed information about edges and instantiate detailed knowledge.

3.2.4. Total Loss Function

The WSM head instantiates fundamental knowledge (K_f) and apparent knowledge (K_a) with the projection loss function (L_{proj}) and the color similarity loss function (L_{colo}), respectively. Through the boundary-sensitive feature of the fully connected CRF, the semantic branch and the awareness-assisted loss function (L_{auxi}) in the MIAA head allow the instantiation of detailed knowledge (K_d). In addition, the total loss function of the MGWI-Net also contains loss functions of the classification head, the box head, and the center-ness head, which are inherited from FCOS [62] and here centrally expressed as the object detection loss function (L_{dete}). Finally, the total loss function of the MGWI-Net can be formulated as a weighted combination of the above loss functions and can be described as follows:

$$L = \lambda_1 L_{dete} + \lambda_2 (L_{proj} + L_{colo}) + \lambda_3 L_{auxi} \tag{9}$$

where λ_1 and λ_2 are the weights of the object detection loss function (L_{dete}) and the mask-aware loss function (composed of two loss functions, L_{proj} and L_{colo}, in the WSM head), respectively, and their settings follow the weight-setting method in CondInst [50], both being set to 1. On the other hand, compared with the object detection and mask-aware tasks, the mask information awareness assistance task plays an auxiliary role in the MGWI-Net, and its importance is relatively low. Considering that the computation of L_{auxi} heavily relies on the output of the WSM head, if the quality of the masks and predicted boxes output by the WSM head is poor during the network learning process, the poor information will be propagated to the MIAA head and lead to poor learning, so it is necessary to control the weight of L_{auxi} to highlight the two essential tasks of detection and mask-awareness. Given the above considerations and the ablation study in Section 4.7, we set the weight parameter λ_3 to 0.1, which is one order of magnitude lower than the weight parameters λ_1 and λ_2, to balance each task.

4. Experiment and Analysis

4.1. Datasets

NWPU VHR-10 instance segmentation dataset: This dataset [21] is an extension of the vanilla NWPU VHR-10 dataset [63,64] with pixel-level annotations and has a total of 800 optical images. It consists of 650 images with objects and 150 images with a pure background, ranging in size from 533 × 597 to 1728 × 1028 pixels. It contains a total of 10 categories, namely airplane (AI), baseball diamond (BD), ground track field (GTF), vehicle (VC), ship (SH), tennis court (TC), harbor (HB), storage tank (ST), basketball court (BC) and bridge (BR). In the experiments, we randomly selected 70% of the images with objects (i.e., 454 images) as the training set and 30% (i.e., 196 images) as the test set. It should be noted that the MGWI-Net uses pixel-level annotations only during the test phase.

SSDD dataset: This dataset [21] consisted of 1160 SAR images with spatial resolutions ranging from 1 to 10 m and covering various polarization methods. The experiments randomly assigned the training and test sets in a 7:3 ratio. The training set consisted of 812 images and the test set of 348. Similarly, while the test phase showed the effect of instance segmentation with pixel-level annotations, the training process in the MGWI-Net did not use pixel-level annotations.

4.2. Implementation Details

The GPU used in the experiments was NVIDIA Tesla V100. The MGWI-Net consists of a feature extraction network and several head networks. The feature extraction network includes the popular ResNet101 [65] and FPN [66]. ResNet101 consists of 100 convolutional layers and a global average pooling layer, extracting feature maps from optical or SAR images. Its specific structure can be found in reference [65]. During the training process of MGWI-Net, ResNet101 is initialized with pre-trained weights from ImageNet. FPN can fuse feature maps of different scales, providing input for the subsequent head networks. Its specific structure can be found in reference [66]. The head networks include a classification head, center-ness head, box head, WSM head, and MIAA head, the specific structures of which are shown in Table 1. The MGWI-Net was optimized by a stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 1×10^{-4}. The base learning rate was 2.5×10^{-3}, which was reduced by a factor of 10 at the 8th and 11th epochs. Due to the considerable variation in object size in remote sensing images, the short edges of the input images were randomly resized to values between 600, 700, 800, 900, and 1000, and the long edges were uniformly set to 1333. The similarity threshold (θ) in the color similarity loss function of the WSM head was set to 0.3. We removed the fully connected CRF at the eighth epoch to prevent the network from overfitting the results of the fully connected CRF processing. In the training phase of the MGWI-Net, we only used box-level annotations. In the testing phase, we verified the segmentation effect of the network by pixel-level annotations.

Table 1. The specific structure of head networks in MGWI-Net. 'Conv3 × 3' and 'Conv1 × 1' denote convolutional layers with the kernel sizes 3 × 3 and 1 × 1, respectively. '×*n*' indicates the number of stacked convolutional layers.

Head Network	Sub-Component	Structure
Classification head	-	(Conv3 × 3) × 4
Center-ness head	-	(Conv3 × 3) × 4
Box head	-	(Conv3 × 3) × 4
WSM head	Mask branch	(Conv3×3) × 4, Conv1 × 1
	Controller	Conv3 × 3
	FCN head	(Conv1 × 1) × 3
MIAA head	Semantic branch	(Conv3 × 3) × 2

4.3. Evaluation Metrics

In this study, the MS COCO [67] dataset evaluation metrics are used to comprehensively and quantitatively evaluate the instance segmentation methods for optical and SAR images in remote sensing. For the mask prediction result (P_{mask}), its intersection over union (IoU) with the ground truth (G_{mask}) is defined as N_{cli}.

$$IoU(P_{mask}, G_{mask}) = \frac{P_{mask} \cap G_{mask}}{P_{mask} \cup G_{mask}}. \tag{10}$$

Based on the IoU, the instance segmentation results can be classified as true positive (TP), false positive (FP), true negative (TN), and false negative (FN). Accuracy and recall are calculated as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

$$Recall = \frac{TP}{TP + FN}. \tag{12}$$

The average precision (AP) for a given IoU threshold can be obtained by

$$AP_{IoU} = \int_0^1 P(r)dr, \tag{13}$$

where $P(r)$ is the precision corresponding to the recall and r is the recall. Thus, AP_{50} and AP_{75} represent the results of the above equations for the IoU thresholds of 0.5 and 0.75, respectively. AP is the average of IoU for 10 thresholds (0.50:0.05:0.95) and can be written as:

$$AP = \frac{1}{10}\sum\nolimits_{IoU=0.5}^{0.95} AP_{IoU}. \tag{14}$$

In addition, the evaluation metrics include AP_S, AP_M and AP_L for small (<32^2-pixel), medium (>32^2-pixel and <96^2-pixel), and large (>96^2-pixel) objects, respectively, which can measure the effectiveness of the mask prediction in all aspects. Finally, because of the general lack of quantitative studies on label production costs in remote sensing, we designed two metrics to compare the label production costs of fully supervised and weakly supervised instance segmentation methods together, namely the annotation time (T_{ann}) and annotation workload (N_{cli}), as detailed in Section 4.5.

4.4. Impact of Meta-Knowledge in MGWI-Net

In this study, we decomposed the prior knowledge in WSIS through the theory of meta-knowledge and the habits of human visual perception into fundamental knowledge (K_f), apparent knowledge (K_a), and detailed knowledge (K_d). Furthermore, we instantiated these meta-knowledge components to construct supervisory information for the model, allowing adaptation from simple box-level annotations to complex instance segmentation. We explored the impact of these three components of meta-knowledge on the NWPU VHR-10 instance segmentation dataset and the SSDD dataset and recorded them in Tables 2–4. We also show the results of the qualitative experiments in Figures 6 and 7.

Table 2. Impact of meta-knowledge on the NWPU VHR-10 instance segmentation dataset.

Fundamental Knowledge	Apparent Knowledge	Detailed Knowledge	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
√			29.8	62.9	25.8	14.7	27.4	39.0
√	√		49.8	80.7	51.0	35.2	46.1	57.4
√	√	√	51.6	81.3	53.3	37.6	48.2	59.1

Table 3. Class-wise impact of meta-knowledge on the NWPU VHR-10 instance segmentation dataset.

Fundamental Knowledge	Apparent Knowledge	Detailed Knowledge	AI	BD	GTF	VC	SH	TC	HB	ST	BC	BR	*AP*
√			0.2	47.8	66.4	16.5	8.1	35.3	11.4	59.4	45.5	7.9	29.8
√	√		15.7	77.0	90.4	40.8	45.9	68.5	12.1	76.8	63.1	8.1	49.8
√	√	√	17.0	77.3	91.9	41.0	50.8	71.2	15.7	76.5	64.6	10.9	51.6

Table 4. Impact of meta-knowledge on the SSDD dataset. The small number of large objects (>96^2 pixels) in the SSDD dataset makes the AP_L subject to considerable uncertainty, so we do not report this metric here.

Fundamental Knowledge	Apparent Knowledge	Detailed Knowledge	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
√			38.1	88.5	21.8	39.8	42.7	-
√	√		51.8	91.9	54.0	52.7	53.2	-
√	√	√	53.0	92.4	57.1	53.7	54.9	-

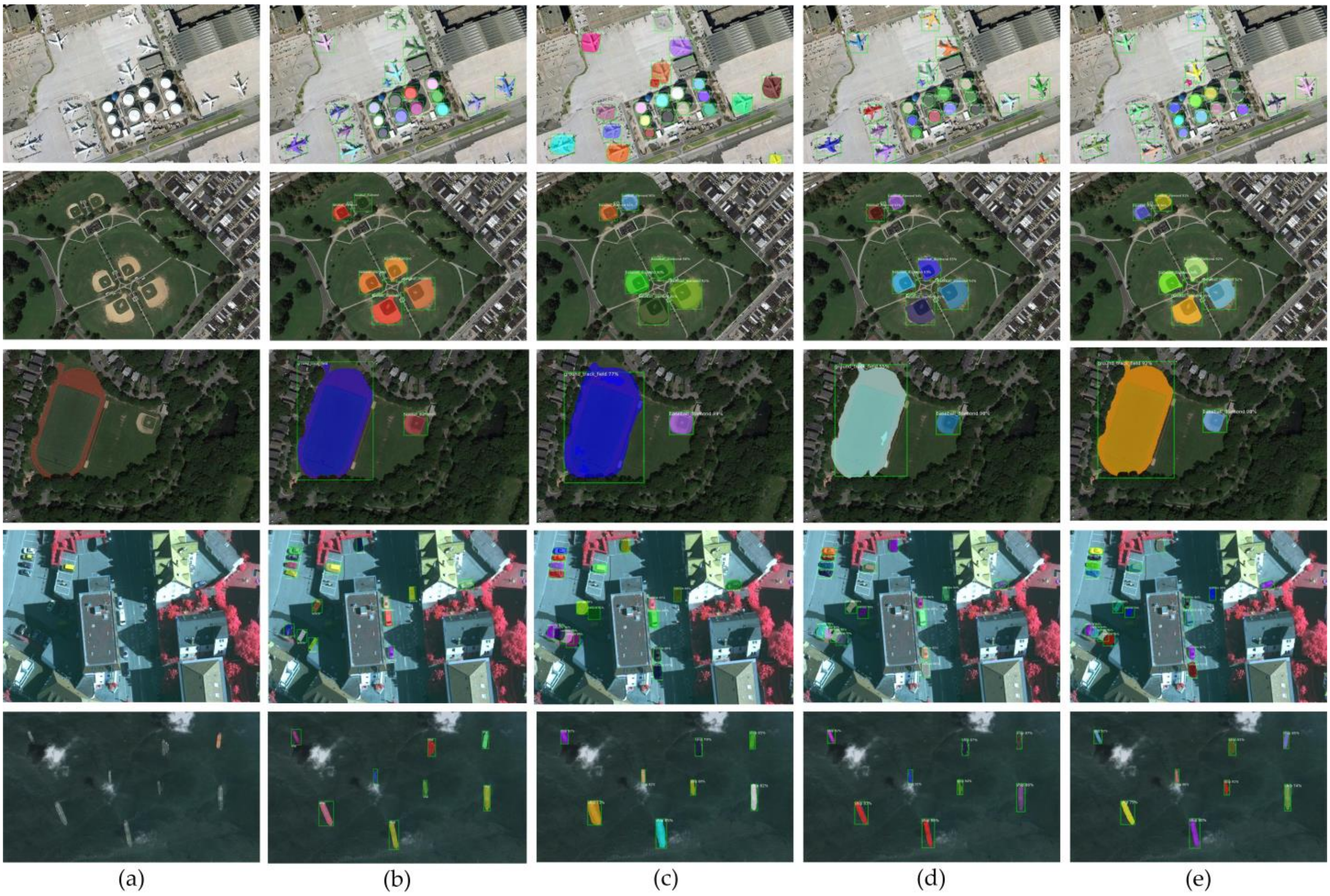


Figure 6. Qualitative experimental results of the impact of meta-knowledge on the NWPU VHR-10 instance segmentation dataset. (**a**,**b**) The original images and the ground truths, respectively. (**c**) The segmentation results guided by fundamental knowledge only. (**d**) The instance segmentation results guided by both fundamental and apparent knowledge. (**e**) The segmentation results for MGWI-Net guided by the three meta-knowledge components together.

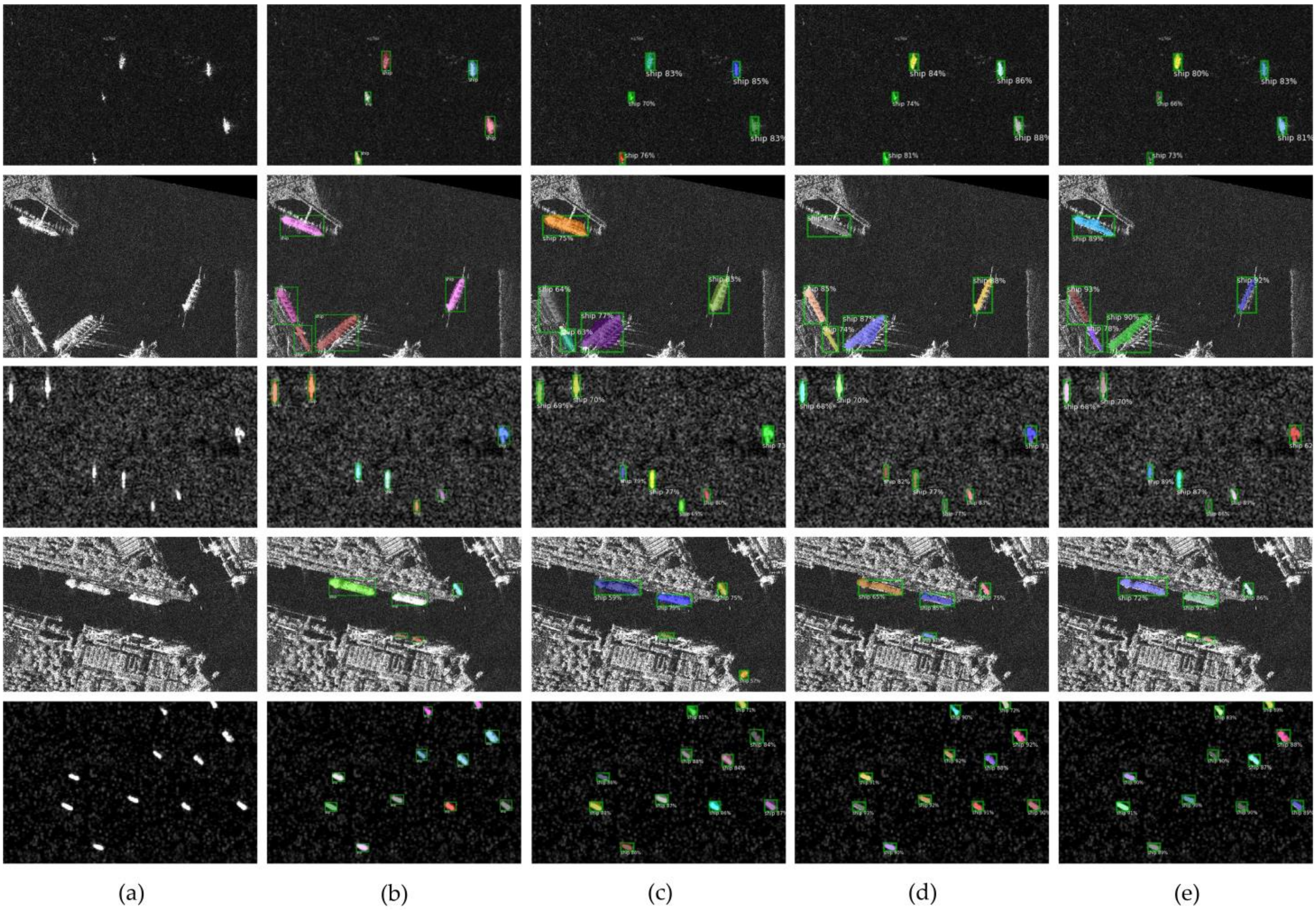


Figure 7. Qualitative experimental results of the impact of meta-knowledge on the SSDD dataset. (**a**,**b**) The original images and the ground truths in the SSDD dataset. (**c**,**d**,**e**) The segmentation results guided by only fundamental, fundamental and apparent knowledge, and the three meta-knowledge components together.

Results of experiments on the NWPU VHR-10 instance segmentation dataset: Table 2 shows the effect of meta-knowledge on the NWPU VHR-10 instance segmentation dataset. The *AP* is 29.8 when the MGWI-Net uses only the projection loss function, indicating that fundamental knowledge is sufficient for coarse segmentation. The addition of the color similarity loss function can increase the *AP*, AP_{50}, and AP_{75} by 20.0, 17.8, and 25.2, respectively, showing that the instantiation of apparent knowledge can significantly improve the segmentation quality. Furthermore, all segmentation metrics are optimized by introducing of the MIAA head, as the instantiation of detailed knowledge allows the network to pay more attention to the detailed information. Table 3 reports the meta-knowledge's impact on each category's segmentation effectiveness in the NWPU VHR-10 instance segmentation dataset. The network has mediocre segmentation effectiveness for most categories when only the fundamental knowledge instantiation component is available. With the introduction of the color similarity loss function, the effectiveness of the segmentation is refined for all categories. It further illustrates the crucial role of apparent knowledge in mask awareness. After adding the MIAA head, higher performance is obtained for all categories except the storage tank (ST), for which there is a slight decrease (0.3 *AP*), demonstrating the role of detailed knowledge in improving segmentation performance. Overall, instantiations of the three meta-knowledge components can form a complete network guidance system, allowing adaptation from simple box-level annotation to a highly sophisticated instance segmentation task.

The results of the qualitative experiment on the NWPU VHR-10 instance segmentation dataset are shown in Figure 6. When the network is guided by fundamental knowledge, the network has only a coarse mask awareness capability, and there is strong over-segmentation. After introducing apparent knowledge, the segmentation results of the MGWI-Net are improved and can distinguish the objects from the background. With the addition of the MIAA head, the network can generate refined masks and achieve higher-quality instance segmentation. In general, by instantiating fundamental, apparent, and detailed knowledge in the absence of pixel-level annotations, the MGWI-Net can accomplish the complex mask-aware task.

Results of experiments on the SSDD dataset: We also verified the effect of meta-knowledge on instance segmentation in SAR images. As shown in Table 4, the MGWI-Net achieves an AP of 38.1 when our network is guided by the projection loss function alone, demonstrating the importance of fundamental knowledge for coarse mask awareness. The AP is increased by 13.7 when fundamental knowledge and apparent knowledge guide the network together, showing that the color information in the images is crucial in enhancing the mask awareness of the network. The mask awareness of the network is further augmented with the introduction of the MIAA head, and the AP_{75} improvement is the most significant. This suggests that detailed information about edges plays a vital role in acquiring high-quality masks. In conclusion, the network can effectively perform instance segmentation of SAR images without pixel-level annotation through the instantiation of three meta-knowledge components.

Figure 7 illustrates the qualitative experimental results of the impact of meta-knowledge on the SSDD dataset. Similarly to the experimental results in the NWPU VHR-10 instance segmentation dataset, in the absence of the color similarity loss function and the MIAA head, the accuracy of the mask segmented by the MGWI-Net is significantly limited, and severe over-segmentation occurs. When fundamental knowledge and apparent knowledge jointly guide the network, the segmentation performance of objects in SAR images is significantly improved. The segmentation performance of the model is further optimized by introducing detailed knowledge. The above experimental results show that our proposed MGWI-Net can segment instances in SAR images with meta-knowledge components.

4.5. Comparison of Annotation Costs

Intuitively, the production process of box-level annotations is more straightforward than that of pixel-level annotations, as no complex masks need to be drawn manually. Given the lack of quantitative research on annotation production costs in remote sensing, we have developed metrics that reflect the annotation cost in terms of time and workload.

Annotation production time: Obviously, the time can give an indication of the annotation production cost. Based on the calculation reported in [41], we split the production time for pixel-level (box-level) annotations into two parts: validating the categories and drawing the masks (bounding boxes). Then, we extracted 50 images from each dataset and drew masks and bounding boxes for all objects using the drawing tool unified in LabelMe [68]. For the images we selected in the NWPU VHR-10 instance segmentation dataset, the average time to obtain a mask was 57.1 s, while the average time to draw a bounding box was 6.6 s. According to [69], the time to validate a category is 1 s. Therefore, the average production time (T_{ann}) for pixel-level annotations per image was 8.6 classes/image × 1 s/class + 6.5 mask/image × 57.1 s/mask = 379.8 s/image. The average production time (T_{ann}) for box-level annotations per image was 8.6 classes/image × 1 s/class + 6.5 mask/image × 6.6 s/mask = 51.5 s/image.

For the SSDD dataset, the time to draw a mask and a bounding box was 43.0 s and 4.8 s, respectively. Therefore, the average time (T_{ann}) for pixel-level annotations per images in the SSDD dataset was 2.2 mask/image × 43.0 s/mask = 94.6 s/image, and the average production time (T_{ann}) for box-level annotations per image was 2.2 mask/image × 4.8 s/box = 10.6 s/image. In summary, the annotation production time for box-level annotations in the NWPU VHR-10 instance segmentation dataset and the SSDD dataset was only

13.6% and 11.2% of the annotation production time for pixel-level annotations, respectively. Thus, the time for producing box-level annotations is much lower than that for producing pixel-level annotations.

Annotation production workload: Although production time is a visual indicator of annotation production cost, it is susceptible to subjective factors such as the skill of the label maker. Since clicks generate pixel-level and box-level annotations, the average number of clicks (N_{cli}) per image is counted to reflect the annotation production workload objectively. As can be seen from Table 5, the average number of clicks for box-level annotation per image on the images we selected on the NWPU VHR-10 instance segmentation dataset was only 11.9% of that for pixel-level annotation. For the SSDD dataset, the number of clicks for box-level annotation was only 14.7% of that for pixel-level annotation, indicating the great advantage of WSIS in reducing the workload of manual annotation. Intuitively, pixel-level annotation requires clicks along a complex object contour for every object in the image, whereas box-level annotation requires only two clicks per object. Compared to standard fully supervised instance segmentation methods such as YOLACT, the mask R-CNN, and CondInst, our proposed MGWI-Net can rely on box-level annotations for mask awareness. It significantly saves time and workload and is significant for applying and promoting instance segmentation methods in remote sensing.

Table 5. Annotation production time and workload for optical and SAR images. The unit of N_{cli} is clicks/image. "Rate" indicates the ratio of box-level annotation metrics to pixel-level annotation metrics.

Dataset	Type of Annotation	T_{ann}	N_{cli}
NWPU VHR-10 instance segmentation dataset	Box-level Pixel-level	51.5 379.8	12.1 101.9
	Rate	13.6%	11.9%
SSDD dataset	Box-level Pixel-level	10.6 94.6	4.5 30.6
	Rate	11.2%	14.7%

4.6. Comparison of Other Methods

The above analysis demonstrates the considerable impact of meta-knowledge on mask awareness and the advantages of the MGWI-Net in terms of annotation production costs. Considering that there is little current research on the WSIS of optical and SAR images in remote sensing, three supervised modes, namely the weakly supervised mode, weakly supervised + fully supervised mode, and fully supervised mode, have been designed for comprehensive comparison with our proposed MGWI-Net. Specifically, the settings details of the three modes are as follows:

- Weakly supervised mode: For conventional fully supervised instance segmentation methods (such as YOLACT [48], the mask R-CNN [43], and CondInst [50]), we consider the region enclosed by the bounding box as one of pseudo-pixel-level annotations to guide the network in instance segmentation training and obtain predicted masks. BoxInst [36], DiscoBox [38], DBIN [40], and our proposed MGWI-Net are dedicated WSIS methods that do not require a pixel-level ground truth and can be trained directly with box-level annotations. Although conventional fully supervised instance segmentation methods use pseudo-pixel-level annotations during training, their ground truth is generated from box-level annotations. Hence, the annotation costs are consistent with those of dedicated WSIS methods. Additionally, as this study did not consider domain adaptation tasks, the DBIN does not have a corresponding domain adaptation structure.
- Weakly supervised + fully supervised mode: We adopted partial pixel-level annotations and some box-level labels to train the network. Note that only pixel-level

annotations were utilized for training the mask branch, while the classification and regression branches were trained using both box-level and pixel-level annotations. Compared with the proposed method, this mode has richer supervision information but requires a greater annotation cost.

- Fully supervised mode: The networks are trained with all pixel-level annotations. Compared with other modes, this mode requires constructing pixel-level annotations for each image, so its supervision information is the richest, and the annotation cost is the highest.

We report the experimental results of the NWPU VHR-10 instance segmentation dataset and the SSDD dataset in Table 6, Table 7 and Table 8 and also show the qualitative experimental results in Figures 8 and 9.

Table 6. Quantitative experimental results compared with those of other methods used on the NWPU VHR-10 instance segmentation dataset. R_{pix} indicates the proportion of annotations at the pixel level.

Supervision Mode	**Method**	R_{pix}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Weakly + fully supervised	YOLACT [48]	25%	15.2	41.2	7.8	7.7	16.8	12.6
	YOLACT [48]	50%	22.5	49.7	17.0	9.6	19.9	31.5
	YOLACT [48]	75%	27.5	54.4	27.4	12.1	25.9	34.2
	Mask R-CNN [43]	25%	25.7	59.4	18.8	16.9	25.3	29.3
	Mask R-CNN [43]	50%	35.5	70.8	31.3	24.6	34.2	39.9
	Mask R-CNN [43]	75%	49.3	82.6	51.7	36.9	47.0	53.9
	CondInst [50]	25%	23.9	59.8	14.8	19.8	23.7	25.3
	CondInst [50]	50%	34.5	73.4	27.6	23.7	34.1	35.9
	CondInst [50]	75%	49.5	85.1	50.3	35.9	48.6	53.7
Fully supervised	YOLACT [48]	100%	35.6	68.4	36.4	14.8	33.3	56.0
	Mask R-CNN [43]	100%	58.8	86.6	65.2	47.1	57.5	62.4
	CondInst [50]	100%	58.5	90.1	62.9	29.4	56.8	71.3
Weakly supervised	YOLACT [48]	0	9.8	32.9	1.3	4.4	11.3	8.0
	Mask R-CNN [43]	0	19.8	54.7	9.7	7.8	19.4	24.6
	CondInst [50]	0	17.1	50.5	6.7	10.7	17.7	18.5
	BoxInst [36]	0	47.6	78.9	49.0	33.8	43.9	55.5
	DiscoBox [38]	0	46.2	79.7	47.4	29.4	42.9	57.1
	DBIN [40]	0	48.3	80.2	50.5	34.5	46.1	57.0
	MGWI-Net	0	51.6	81.3	53.3	37.6	48.2	59.1

Results of experiments on the NWPU VHR-10 instance segmentation dataset: As shown in Table 6, compared to conventional instance segmentation methods in the weakly supervised mode, MGWI-Net achieves 31.8, 26.6, and 43.6 higher AP, AP_{50}, and AP_{75} than the mask R-CNN does in the weakly supervised mode in the NWPU VHR-10 instance segmentation dataset, respectively. This indicates that our proposed method has better mask awareness than the conventional method does under box-level annotation conditions. Compared to dedicated WSIS methods, the MGWI-Net achieves an AP value that is 3.3 times higher than that of the best-performing DBIN, which further demonstrates the advantage of our proposed meta-knowledge-guided WSIS network in instance segmentation. Table 6 shows the segmentation results for each category in the NWPU VHR-10 instance segmentation dataset. The MGWI-Net achieves the best segmentation for all categories except for harbor (HB) and baseball diamond (BD) (which achieved suboptimal results with a slight disadvantage of 0.7 and 0.4, respectively), demonstrating its superior segmentation capability compared to that of the conventional methods and the dedicated WSIS methods under the box-level annotation condition.

Table 7. Class-wise quantitative experimental results compared with those of other methods used on the NWPU VHR-10 instance segmentation dataset.

Supervision Mode	Method	R_{pix}	AI	BD	GTF	VC	SH	TC	HB	ST	BC	BR	AP
Weakly + fully supervised	YOLACT [48]	25%	0	39.1	14.2	12.2	1.5	7.6	12.2	49.8	10.3	1.4	15.2
	YOLACT [48]	50%	0.1	55.0	49.9	11.7	4.6	17.0	19.3	52.5	13.6	1.2	22.5
	YOLACT [48]	75%	0.7	64.0	62.9	19.4	5.9	19.8	17.4	60.7	16.9	7.2	27.5
	Mask R-CNN [43]	25%	0.1	34.2	37.4	12.5	4.7	39.3	22.5	56.6	42.0	7.8	25.7
	Mask R-CNN [43]	50%	8.8	49.7	50.8	28.1	14.9	44.1	30.6	59.4	52.1	16.7	35.5
	Mask R-CNN [43]	75%	27.6	71.2	68.6	40.5	30.4	63.2	33.0	70.1	64.7	23.9	49.3
	CondInst [50]	25%	0	37.5	35.7	18.2	3.0	36.6	18.2	53.8	32.5	3.5	23.9
	CondInst [50]	50%	14.8	49.0	45.6	23.0	12.8	45.6	28.4	57.1	56.8	11.3	34.5
	CondInst [50]	75%	30.9	68.4	64.5	41.5	31.8	68.4	30.3	65.0	66.0	27.8	49.5
Fully supervised	YOLACT [48]	100%	8.2	70.5	70.8	22.7	21.5	24.3	34.8	63.4	26.5	13.5	35.6
	Mask R-CNN [43]	100%	35.3	78.8	84.8	46.1	50.2	72.0	48.1	80.9	64.2	28.0	58.8
	CondInst [50]	100%	26.7	77.7	89.1	46.2	46.1	69.7	46.8	73.4	74.0	35.4	58.5
Weakly supervised	YOLACT [48]	0	0	20.7	12.1	4.8	0.1	9.6	2.1	33.5	14.9	0.1	9.8
	Mask R-CNN [43]	0	0	33.3	34.2	8.0	2.3	21.5	16.4	48.6	26.9	6.6	19.8
	CondInst [50]	0	0	30.7	26.8	6.6	1.1	19.2	14.2	46.1	23.1	3.1	17.1
	BoxInst [36]	0	12.5	76.6	89.7	38.0	47.9	65.5	11.3	75.4	58.9	6.8	47.6
	DiscoBox [38]	0	12.0	77.7	91.5	33.7	42.8	64.3	10.6	74.6	57.9	6.0	46.2
	DBIN [40]	0	14.0	77.1	91.2	37.8	48.6	67.8	13.0	75.2	61.9	5.4	48.3
	MGWI-Net	0	17.0	77.3	91.9	41.0	50.8	71.2	15.7	76.5	64.6	10.9	51.6

Table 8. Quantitative experimental results compared with those of other methods for the SSDD dataset.

Supervision Mode	Method	R_{pixel}	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Weakly + fully supervised	YOLACT [48]	25%	17.4	59.0	1.5	19.7	21.0	-
	YOLACT [48]	50%	28.6	76.7	9.0	32.1	34.2	-
	YOLACT [48]	75%	39.0	79.9	32.5	40.3	45.5	-
	Mask R-CNN [43]	25%	22.8	72.4	6.3	27.2	28.5	-
	Mask R-CNN [43]	50%	39.3	86.2	28.0	42.7	44.4	-
	Mask R-CNN [43]	75%	54.6	90.2	63.0	56.6	57.1	-
	CondInst [50]	25%	18.6	65.7	2.7	22.1	23.7	-
	CondInst [50]	50%	38.4	87.4	28.6	41.3	43.8	-
	CondInst [50]	75%	54.1	93.0	59.6	54.6	56.8	-
Fully supervised	YOLACT [48]	100%	44.6	86.6	41.0	45.3	48.5	-
	Mask R-CNN [43]	100%	64.2	94.9	80.1	62.0	64.7	-
	CondInst [50]	100%	63.0	95.9	78.4	63.7	63.6	-
Weakly supervised	YOLACT [48]	0	12.4	49.4	0.6	15.9	17.3	-
	Mask R-CNN [43]	0	15.5	61.0	1.6	20.2	21.1	-
	CondInst [50]	0	14.8	59.1	1.4	17.7	19.6	-
	BoxInst [36]	0	49.9	90.1	52.7	50.6	52.3	-
	DiscoBox [38]	0	48.4	90.2	50.4	47.2	50.6	-
	DBIN [40]	0	50.6	91.7	52.8	51.3	52.0	-
	MGWI-Net	0	53.0	92.4	57.1	53.7	54.9	-

In addition to the weakly supervised model, we also investigated the performance of the conventional instance segmentation method under different percentages of pixel-level annotation for further comparison with our proposed MGWI-Net. As shown in Table 6, MGWI-Net achieved an *AP* of 51.6 with no pixel-level annotations at all, which is not only much higher than the segmentation results of the conventional methods with 25% and 50% pixel-level annotations but also better than the segmentation results of the conventional methods with 75% pixel-level annotations. Table 7 gives the segmentation results of the conventional segmentation method for each category in the weakly supervised + fully supervised mode. In contrast, our MGWI-Net can outperform this method with 75% pixel-level labels in many categories by relying on box-level annotations alone. We also compared the MGWI-Net with YOLACT, the mask R-CNN, and CondInst in the fully supervised mode. The AP of the proposed MGWI-Net is much higher than that of the fully supervised YOLACT and can reach 87.8% of that of the fully supervised mask R-CNN. Combined with the annotation production cost analysis in the previous section, we find that MGWI-Net can achieve 87.8% of the instance segmentation performance of the fully supervised mask R-CNN with 13.6% of the annotation production time.

Figure 8 shows the segmentation results of the MGWI-Net compared with those of the weakly supervised, fully supervised, and weakly + fully supervised approaches. Our method can not only effectively segment the objects such as baseball diamonds, ground track fields, and ships in optical images but also achieve good segmentation results for airplanes with complex contours and bridges with large sizes. In contrast, YOLACT, the mask R-CNN, and CondInst in the weakly supervised mode suffer from significant over-segmentation due to the lack of pixel-level annotations. Compared with the dedicated WSIS methods, our proposed MGWI-Net-segmented masks not only can effectively cover the objects in the image but also have good edge detail, further demonstrating the excellent performance of this method. The fully supervised approaches have the best instance segmentation performance as they depend on all pixel-level annotations. The instance segmentation in the weakly supervised + fully supervised mode gradually improved with increasing pixel-level annotations. Overall, the MGWI-Net can make the instance segmentation of remote sensing images no longer rely on complex pixel-level methods and achieve a balance between annotation production costs and segmentation results.

Results of experiments on the SSDD dataset: Like the NWPU VHR-10 instance segmentation dataset, we compared the results of the MGWI-Net with those of the weakly supervised, weakly supervised + fully supervised, and fully supervised modes used on the SSDD dataset. From Table 8, our MGWI-Net has higher AP, AP_{50}, and AP_{75} values than the weakly supervised mask R-CNN does, which shows that the meta-knowledge can guide our MGWI-Net to adapt to the mask-aware task. We also find that the proposed MGWI-Net outperformed the dedicated WSIS method in terms of AP when used on the SSDD dataset, which indicates that the method competes with not only the conventional methods but also the dedicated WSIS methods.

Furthermore, the MGWI-Net also outperforms the segmentation results of conventional methods with 25% and 50% pixel-level annotations and is close to those of themask R-CNN and CondInst in the 75% pixel-level annotations' condition. This further demonstrates the excellent performance of our method. In addition, the MGWI-Net has an AP of 8.4 times higher than that of the fully supervised YOLACT and can achieve 82.6% of that of the fully supervised mask R-CNN. Thus, the MGWI-Net can have a performance of 82.6% of the fully supervised mask R-CNN's instance segmentation performance when used on the SSDD dataset with 11.2% of the annotation production time.

Figure 9 shows the qualitative experimental results of the MGWI-Net being used on the SSDD dataset compared with the segmentation effect of the fully supervised mode, weakly supervised mode, and weakly supervision + fully supervision mode methods. Among them, (a) and (b) show that our MGWI-Net can obtain high-quality masks close to those of fully supervised methods. (c) and (d) compare the MGWI-Net and conventional methods in the weakly supervised mode. Due to the lack of pixel-level labels, the segmentation effect of traditional methods in the weakly supervised mode is significantly worse. (e) and (f) are the segmentation results of our MGWI-Net and other dedicated WSIS methods, showing that these methods can accurately segment ships in SAR images without relying on pixel-level annotations. (g) and (i) show that the MGWI-Net significantly outperforms the weakly supervised + fully supervised mode approaches of the 25% and 50% pixel-level annotation conditions and is comparable to the mask R-CNN and CondInst of the 75% pixel-level annotation condition.

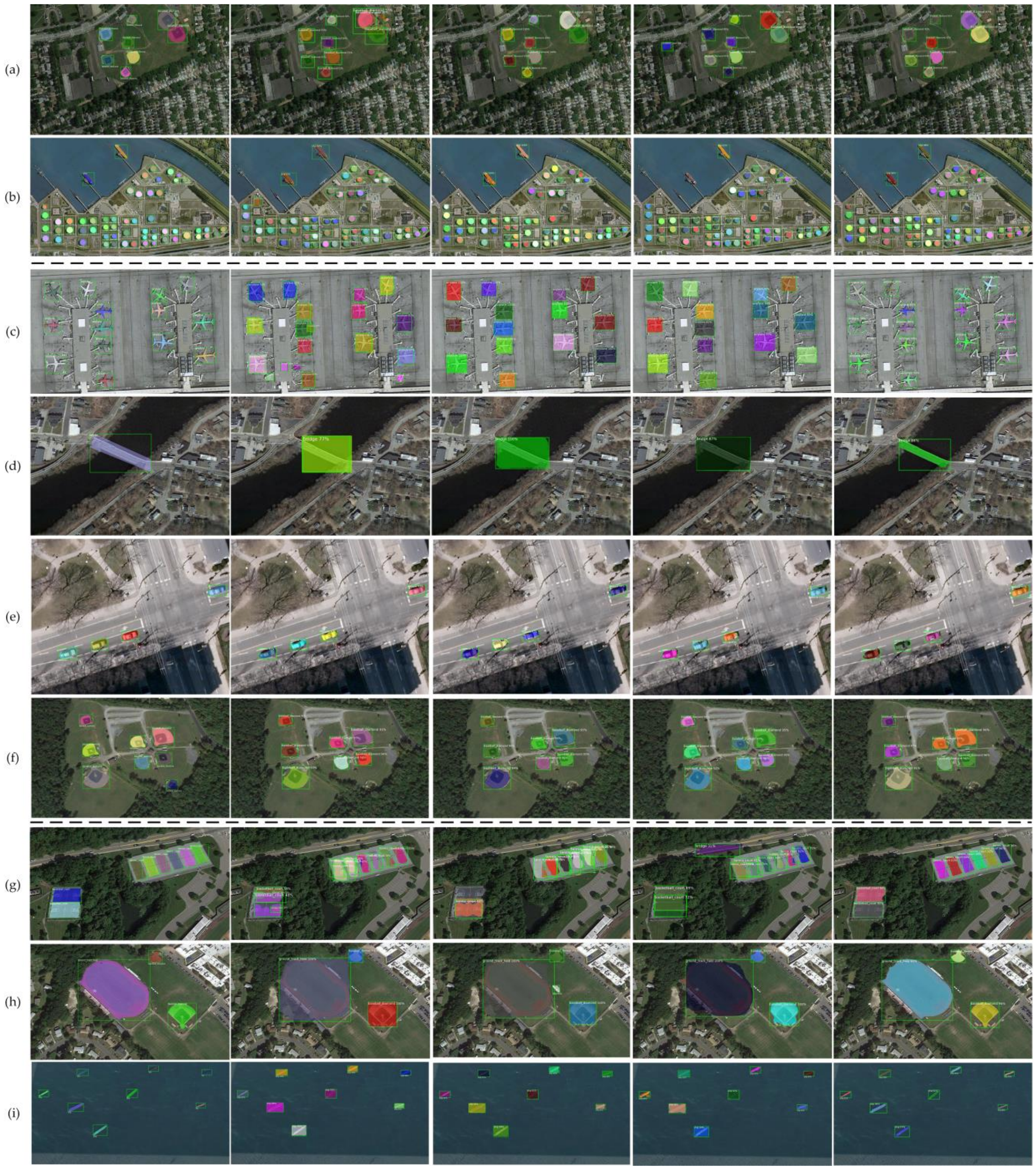


Figure 8. Comparison of qualitative experimental results with those of other methods used on the NWPU VHR-10 instance segmentation dataset. (**a**,**b**) Comparisons between the MGWI-Net and the fully supervised model approach, where the second to fourth columns show the results of YOLACT, mask R-CNN, and CondInst in the fully supervised mode, respectively. (**c**,**d**) Comparison of the MGWI-Net and the conventional fully supervised instance segmentation methods in the weakly supervised mode. The second to fourth columns are the segmentation results of YOLACT, mask R-CNN, and CondInst in the weakly supervised mode, respectively. (**e**,**f**) Comparison of the MGWI-Net with the dedicated WSIS methods, where the second to fourth columns are the segmentation results of BoxInst, DiscoBox, and DBIN, respectively. (**g**–**i**) Comparisons between the MGWI-Net and the weakly supervised + fully supervised methods, where the second to fourth columns of (**g**) show the segmentation effects of YOLACT with 25%, 50%, and 75% pixel-level labels, respectively, the second to fourth columns of (**h**) are the processing results of the mask R-CNN with 25%, 50%, and 75% pixel-level labels, respectively, and the second to fourth columns of (**i**) show the results of 25%, 50% and 75% pixel-level labels of CondInst, respectively. The first and fifth columns of (**a**–**i**) show the ground truths and the MGWI-Net's segmentation results, respectively.

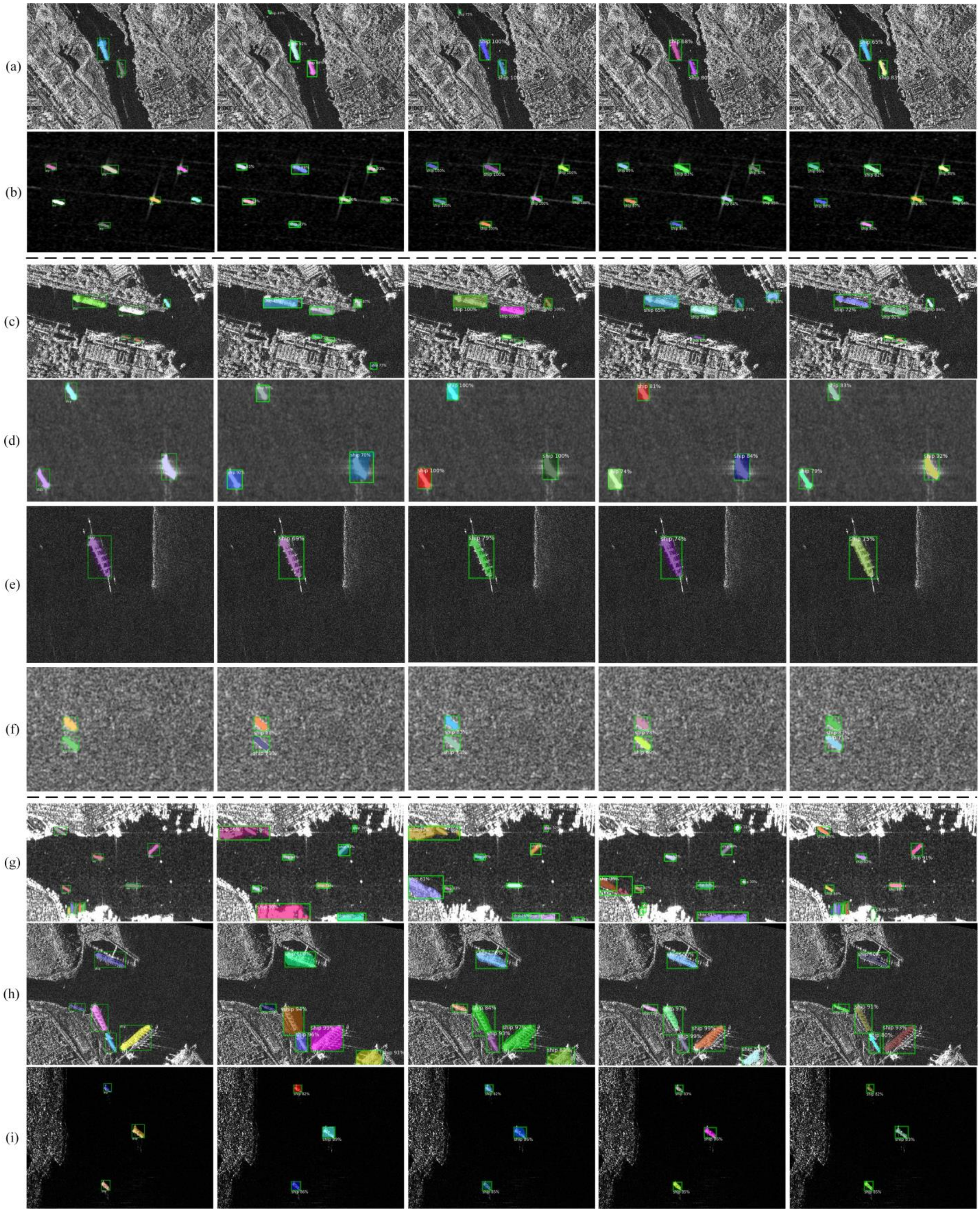


Figure 9. Comparison of qualitative experimental results with other methods in the SSDD dataset. (**a**,**b**) Comparisons of the MGWI-Net with the fully supervised approaches. (**c**–**f**) Comparisons of the MGWI-Net with the weakly supervised approaches. (**g**–**i**) Comparisons of the MGWI-Net with the fully + weakly supervised approaches. The layout of the columns is the same as that in Figure 8.

Considering the frequent presence of various interference factors in SAR images, such as speckle noise and geometric distortion, we also present some examples of how the interference factors affect the segmentation performance of the MGWI-Net. In Figure 10a, there is a large amount of speckle noise, which blurs the ship, especially at the edges. Our proposed MGWI-Net can accurately locate and perceive the ship's mask, but the edge-processing effect is generally unsatisfactory. In Figure 10b, there is some speckle noise, and the ships occupy small areas in the image, resulting in missed segmentation for the ships in the red ellipses. The texture of the ship in Figure 10c is not uniform, so the quality of the segmented mask is not high. Overall, our MGWI-Net can perceive the masks of ships in SAR images without pixel-level annotations, but interference factors also affect the segmentation effect.

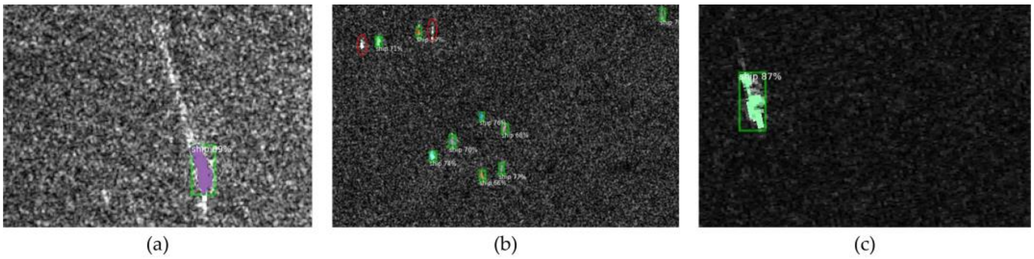


Figure 10. Some examples of how the interference factors affect the segmentation performance of MGWI-Net. (**a**,**b**) The instance segmentation results of MGWI-Net for images with speckle noise. The ships that miss segmentation in (**b**) are highlighted with red ellipses. (**c**) The instance segmentation results of MGWI-Net for ships with uneven texture.

Furthermore, we noticed that the SSDD dataset contains various polarization modes (HH, VV, HV, and VH), which may have a certain impact on segmentation. Since not knowing the specific polarization mode of each image makes quantitative analysis difficult, we conducted a quantitative analysis based on the characteristics of different polarization modes and the specific ship segmentation task. Generally, HH and VV polarization modes have better reflection capabilities for targets with horizontal and vertical surfaces, respectively. In contrast, HV and VH polarization modes can provide reflection signals in different directions, thereby providing a more comprehensive understanding of the characteristics of objects. In addition, HV and VH polarization modes may be more suitable for ship segmentation. The impact of different polarization modes on ship instance segmentation also depends on internal factors (such as the ship's material) and external factors (such as the marine environment). In future work, we will quantitatively explore the impact of polarization modes on instance segmentation, providing more comprehensive technical support for interpreting SAR images.

Model Parameters and Inference Speed: The above experiments on the NWPU VHR-10 instance segmentation dataset and the SSDD dataset show that the MGWI-Net can balance the visual processing results and annotation costs. We further tested the model parameters and inference speed of the MGWI-Net to reflect its performance fully, and the results are recorded in Table 9.

As seen from Table 9, the parameters of the MGWI-Net are lower than those of the classical mask R-CNN, and its network inference speed is close to that of CondInst. Furthermore, it is clear from the previous experimental results and analysis that the segmentation effects of YOLACT are much worse than those of our proposed MGWI-Net. However, YOLACT achieves optimal results in terms of parameters and inference speed. It is worth noting that the MGWI-Net's model parameters and inference speed are comparable to those of BoxInst and DBIN. DiscoBox has a slightly faster inference speed but larger parameters. Therefore, the MGWI-Net can rely on box-level annotations for high-quality instance segmentation while maintaining suitable model parameters and inference speed.

Table 9. Comparison of model parameters and inference speed between MGWI-Net and other methods. "FPS-NW10" and "FPS-SSDD" denote the frames per second (FPS) in the NWPU VHR-10 instance segmentation and SSDD datasets, respectively.

Method	Parameters/M	FPS-NW10	FPS-SSDD
YOLACT [48]	34.8	16.4	22.8
Mask R-CNN [43]	63.3	13.5	19.7
CondInst [50]	53.5	10.6	15.7
BoxInst [36]	53.5	10.6	15.6
DiscoBox [38]	65.0	11.0	16.5
DBIN [40]	55.6	10.1	15.3
MGWI-Net	53.7	10.4	15.4

4.7. Ablation Study

We conducted ablation studies to explore better the impact of some key parameters in the model. Firstly, we conducted ablation studies on the effect of the similarity threshold (θ) in the color similarity loss function on the NWPU VHR-10 instance segmentation dataset and the SSDD dataset to investigate its effect on segmentation performance. The main experimental results are shown in Table 10. Secondly, we explore the weight (λ_3) of the awareness-assisted loss function (L_{auxi}) in the MIAA head, and the experimental results are recorded in Table 11.

Table 10. Ablation study of the effect of the similarity threshold (θ) on the NWPU VHR-10 instance segmentation and SSDD datasets.

Dataset	Similarity Threshold θ	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
NWPU VHR-10 instance segmentation dataset	0	9.9	34.4	1.4	9.9	11.3	9.7
	0.1	50.7	81.0	51.7	36.8	47.6	59.2
	0.2	51.4	80.8	53.4	37.1	48.1	58.7
	0.3	51.6	81.3	53.3	37.6	48.2	59.1
	0.4	50.9	81.2	52.2	37.5	47.4	58.8
SSDD dataset	0	11.0	47.6	2.1	16.6	18.2	-
	0.1	52.6	91.4	56.2	52.8	55.0	-
	0.2	53.3	91.9	57.5	53.5	55.6	-
	0.3	53.0	92.4	57.1	53.7	54.9	-
	0.4	52.2	92.0	56.8	53.2	54.1	-

Table 11. The values of AP obtained by MGWI-Net under different λ_3 values. "NW10" is the NWPU VHR-10 instance segmentation dataset.

Dataset	λ_3 = 0.1	λ_3 = 0.5	λ_3 = 1.0	λ_3 = 1.5
NW10	51.6	51.4	50.6	49.9
SSDD	53.0	52.9	52.5	51.4

As shown in Table 10, when the similarity threshold (θ) was 0, all pixel position pairs belonged to the same object. Therefore, the segmentation metrics under this condition unsurprisingly achieved the lowest value. When the θ increased to 0.1, the segmentation performance was significantly improved. When the similarity threshold (θ) continued to increase to 0.2, 0.3, and 0.4, the MGWI-Net achieved good segmentation results, indicating that our proposed method is not sensitive to this hyperparameter. From the segmentation performance of the MGWI-Net when used on the two datasets, the θ values of 0.2 or 0.3 were suitable, and the θ in this study was set to 0.3.

As shown in Table 11, when the weight (λ_3) was 0.1 or 0.5, the MGWI-Net achieved good instance segmentation results, indicating that the model could effectively balance the relationship between the awareness-assisted loss function and other loss functions.

However, when the weight (λ_3)was set to 1.5, the segmentation performance decreased, reflecting that setting an enormous weight for the awareness-assisted loss function is not conducive to performance improvement.

5. Discussion

This study regards mask awareness in WSIS as a task adaptation problem. Combining the meta-knowledge theory and human visual perception habits, we scientifically deconstructed and instantiated prior knowledge in mask awareness to enable the network to adapt from simple box-level annotations to complex instance segmentation. Specifically, we proposed the MGWI-Net to achieve pixel-by-pixel interpretation of optical and SAR images in remote sensing. This network can provide similar instance segmentation effects with a much shorter annotation production time and workload than the fully supervised methods. This work can provide inexpensive and convenient technical support for applying and promoting instance segmentation methods for optical and SAR images.

However, this work also has some limitations that require further improvement and optimization. Firstly, the MGWI-Net distinguishes objects from the background based on the color similarity between pixel positions, which requires some difference between the objects and the background. Therefore, this method has limitations in terms of segmenting objects with high similarity to the background, such as airplanes. Secondly, the impact of polarization modes on segmentation results needs to be further quantitatively analyzed. Thirdly, the proposed method relies on box-level annotations, which still require some annotation costs and may affect the method's convenience in practical applications, even if the annotation production time is only one-eighth of that of pixel-level annotations. Finally, since our method is trained without pixel-level annotations, the segmentation accuracy is not as good as that of the fully supervised methods, which may not meet the requirements of some fields that demand high-quality masks, such as medical image processing and autonomous driving.

In future work, we plan to study more aspects of meta-knowledge, such as object texture and shape, and incorporate them into the network design process to overcome the limitations in segmenting objects with high similarity to the background and further improve the quality of segmentation masks. Additionally, for SAR image instance segmentation, we will focus on reducing speckle noise in SAR images and improving the network's segmentation performance for objects with uneven textures through data augmentation or other methods. We will also quantitatively explore the effect of the polarization mode on instance segmentation and consider fusing different polarization information to enhance the instance segmentation of ships, providing more comprehensive technical support for interpreting SAR images. Furthermore, we plan to extend the meta-knowledge theory to more sophisticated image-level supervised instance segmentation paradigms and cross-domain adaptive methods in future work, further reducing the dependency of instance segmentation on annotations. Finally, we will actively extend the idea of meta-knowledge to other fields and tasks in future work and design meta-knowledge suitable for the actual situation of different fields and tasks.

6. Conclusions

In this work, we fully balance the visual processing requirements and the annotation costs and introduce the box-supervised segmentation paradigm to the optical and SAR image interpretation work. We first decompose prior knowledge in mask awareness through the meta-knowledge theory and the habits of human visual perception into three meta-knowledge components: fundamental knowledge, apparent knowledge, and detailed knowledge. Subsequently, a meta-knowledge-guided weakly supervised instance segmentation network (MGWI-Net) is proposed. The weakly supervised mask (WSM) head of the MGWI-Net can instantiate fundamental and apparent knowledge, and then guide the network to perceive instance masks with box-level annotations. Furthermore, the mask information awareness assist (MIAA) head can implicitly guide the network to learn detailed information about edges using the boundary-sensitive feature of the fully connected CRF, which facilitates the instantiation of detailed knowledge. The experimental results show that the MGWI-Net can efficiently perform mask awareness for optical and SAR images, achieving instance segmentation results close to those of fully supervised methods in about one-eighth of the annotation production time. This work can provide low-cost and convenient technical support for applying and extending instance segmentation methods for optical and SAR images.

Author Contributions: Conceptualization, M.C. and Z.P.; methodology, M.C. and Y.Z.; software, M.C., E.C. and Y.H.; validation, M.C., Y.Z. and E.C.; formal analysis, M.C. and Y.X.; investigation, Y.H. and Z.P.; resources, Z.P.; data curation, Y.Z. and E.C.; writing—original draft preparation, M.C., Y.Z. and Y.H.; writing—review and editing, Z.P. and Y.X.; visualization, M.C. and Y.X.; supervision, Z.P.; project administration, M.C.; funding acquisition, Z.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Amitrano, D.; Di Martino, G.; Guida, R.; Iervolino, P.; Iodice, A.; Papa, M.N.; Riccio, D.; Ruello, G. Earth Environmental Monitoring Using Multi-Temporal Synthetic Aperture Radar: A Critical Review of Selected Applications. *Remote Sens.* **2021**, *13*, 604. [CrossRef]
2. Liu, C.; Xing, C.; Hu, Q.; Wang, S.; Zhao, S.; Gao, M. Stereoscopic Hyperspectral Remote Sensing of the Atmospheric Environment: Innovation and Prospects. *Earth Sci. Rev.* **2022**, *226*, 103958. [CrossRef]
3. Wu, Z.; Hou, B.; Ren, B.; Ren, Z.; Wang, S.; Jiao, L. A Deep Detection Network Based on Interaction of Instance Segmentation and Object Detection for SAR Images. *Remote Sens.* **2021**, *13*, 2582. [CrossRef]
4. Zhu, M.; Hu, G.; Li, S.; Zhou, H.; Wang, S.; Feng, Z. A Novel Anchor-Free Method Based on FCOS + ATSS for Ship Detection in SAR Images. *Remote Sens.* **2022**, *14*, 2034. [CrossRef]
5. Bühler, M.M.; Sebald, C.; Rechid, D.; Baier, E.; Michalski, A.; Rothstein, B.; Nübel, K.; Metzner, M.; Schwieger, V.; Harrs, J.-A.; et al. Application of Copernicus Data for Climate-Relevant Urban Planning Using the Example of Water, Heat, and Vegetation. *Remote Sens.* **2021**, *13*, 3634. [CrossRef]
6. Yu, D.; Ji, S.; Li, X.; Yuan, Z.; Shen, C. Earthquake Crack Detection from Aerial Images Using a Deformable Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
7. Zhang, T.; Zhang, X.; Ke, X.; Liu, C.; Xu, X.; Zhan, X.; Wang, C.; Ahmad, I.; Zhou, Y.; Pan, D.; et al. HOG-ShipCLSNet: A Novel Deep Learning Network with HOG Feature Fusion for SAR Ship Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–22. [CrossRef]
8. Liu, X.; Huang, Y.; Wang, C.; Pei, J.; Huo, W.; Zhang, Y.; Yang, J. Semi-Supervised SAR ATR via Conditional Generative Adversarial Network with Multi-Discriminator. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Brussels, Belgium, 11–16 July 2021; pp. 2361–2364. [CrossRef]
9. Hao, S.; Wu, B.; Zhao, K.; Ye, Y.; Wang, W. Two-Stream Swin Transformer with Differentiable Sobel Operator for Remote Sensing Image Classification. *Remote Sens.* **2022**, *14*, 1507. [CrossRef]
10. Miao, W.; Geng, J.; Jiang, W. Semi-Supervised Remote-Sensing Image Scene Classification Using Representation Consistency Siamese Network. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
11. Chen, S.-B.; Wei, Q.-S.; Wang, W.-Z.; Tang, J.; Luo, B.; Wang, Z.-Y. Remote Sensing Scene Classification via Multi-Branch Local Attention Network. *IEEE Trans. Image Process.* **2022**, *31*, 99–109. [CrossRef]
12. Shi, C.; Fang, L.; Lv, Z.; Shen, H. Improved Generative Adversarial Networks for VHR Remote Sensing Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
13. Zhang, T.; Zhang, X.; Ke, X. Quad-FPN: A Novel Quad Feature Pyramid Network for SAR Ship Detection. *Remote Sens.* **2021**, *13*, 2771. [CrossRef]
14. Chen, S.; Zhang, J.; Zhan, R.; Zhu, R.; Wang, W. Few Shot Object Detection for SAR Images via Feature Enhancement and Dynamic Relationship Modeling. *Remote Sens.* **2022**, *14*, 3669. [CrossRef]
15. Zhang, R.; Zhang, X.; Zheng, Y.; Wang, D.; Hua, L. MSCNet: A Multilevel Stacked Context Network for Oriented Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5066. [CrossRef]

16. Wang, J.; Gong, Z.; Liu, X.; Guo, H.; Yu, D.; Ding, L. Object Detection Based on Adaptive Feature-Aware Method in Optical Remote Sensing Images. *Remote Sens.* **2022**, *14*, 3616. [CrossRef]
17. Liu, B.; Hu, J.; Bi, X.; Li, W.; Gao, X. PGNet: Positioning Guidance Network for Semantic Segmentation of Very-High-Resolution Remote Sensing Images. *Remote Sens.* **2022**, *14*, 4219. [CrossRef]
18. Feng, M.; Sun, X.; Dong, J.; Zhao, H. Gaussian Dynamic Convolution for Semantic Segmentation in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 5736. [CrossRef]
19. Kong, Y.; Li, Q. Semantic Segmentation of Polarimetric SAR Image Based on Dual-Channel Multi-Size Fully Connected Convolutional Conditional Random Field. *Remote Sens.* **2022**, *14*, 1502. [CrossRef]
20. Colin, A.; Fablet, R.; Tandeo, P.; Husson, R.; Peureux, C.; Longépé, N.; Mouche, A. Semantic Segmentation of Metoceanic Processes Using SAR Observations and Deep Learning. *Remote Sens.* **2022**, *14*, 851. [CrossRef]
21. Su, H.; Wei, S.; Liu, S.; Liang, J.; Wang, C.; Shi, J.; Zhang, X. HQ-ISNet: High-Quality Instance Segmentation for Remote Sensing Imagery. *Remote Sens.* **2020**, *12*, 989. [CrossRef]
22. Zeng, X.; Wei, S.; Wei, J.; Zhou, Z.; Shi, J.; Zhang, X.; Fan, F. CPISNet: Delving into Consistent Proposals of Instance Segmentation Network for High-Resolution Aerial Images. *Remote Sens.* **2021**, *13*, 2788. [CrossRef]
23. Zhang, T.; Zhang, X.; Zhu, P.; Tang, X.; Li, C.; Jiao, L.; Zhou, H. Semantic Attention and Scale Complementary Network for Instance Segmentation in Remote Sensing Images. *IEEE Trans. Cybern.* **2022**, *52*, 10999–11013. [CrossRef] [PubMed]
24. Zhao, D.; Zhu, C.; Qi, J.; Qi, X.; Su, Z.; Shi, Z. Synergistic Attention for Ship Instance Segmentation in SAR Images. *Remote Sens.* **2021**, *13*, 4384. [CrossRef]
25. Shi, F.; Zhang, T. An Anchor-Free Network with Box Refinement and Saliency Supplement for Instance Segmentation in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
26. Ke, X.; Zhang, X.; Zhang, T. GCBANet: A Global Context Boundary-Aware Network for SAR Ship Instance Segmentation. *Remote Sens.* **2022**, *14*, 2165. [CrossRef]
27. Fan, F.; Zeng, X.; Wei, S.; Zhang, H.; Tang, D.; Shi, J.; Zhang, X. Efficient Instance Segmentation Paradigm for Interpreting SAR and Optical Images. *Remote Sens.* **2022**, *14*, 531. [CrossRef]
28. Uijlings, J.R.R.; van de Sande, K.E.A.; Gevers, T.; Smeulders, A.W.M. Selective Search for Object Recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
29. Pont-Tuset, J.; Arbelaez, P.; Barron, J.T.; Marques, F.; Malik, J. Multiscale Combinatorial Grouping for Image Segmentation and Object Proposal Generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 128–140. [CrossRef]
30. Zhou, Y.; Zhu, Y.; Ye, Q.; Qiu, Q.; Jiao, J. Weakly Supervised Instance Segmentation Using Class Peak Response. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 3791–3800. [CrossRef]
31. Ahn, J.; Cho, S.; Kwak, S. Weakly Supervised Learning of Instance Segmentation with Inter-Pixel Relations. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2204–2213. [CrossRef]
32. Zhu, Y.; Zhou, Y.; Xu, H.; Ye, Q.; Doermann, D.; Jiao, J. Learning Instance Activation Maps for Weakly Supervised Instance Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3111–3120. [CrossRef]
33. Ge, W.; Guo, S.; Huang, W.; Scott, M.R. Label-PEnet: Sequential Label Propagation and Enhancement Networks for Weakly Supervised Instance Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 3345–3354. [CrossRef]
34. Arun, A.; Jawahar, C.V.; Kumar, M.P. Weakly Supervised Instance Segmentation by Learning Annotation Consistent Instances. In Proceedings of the European Conference on Computer Vision (ECCV), 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 254–270. [CrossRef]
35. Hsu, C.-C.; Hsu, K.-J.; Tsai, C.-C.; Lin, Y.-Y.; Chuang, Y.-Y. Weakly Supervised Instance Segmentation Using the Bounding Box Tightness Prior. In Proceedings of the 2019 Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; pp. 6586–6597. [CrossRef]
36. Tian, Z.; Shen, C.; Wang, X.; Chen, H. BoxInst: High-Performance Instance Segmentation with Box Annotations. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5443–5452. [CrossRef]
37. Hao, S.; Wang, G.; Gu, R. Weakly Supervised Instance Segmentation Using Multi-Prior Fusion. *Comput. Vis. Image Underst.* **2021**, *211*, 103261. [CrossRef]
38. Lan, S.; Yu, Z.; Choy, C.; Radhakrishnan, S.; Liu, G.; Zhu, Y.; Davis, L.S.; Anandkumar, A. DiscoBox: Weakly Supervised Instance Segmentation and Semantic Correspondence from Box Supervision. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 3386–3396. [CrossRef]
39. Wang, X.; Feng, J.; Hu, B.; Ding, Q.; Ran, L.; Chen, X.; Liu, W. Weakly-Supervised Instance Segmentation via Class-Agnostic Learning with Salient Images. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10225–10235. [CrossRef]
40. Li, Y.; Xue, Y.; Li, L.; Zhang, X.; Qian, X. Domain Adaptive Box-Supervised Instance Segmentation Network for Mitosis Detection. *IEEE Trans. Med. Imaging.* **2022**, *41*, 2469–2485. [CrossRef]

41. Bellver, M.; Salvador, A.; Torres, J.; Giro-i-Nieto, X. Budget-Aware Semi-Supervised Semantic and Instance Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 93–102. [CrossRef]
42. Krähenbühl, P.; Koltun, V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. *arXiv* **2012**, arXiv:1210.5644.
43. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
44. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768. [CrossRef]
45. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162. [CrossRef]
46. Chen, K.; Ouyang, W.; Loy, C.C.; Lin, D.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; et al. Hybrid Task Cascade for Instance Segmentation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4969–4978. [CrossRef]
47. Liu, J.-J.; Hou, Q.; Cheng, M.-M.; Wang, C.; Feng, J. Improving Convolutional Networks with Self-Calibrated Convolutions. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10093–10102. [CrossRef]
48. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT: Real-Time Instance Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 9156–9165. [CrossRef]
49. Xie, E.; Sun, P.; Song, X.; Wang, W.; Liang, D.; Shen, C.; Luo, P. PolarMask: Single Shot Instance Segmentation with Polar Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12193–12202. [CrossRef]
50. Tian, Z.; Shen, C.; Chen, H. Conditional Convolutions for Instance Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 254–270. [CrossRef]
51. Wang, X.; Kong, T.; Shen, C.; Jiang, Y.; Li, L. SOLO: Segmenting Objects by Locations. In Proceedings of the European Conference on Computer Vision (ECCV), 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 649–665. [CrossRef]
52. Wang, X.; Zhang, R.; Kong, T.; Li, L.; Shen, C. SOLOv2: Dynamic and Fast Instance Segmentation. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020; pp. 17721–17732. [CrossRef]
53. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
54. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440. [CrossRef]
55. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, NSW, Australia, 6–11 August 2017; pp. 1126–1135. [CrossRef]
56. Hospedales, T.M.; Antoniou, A.; Micaelli, P.; Storkey, A.J. Meta-Learning in Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 5149–5169. [CrossRef]
57. Zhang, J.; Zhang, X.; Zhang, Y.; Duan, Y.; Li, Y.; Pan, Z. Meta-Knowledge Learning and Domain Adaptation for Unseen Background Subtraction. *IEEE Trans. Image Process.* **2021**, *30*, 9058–9068. [CrossRef]
58. Tonioni, A.; Rahnama, O.; Joy, T.; Di Stefano, L.; Ajanthan, T.; Torr, P.H.S. Learning to Adapt for Stereo. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9653–9662. [CrossRef]
59. Huisman, M.; van Rijn, J.N.; Plaat, A. A Survey of Deep Meta-Learning. *Artif. Intell. Rev.* **2021**, *54*, 4483–4541. [CrossRef]
60. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571. [CrossRef]
61. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef] [PubMed]
62. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; p. 9626. [CrossRef]
63. Su, H.; Wei, S.; Yan, M.; Wang, C.; Shi, J.; Zhang, X. Object Detection and Instance Segmentation in Remote Sensing Imagery Based on Precise Mask R-CNN. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1454–1457. [CrossRef]
64. Cheng, G.; Zhou, P.; Han, J. Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
65. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
66. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [CrossRef]

67. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the European Conference on Computer Vision (ECCV), 13th European Conference, Zurich, Switzerland, 6–12 September 2014; pp. 740–755. [CrossRef]
68. Wada, K. Labelme: Image Polygonal Annotation with Python. Available online: https://github.com/wkentaro/labelme (accessed on 20 July 2018).
69. Bearman, A.; Russakovsky, O.; Ferrari, V.; Fei-Fei, L. What's the Point: Semantic Segmentation with Point Supervision. In Proceedings of the European Conference on Computer Vision (ECCV), 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 740–755. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.