remote sensing

MDPI

Article

Efficient Hybrid Supervision for Instance Segmentation in Aerial Images

Linwei Chen [1], Ying Fu [1,*], Shaodi You [2] and Hongzhe Liu [3]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China; chenlinwei@bit.edu.cn
[2] Computer Vision Research Group, Institute of Informatics, University of Amsterdam, 1098 XH Amsterdam, The Netherlands; s.you@uva.nl
[3] Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China; liuhongzhe@buu.edu.cn
* Correspondence: fuying@bit.edu.cn

Abstract: Instance segmentation in aerial images is of great significance for remote sensing applications, and it is inherently more challenging because of cluttered background, extremely dense and small objects, and objects with arbitrary orientations. Besides, current mainstream CNN-based methods often suffer from the trade-off between labeling cost and performance. To address these problems, we present a pipeline of hybrid supervision. In the pipeline, we design an ancillary segmentation model with the bounding box attention module and bounding box filter module. It is able to generate accurate pseudo pixel-wise labels from real-world aerial images for training any instance segmentation models. Specifically, bounding box attention module can effectively suppress the noise in cluttered background and improve the capability of segmenting small objects. Bounding box filter module works as a filter which removes the false positives caused by cluttered background and densely distributed objects. Our ancillary segmentation model can locate object pixel-wisely instead of relying on horizontal bounding box prediction, which has better adaptability to arbitrary oriented objects. Furthermore, oriented bounding box labels are utilized for handling arbitrary oriented objects. Experiments on iSAID dataset show that the proposed method can achieve comparable performance (32.1 *AP*) to fully supervised methods (33.9 *AP*), which is obviously higher than weakly supervised setting (26.5 *AP*), when using only 10% pixel-wise labels.

Keywords: hybrid supervision; instance segmentation; aerial images; labeling cost

Citation: Chen, L.; Fu, Y.; You, S.; Liu, H. Efficient Hybrid Supervision for Instance Segmentation in Aerial Images. *Remote Sens.* **2021**, *13*, 252. https://doi.org/10.3390/rs13020252

Received: 5 December 2020
Accepted: 6 January 2021
Published: 13 January 2021

1. Introduction

Instance segmentation in aerial images is an important task, which benefits various applications, e.g., monitoring of land changes [1], urban management [2] and traffic monitoring [3]. With the fast development of deep convolutional neural networks (CNN), the CNN-based instance segmentation methods are able to reach higher performance. However, the prerequisite for their performance is the availability of large-scale image dataset with accurate manually annotated labels. Current mainstream fully supervised instance segmentation methods [4–9] need instance-level pixel-wise labels for training. Specifically, labeling a bounding box on an object takes 10.2 s on average while labeling a segmentation annotation takes 79 s, which is about 8× slower [10]. Furthermore, aerial images usually have a wide range of view, which means they contain much more objects of interest than natural images. This leads to higher labeling cost. Taking the iSAID dataset [11] as an example, it is a large-scale aerial image dataset where each image has 233.6 instances on average. In comparison, there are only 2.8 instances per image in natural image dataset PASCAL VOC 2012 [12]. It means labeling cost of aerial images is two magnitude higher than natural images.

Although fully supervised instance segmentation methods [4–9] for natural images can be extended to aerial images, their labeling cost is unacceptable. Currently, weakly supervised instance segmentation methods that use economic image-level labels [13–15] or bounding box labels [16–18] have been proposed for natural scenes. To achieve better performance, some of them [16,17] use bounding box labels and rely on hand-crafted heuristics (e.g., Grabcut [19] and MCG [20]) to infer an object mask inside a bounding box as pixel-wise label for training. Nonetheless, their performance is far behind fully supervised methods due to inaccurate labels, and they cannot work well with aerial images because they do not consider challenges in aerial images which do not exist in natural scenes. Specifically, there are three key challenges in aerial images, as illustrated in Figure 1, first, cluttered background results in severe false positives, especially for small objects. Second, objects (e.g., vehicles) can be extremely small and dense, meanwhile, huge objects (e.g., ground track field) can cover a huge area. Third, while objects often appear with horizontal orientation in natural images, they can be with arbitrary orientations in aerial images. These challenges make it hard for hand-crafted heuristics (e.g., Grabcut [19] and MCG [20]) to obtain accurate object masks from aerial images for training.

In this paper, we aim to achieve satisfactory performance while keeping low labeling cost for aerial image instance segmentation. To this end, we present a pipeline of hybrid supervision that takes advantage of low labeling cost from bounding box labels and high accuracy from pixel-wise labels. It only uses 5–20% of images with instance-level pixel-wise labels and the rest of training images only have bounding box labels. The pipeline consists of an ancillary segmentation model and a primary instance segmentation model. The ancillary segmentation model is designed to generate accurate pseudo pixel-wise labels from real-world aerial images. A small portion of pixel-wise labels is enough for it to learn to recognize the shape of objects. After obtaining accurate pseudo pixel-wise labels with the help of the ancillary segmentation model, the primary instance segmentation model can be easily trained with large amounts of pseudo pixel-wise labels as well as a small portion of pixel-wise labels in a hybrid way and reach similar performance to fully supervised setting.

To address three key challenges in aerial images, we add two simple but effective modules to the ancillary segmentation model. The bounding box attention module is proposed to suppress the noise in clutter background and deals with densely distributed small objects, and the bounding box filter module is designed to suppress false positives caused by cluttered background. To handle the arbitrary oriented objects in aerial images, we adopt oriented bounding box instead of horizontal box to help these two modules to work better when arbitrary orientated objects show up.

In summary, our main contributions of this work are that

- The proposed method can well balance performance and labeling cost for instance segmentation in aerial images.
- We propose a pipeline that consists of an ancillary segmentation model and a primary instance segmentation model. The ancillary segmentation model with bounding box attention module, bounding box filter module and oriented bounding box labels can effectively address the specific challenges in aerial images, i.e., cluttered background, extremely dense and small objects, and objects with arbitrary orientations.
- We evaluate our method and achieve 32.1 AP on challenging iSAID dataset [11] using 10% pixel-wise labels, which is comparable to fully supervised method 33.9 AP and much better than weakly supervised setting 26.5 AP.

2. Related Work

In this section, we review segmentation methods in natural images and the segmentation methods in aerial images, and the progress of aerial image datasets.

Segmentation in natural images. Semantic segmentation is a dense prediction task that assigns each pixel a semantic category without identifying different instances. An early CNN-based method was proposed by Long et al. [21]. Furthermore, later works improve

performances by using multi-scale method [22–26], contextual information [27–30], and feature fusion [26,30].

Compared with semantic segmentation, instance segmentation further identifies each object instance in the same semantic category. Most mainstream methods use object proposals where objects are detected with candidate bounding boxes and then segmented with a binary mask. These methods can be further divided into two categories in terms of the proposal methods. One is two-stage object detection framework based [4–6,31], the other is one-stage object detection framework based [7–9,32]. The two-stage methods rely on two-stage detectors [33,34], which usually first employ region proposal techniques to obtain regions of interest, and then extract the features of the regions and obtain the predictions of categories, bounding boxes and shapes. The one-stage methods rely on one-stage detectors [35–38] which require only a single pass through the neural network and directly obtain the predictions.

The instance-level pixel-wise labels for segmentation is labor intensive. Therefore, some recent researchers exploit weakly supervised method, which only requires image-level labels [13–15] or bounding box labels [16–18]. Khoreva et al. [16] and Li et al. [17] use GrabCut [19] and MCG [20] to propose the pseudo pixel-wise labels of objects, and then refine them with iterative label refinement mechanism. Hsu et al. [18] learn a CNN-based model in an end-to-end fashion by using the bounding box tightness prior and multiple-instance learning. Although these weakly supervised methods require less labeling cost, their performance is much inferior to fully supervised methods [13–16,18].

Segmentation in aerial images. The state-of-the-art end-to-end aerial image semantic segmentation models are mostly inspired by the idea of fully convolutional networks [21], which generally consist of an encoder-decoder architecture [21]. Sherrah et al. [39] utilize a recurrent network in fully convolutional network which fuses multi-level features with boundary-aware features to achieve better inferences. Ghosh et al. [40] stack U-Nets architecture to merge high-resolution details and long distance context information at low-resolution image. Hamaguchi et al. [41] introduce local feature extraction module to aggregate local features with decreasing dilation factor.

As for instance segmentation, the relevant datasets for aerial images are less than natural image datasets, and related researches mainly focus on segmenting one particular type of object, e.g., vehicle [42] or ship [43]. Mou et al. [42] introduce a unified multi-task learning network that can simultaneously segment vehicle regions and detect semantic boundaries. Feng et al. [43] address dense object detection issue by applying a sequence of dilation convolution blocks to progressively learn multi-scale context information and avoid confusion between objects of the same class. To our best knowledge, there is still no existing studies for weakly supervised aerial image instance segmentation on multiple types of objects.

Aerial image datasets. Recently, some well-annotated aerial image datasets for object detection [44,45] and semantic segmentation [46,47] have been introduced, which encourage the advancements in aerial images for earth observation. However, these datasets do not provide accurate pixel-wise labels for each object instance in an aerial image, so they are not suitable for instance segmentation task. There does exist a few publicly available instance segmentation datasets [11,48], but some of them typically focus on a single object category, e.g., [48] only labels building footprints. Currently, the only aerial image dataset with instance-level pixel-wise labels of multiple categories is iSAID [11]. It contains annotations for 655,451 instances of 15 important categories in 2806 high spatial resolution images. Moreover, iSAID dataset [11] exhibits the following distinctive characteristics: (1) images were collected from multiple sensors and platforms, scenes in these images are varying and have complex contextual information; (2) it has huge object scale variation, the small, medium and large objects, often show in the same image; (3) it depicts real-life aerial conditions, the distribution of objects is imbalanced and uneven, the orientation of objects are arbitrary. All of these characteristics make the instance segmentation task on iSAID dataset [11] challenging.

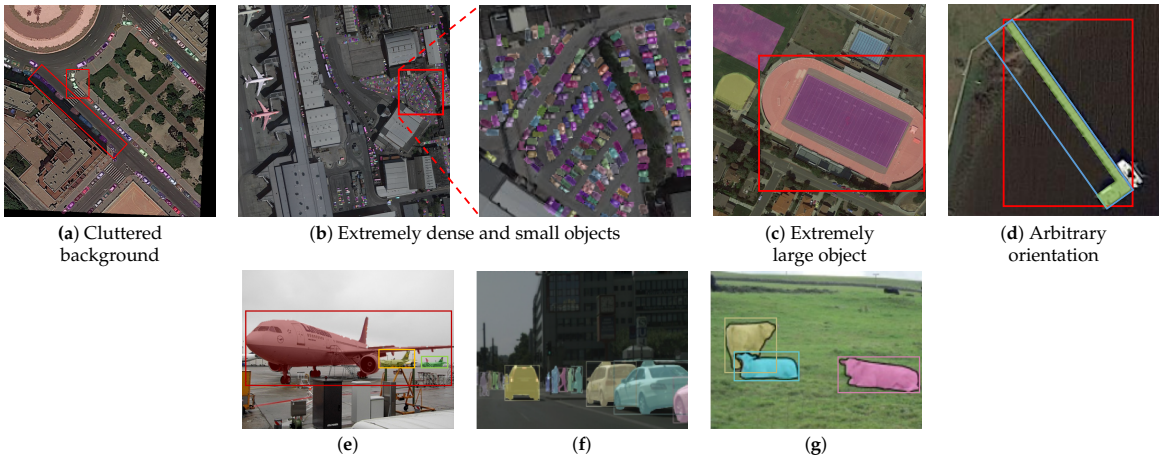
3. Hybrid Supervision for Instance Segmentation in Aerial Images

In this section, we first formulate the instance segmentation task in aerial images and provide more insights on the choice of label types and three challenges. Then, our proposed method and the implementation details are introduced.

3.1. Motivation

In this paper, we focus on instance segmentation in aerial images that locates the objects of interest (Figure 1a–d, e.g., aeroplanes, vehicles and harbors, etc.) with pixel-level accuracy.

Our goal is to balance performance and labeling cost. To this end, we analyze three types of labels that are commonly used and we also analyze the key challenges in aerial images that do not exist in natural images.



(**a**) Cluttered background (**b**) Extremely dense and small objects (**c**) Extremely large object (**d**) Arbitrary orientation (**e**) (**f**) (**g**)

Figure 1. Challenges in aerial images. (**a**–**d**) are aerial images from the iSAID Dataset [11]. (**a**) Cluttered background. A line of cars are in the shadow of the buildings and the white cars are surrounded by white zebra crossings. (**b**) Small vehicles that are extremely densely distributed. (**c**) The extremely large ground track field. (**d**) Arbitrary oriented harbor. (**e**–**g**) are natural images from PASCAL VOC 2012 [49], Cityscapes [50] and COCO [51] respectively, where objects are much less and have a similar size.

Choice of label types. To reduce labeling cost of instance segmentation in aerial images, it is natural to use economic image-level labels or bounding box labels to replace expensive instance-level pixel-wise labels. However, uncertain locations of objects in image-level labels and uncertain shapes of objects in bounding box labels definitely harm the learning of instance segmentation model and lead to inferior performance. In order to solve this problem, as shown in Figure 2, we have carefully analyzed three types of labels which are commonly used.

(**a**) Image (**b**) Image-level label (**c**) Bounding box label (**d**) Instance-level pixel-wise label

Figure 2. The example of three types of labels.

- Image-level labels only provide the information about categories of objects in images and cannot indicate specific location of each object. They are usually used for image classification [52] but can be hardly used for instance segmentation task in aerial images especially when objects of interest are small and densely distributed.
- Bounding box labels can provide the information about categories, and locations of objects. They are usually used for object detection [12]. Nonetheless, they do not contain the information about shapes of objects, which are important for instance segmentation task.
- Instance-level pixel-wise labels contain rich information about category, location and shape of each interested object. But they are expensive to obtain. They are usually used for instance segmentation [12,50,51].

According to these analyses, the combination of a few pixel-wise labeled samples and a dominant majority of bounding box labeled samples is an optimal choice to reduce labeling cost while keeping satisfactory performance. Using dominant majority of bounding box labeled samples can save $\sim 7\times$ labeling cost compared with pixel-wise labels [10], and the usage of a few pixel-wise labeled samples can provide the knowledge about shapes of objects, which can be beneficial to obtain the shape from bounding box labels and achieve better performance.

Key challenges. While weakly supervised instance segmentation in natural images has been well exploited [13–16,18], those methods cannot be naively adopted to aerial images, due to three challenges of weakly supervised instance segmentation in aerial images.

- Cluttered background. Aerial images can cover various scenes rather than specific scenes, e.g., cities, oceans and field. Furthermore, other factors like trees and shadows of buildings can also disturb the detection and segmentation. Therefore, the background (area without interested objects) can be highly diverse and cause false positives easily. Taking Figure 1a as an example, a line of cars in the shadow of the buildings are easy to ignore, and the shape of white cars surrounded by zebra crossings are difficult to obtain accurately.
- Extremely dense and small objects. Aerial images are taken from a much longer distance than natural images, which results in an extremely dense distribution of small objects. For example, as shown in Figure 1b, many small vehicles are concentrated in specific area, sizes of these objects are smaller than 10 pixels in the aerial image. At the same time, there also exists extremely large objects, as shown in Figure 1c, making object detection more complex and challenging.
- Arbitrary object orientation. In contrast to conventional datasets for instance segmentation [12,50,51], where objects are generally oriented upward due to gravity, the orientation of object in aerial images is arbitrary as shown in Figure 1d.

3.2. Formulation

The overview of our proposed method is illustrated in Figure 3. It adopts a small portion of pixel-wise labeled samples (i.e., fully labeled images) and a dominant amount of bounding box labeled samples (i.e., weakly labeled images).

In this part, we first introduce the idea of hybrid supervision, then the design of ancillary segmentation model and how we address the challenges in aerial images are described. As for instance segmentation model in our pipeline, we use vanilla Mask R-CNN [4] and CenterMask [8] for experiments separately, please refer to [4,8] for more details. Notice that they work independently after training and can be replaced by any instance segmentation models [5,6].

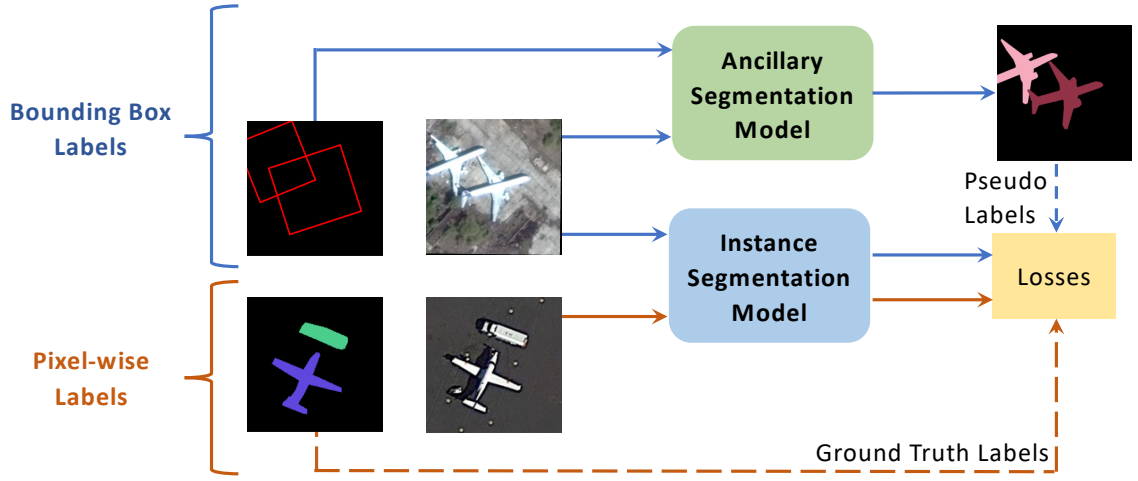


Figure 3. The overview of our pipeline. The ancillary segmentation model is first trained with a small portion of pixel-wise labeled samples, so as to learn to predict high quality pseudo pixel-wise labels on the weakly labeled samples. Then, the instance segmentation model is optimized with the combination of pseudo labels generated by the ancillary segmentation model and the pixel-wise labeled samples. Notice that the instance segmentation model works independently after training.

Hybrid supervision. As shown in Figure 3, our method uses both fully labeled and weakly labeled images for learning in a hybrid way. Our work differs from previous methods [16–18,53] in four significant aspects:

Firstly, there have been some existing works [17,54,55] that provide experiment results on utilizing both fully labeled and weakly labeled images for learning. Nevertheless, [54,55] can only deal with semantic segmentation task. [17] simply uses the GrabCut [19] and MCG [20] to obtain pseudo pixel-wise labels from bounding boxes and cannot utilize fully labeled images to refine the pseudo labels which limits its performance. We replace hand-crafted methods [19,20] and iterative mechanism [16,17] with a CNN-based ancillary segmentation model for extracting high quality pseudo instance-level pixel-wise labels on weakly labeled images. The pseudo labels we obtain can be seen as a hybrid of knowledge about shapes of objects from fully labeled images and the ground truth information about locations of objects from weakly labeled images, which provide richer information than original bounding box labels and help to reach better performance.

Secondly, weak supervision [16,18] uses only weakly labeled images but suffers from uncertain shapes of objects in labels. We utilize a small portion of fully labeled images to provide the knowledge about shapes of objects, avoiding the problem of uncertain shapes in weak supervision [16,18].

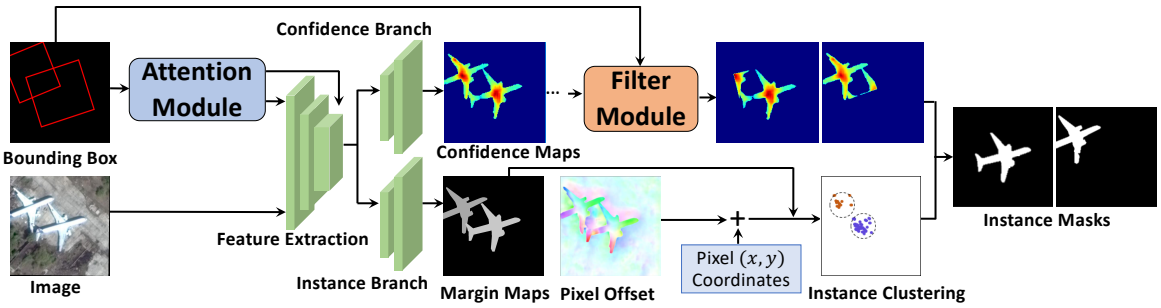
Thirdly, semi-supervision [53] utilizes fully labeled images and predicts pseudo labels on unlabeled images for learning, however, noise in pseudo labels (e.g., false positives in cluttered background, objects with wrong category labels) harms the learning of models. In our pipeline, the ancillary segmentation model is able to fully use the bounding box labels to suppress the noise in pseudo labels.

Finally, previous works [16–18,53] only consider natural images, our method is carefully designed to deal with the challenges in aerial image instance segmentation, i.e., cluttered background, extremely dense and small objects, and objects with arbitrary orientations, which do not exist in natural images.

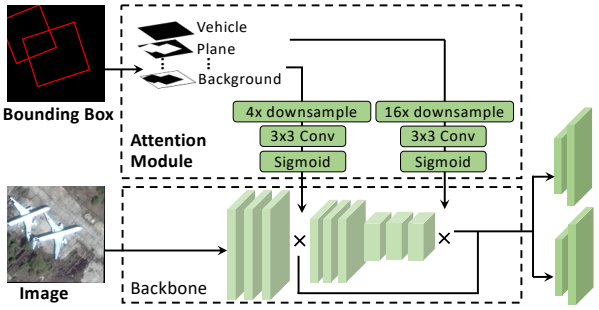
Ancillary segmentation model. To implement the method of hybrid supervision, we design an ancillary segmentation model based on DeepLabv3+ [22], which is a reliable semantic segmentation network. The spatial-invariant property of fully convolutional networks makes DeepLabv3+ [22] unable to distinguish different instances that distribute in different location. To adopt it to instance segmentation, we modify DeepLabv3+ [22] into two decoder branches and utilize spatial-embedding loss introduced by Neven et al. [56] for training. The spatial-embedding loss [56] avoids the problem of spatial-invariance

by assigning each pixel a spatial coordinate and learning position-relative offset vectors. Therefore, the resulting combination of pixel coordinate and offset vector points to its corresponding instance center.

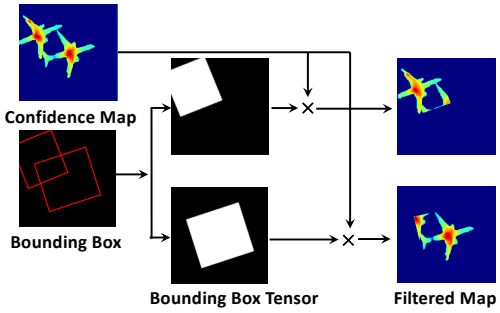
As shown in Figure 4a, after extracting features by backbone Xception-65 in DeepLabv3+ [22], two decoder branches are followed. The confidence branch decoder consists of a deconvolutional layer, a residual block [57], a deconvolutional layer and a sigmoid layer. Furthermore, the instance branch decoder is similar, except that margin maps need no sigmoid layer and pixel offset maps need tanh layer to predict offset value in $[-1, 1]$. Notice that the difference in coordinate between two neighboring pixels is $1/800$, both in x and y direction, so each pixel can point at most 800 pixels away. Deconvolutional layers in decoders are for $2\times$ upsampling and the first deconvolutional layer is followed by a batch normalization layer and a ReLU layer.



(**a**) Overview of ancillary segmentation model



(**b**) Bounding box attention module



(**c**) Bounding box filter module

Figure 4. The illustration of ancillary segmentation model. (**a**) Pixel offset map is color-encoding where the color indicates the angle of the vector and the saturation indicates the distance to the center. The "+" in (**a**) indicates pixel-wise addition. The "×" in (**b**,**c**) indicates element-wise multiplication.

The confidence branch is used to obtain confidence map for each category. The pixel i is more likely to be foreground if its confidence value d_i is close to 1, and pixel with $d_i \leq 0.5$ is regarded as background. The instance branch predicts pixel offset maps and margin maps. Pixel offset value (o_{ix}, o_{iy}) in pixel offset maps is used to obtain the predicted center (c_{ix}, c_{iy})

$$\begin{aligned} c_{ix} &= o_{ix} + e_{ix}, \\ c_{iy} &= o_{iy} + e_{iy}, \end{aligned} \tag{1}$$

where (e_{ix}, e_{iy}) is the coordinates of pixel. The predicted centers of pixels that belong to the same instance should be close to each other, which can be recognized by clustering. Considering that different objects have different size and shape, they need object-specific clustering margin. Otherwise, if clustering margin is kept the same for all objects, two small objects that are next to each other may be clustered into one object, since margin is relative large for them. Furthermore, big objects may be clustered into more than one objects, because pixels far away from the center may not be able to point into this small

region around the center. To handle this, margin maps are learnt for predicting instance margin values $(\sigma_{ix}, \sigma_{iy})$ in (x, y) direction for each object, which are adapted to size and shape of each object. The confidence branch and the instance branch are optimized jointly to achieve best performance as described in Section 3.3.

During inference, we sequentially cluster the foreground pixels (whose confidence value $d_i > 0.5$) to different instance objects for each category-specific confidence map. The cluster procedure is first to choose the pixel with the highest confidence value in confidence map and then use corresponding predicted center as the center of instance S_k $(\hat{C}_{kx}, \hat{C}_{ky})$. The corresponding instance margin values $(\hat{\sigma}_{kx}, \hat{\sigma}_{ky})$ are also kept. By using this center and accompanying margin, we cluster the i-th foreground pixel into instance S_k, if the i-th predicted center (c_{ix}, c_{iy}) is close to S_k, the distance is measured by gaussian function,

$$i \in S_k \Longleftrightarrow \exp(-\frac{(c_{ix} - \hat{C}_{kx})^2}{2\hat{\sigma}_{kx}^2} - \frac{(c_{iy} - \hat{C}_{ky})^2}{2\hat{\sigma}_{ky}^2}) > 0.5. \quad (2)$$

The spatial-embedding loss [56] is an excellent work designed for predicting high resolution instance segmentation results in urban street scenes. We utilize it for helping our ancillary segmentation model to predict pseudo pixel-wise labels in high resolution and instance-level, which are important for the learning of primary instance segmentation model. However, it is not enough to deal with the challenges in aerial images and the task of generating pseudo pixel-wise labels. To make our ancillary segmentation model more robust for extracting pseudo labels from aerial images, we add a bounding box attention module and bounding box filter module to it, and the usage of an oriented bounding box helps these two modules work better when arbitrary orientated objects show up.

3.3. Design and Learning Details

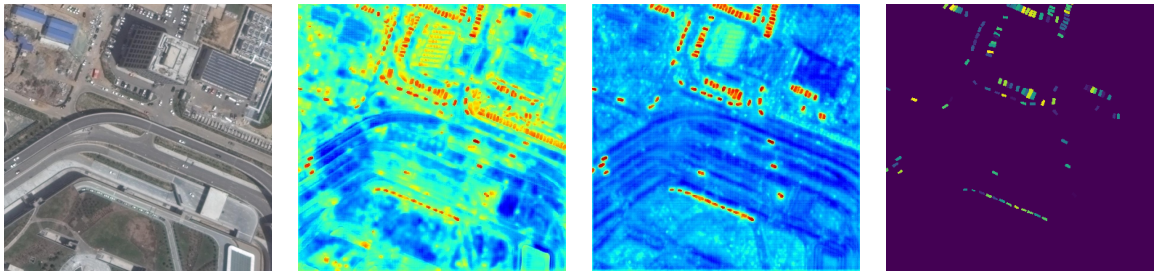
Bounding box attention module. In our analysis, there are two main obstacles in predicting accurate pseudo pixel-wise labels for small objects, insufficient object feature information and cluttered background. Small objects lose most of their feature information in deep layers due to the use of the pooling layer, which makes it hard to localize them. Meanwhile, cluttered backgrounds may introduce false positives due to its similarity with foreground objects.

We notice that bounding box labels contain the information about locations of small objects, and they also provide the information about the background. Therefore, we regard the bounding box labels as feature maps containing localization information and present a bounding box attention module for encoding bounding boxes to attention maps, as shown in Figure 4b. Though its structure is very simple, it can fully utilize the information of bounding boxes to adjust the feature maps of backbone, e.g., highlight the features of small objects within bounding boxes and decay the features in the area of background.

In detail, the bounding box attention module first converts the bounding box labels to feature maps with $N + 1$ channels, where N is the number of categories and 1 represents background. If a given pixel belongs to a bounding box of specific class, its corresponding category channel is set to 1 and the background channel is set to 0. Please notice that a pixel can belong to multiple bounding boxes. If a given pixel does not belong to any bounding boxes, the background channel is set to 1 and other channels are set to 0. Then the bounding box attention module converts the bounding box feature maps to attention maps which are fused with the feature maps of backbone Xception-65 [22] using element-wise multiplication. Moreover, the attention maps are generated in $4\times$ downsampling scale and $16\times$ downsampling scale, which are adapted to both small objects and large objects.

Figure 5 visualizes a confidence map for small vehicles, due to the complexity of real-world aerial images, excessive noise can overwhelm the object information. As shown in Figure 5b, some small objects have relative low confidence score and cluttered background introduce severe false positives. After we add the bounding box attention module to

ancillary segmentation model, in Figure 5c, most of small objects are clearly located and noise caused by cluttered background are largely suppressed.

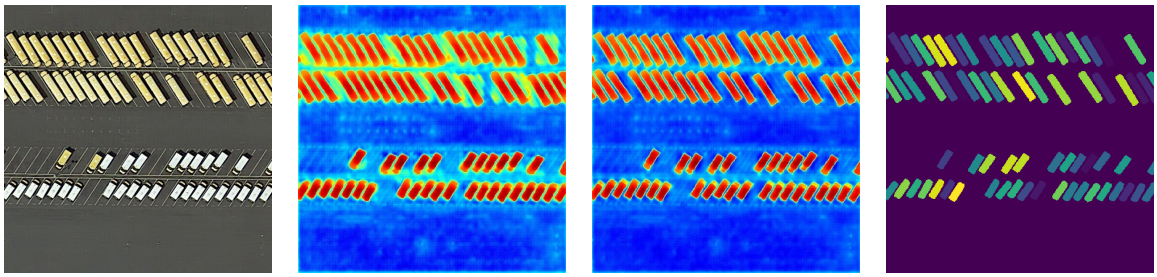


(**a**) Image (**b**) Baseline (**c**) +Bounding box attention module (**d**) Ground truth

Figure 5. Visualization of bounding box attention module. The (**b**,**c**) are visualized confidence map for small vehicles. It can be seen in (**c**) that false positives are largely suppressed and most of small vehicles in image are clearly identified with the help of bounding box attention module.

Bounding box filter module. The original clustering procedure in [56] easily leads to inaccurate segmentation results when objects are densely distributed or background is cluttered. For example, two objects that are very close can be clustered as one object, and cluttered background may lead to false positives. To address this problem and further improve the quality of generated pseudo pixel-wise labels, we introduce the bounding box filter module. As shown in Figure 4c, it utilizes the bounding box labels to output confidence map for each object sequentially before clustering procedure. For each object, it separates objects and filters out noise of confidence map in area outside the bounding box, so as to prevent the situation that two objects are clustered as one. Furthermore, background area outside bounding box is not involved in clustering procedure, so the false positives are avoided. After clustering procedure, we choose the instance mask that fits the bounding box best as optimal pseudo pixel-wise label of object. We view the pseudo pixel-wise label along with corresponding bounding box label as ground truth for the training of primary instance segmentation model.

Usage of oriented bounding box. Our ancillary segmentation model is segmentation-based and do not rely on horizontal bounding box prediction to locate object. Therefore, it has better adaptability to arbitrary oriented objects. For better dealing with arbitrary oriented objects, we adopt oriented bounding box inspired from [44,58,59], which tackle objects with arbitrary orientations in detection task. Notice that we use it for solving instance segmentation in aerial images with hybrid supervision, which is more challenging and not a naive extension. Moreover, we do not use oriented bounding box labels directly for training, they are utilized for obtaining pseudo pixel-wise labels with our ancillary segmentation model. Compared with horizontal bounding box, oriented bounding box can fit object with arbitrary orientation better. It helps bounding box attention module to generate more accurate attention maps (see Figure 6), which leads to better results. Furthermore, the bounding box filter module can better separate objects whose orientations are arbitrary. Therefore, our ancillary segmentation is able to obtain more accurate shape of object and this leads to better performance.



(**a**) Image (**b**) Using horizontal bounding box (**c**) Using oriental bounding box (**d**) Ground truth

Figure 6. Visualization of bounding box attention modules with horizontal and oriental bounding box. The (**b**,**c**) are visualized confidence map for large vehicles. It can be seen in (**c**) that arbitrary oriented objects are better identified when the bounding box attention module takes oriental bounding box as input.

Loss function. During training, we utilize aforementioned spatial-embedding loss [56] for learning. Specifically, the confidence branch is optimized with pixel-wise L2 loss $\mathcal{L}_{conf}$, and ϕ_k is a gaussian function representing the distance to instance S_k with probability

$$\mathcal{L}_{conf} = \frac{1}{M}\sum_{i}^{M} \mathbb{1}_{\{i \in S_k\}} \|d_i - \phi_k(c_{ix}, c_{iy})\|^2 + \mathbb{1}_{\{i \in bg\}} \|d_i - 0\|^2, \tag{3}$$

$$\phi_k(c_{ix}, c_{iy}) = \exp(-\frac{(c_{ix} - C_{kx})^2}{2\sigma_{kx}^2} - \frac{(c_{iy} - C_{ky})^2}{2\sigma_{ky}^2}), \tag{4}$$

where M is the number of pixels, $\mathbb{1}$ is indicator function, d_i is confidence value and bg represents background. $\phi_k(c_{ix}, c_{iy})$ usually has higher value if pixel i is close to the center of object. Therefore higher value in confidence maps indicates that the corresponding pixel is closer to a center of object. The (C_{kx}, C_{ky}) is the center of instance object S_k

$$C_{kx} = \frac{1}{|S_k|}\sum_{i \in S_k} c_{ix}, \quad C_{ky} = \frac{1}{|S_k|}\sum_{i \in S_k} c_{iy}, \tag{5}$$

and $(\sigma_{kx}, \sigma_{ky})$ is defined as

$$\sigma_{kx} = \frac{1}{|S_k|}\sum_{i \in S_k} \sigma_{ix}, \quad \sigma_{ky} = \frac{1}{|S_k|}\sum_{i \in S_k} \sigma_{iy}. \tag{6}$$

The instance branch can be optimized with cross entropy loss

$$\mathcal{L}_{inst} = \frac{1}{M}\sum_{i}^{M} \mathbb{1}_{\{i \in S_k\}}(-\log \phi_k(c_{ix}, c_{iy})) + \mathbb{1}_{\{i \in bg\}}(-\log(1 - \phi_k(c_{ix}, c_{iy}))). \tag{7}$$

Furthermore, to ensure that $(\sigma_{ix}, \sigma_{iy})$ is close to $(\sigma_{kx}, \sigma_{ky})$, a smoothness term is added

$$\mathcal{L}_{smooth} = \frac{1}{|S_k|}\sum_{i \in S_k} \|\sigma_{ix} - \sigma_{kx}\|^2 + \|\sigma_{iy} - \sigma_{ky}\|^2. \tag{8}$$

So the total loss function is

$$\mathcal{L}_{total} = \lambda_{conf}\mathcal{L}_{conf} + \lambda_{inst}\mathcal{L}_{inst} + \lambda_{smooth}\mathcal{L}_{smooth}, \tag{9}$$

where λ_{conf}, λ_{inst}, λ_{smooth} are the constant coefficient, we choose $\lambda_{conf} = 1$, $\lambda_{inst} = 10$, $\lambda_{smooth} = 1$ for balancing three losses.

4. Experimental Results

Our method is systematically evaluated on the challenging iSAID dataset [11]. In this section, the dataset and evaluation metric are first introduced. Then, we describe the implement details. After that, we quantitatively and qualitatively evaluate our method. Finally, the ablation studies are performed to analyze the proposed modules.

4.1. Dataset and Evaluation Metric

Dataset. The iSAID dataset [11] contains 2806 original high spatial resolution images. These images are collected from multiple sensors and platforms with multiple resolutions. The original spatial resolution ranges from ∼800 × 800 to ∼4000 × 13,000. The predefined training set consists of 1411 images, while validation set contains 458 images and test set has 937 images. For instance segmentation task, the iSAID dataset provides 655,451 instances annotations over 15 different categories of object, which is the largest dataset for instance segmentation in high spatial resolution remote sensing images. As the official evaluation server of iSAID is still improving, the results on testing set are unavailable, we evaluate our method on iSAID validation set in the following.

Evaluation metric. We use the standard COCO [51] metrics, i.e., AP (averaged over IoU threshold from 0.5 to 0.95 with stride of 0.05), AP_{50} (IoU threshold is 0.5), AP_{75} (IoU threshold is 0.75), AP_S, AP_M and AP_L, where S, M and L represent small (area: 10–144 pixels), medium (area: 144 to 1024 pixels) and large objects (area: 1024 pixels and above), respectively.

4.2. Implementation Details

There are three steps in the whole training procedure. First, we train the ancillary segmentation model with a small portion of images with instance-level pixel-wise labels. Notice that both horizontal and oriented bounding box labels can be obtained from pixel-wise labels by using extreme points or algorithm of finding minimum area rectangle, which needs no extra labeling cost. We input both bounding box labels and images to ancillary segmentation model and take pixel-wise labels as the supervisory signal. Second, the ancillary segmentation model generates high quality pseudo instance-level pixel-wise labels from bounding box labels on weakly labeled images. Third, the primary instance segmentation model can be trained with a combination of images with instance-level pixel-wise labels and images with generated pseudo instance-level pixel-wise labels. Here, we describe the implementation details of ancillary segmentation model and instance segmentation model in Figure 3.

Training procedure of ancillary segmentation model. For optimizing intersection-over-union of each instance, we follow [56] to use Lovász-hinge loss [60] rather than standard cross entropy loss in $\mathcal{L}_{inst}$. The backbone of the ancillary segmentation model is based on Xception-65 [22], which is pre-trained on the ImageNet dataset [52]. We first pre-train our model on 448 × 448 crops, which are taken out of the original 800 × 800 training images. Notice that each 448 × 448 image patch is centered around an object. In this way, we avoid spending too much computation time on background images patches without any instances, which leads to shorter training time. The training iterations and batch-size for pre-training are 40 k and 8, respectively.

Then, we finetune the ancillary segmentation model for another 20 k iterations on 800 × 800 crops with a batch-size of 2 to increase performance on the bigger objects which cannot fit completely within the 448 × 448 image patch. The batch normalization statistics are kept fixed during this stage for better convergence. We use the ADAM optimizer [61] and polynomial learning rate decay $(1 - \frac{iter}{max\ iter})^{0.9}$. The initial learning rate is 5×10^{-4}, which is later decreased to 5×10^{-5} for finetuning.

The ancillary segmentation model is optimized on two NVIDIA GeForce GTX 1080 Ti GPUs for roughly two days. Next to random cropping, we also apply random horizontal mirroring and vertical mirroring as data-augmentation.

Training procedure of instance segmentation model. We use off-the-shelf Mask R-CNN [4] and CenterMask [8] as our primary instance segmentation model, and the implementation is based on *Detectron2*. The training data consists of all bounding box labels, a small portion of pixel-wise labels and pseudo labels generated by our ancillary segmentation model.

In our implementation, Mask R-CNN [4] adopts ResNet-101 [57] as the backbone network and CenterMask [8] adopts ResNeXt-101 [62] as the backbone network. For the training setting, ResNet-101 [57] and ResNeXt-101 are both pre-trained on the ImageNet dataset [52]. The batch size, learning rate, weight decay, momentum, and the number of the iterations are set to 8, 1×10^{-2}, 1×10^{-4}, 0.9 and 180 k, respectively. The learning rate drops to 1×10^{-3} at the 60k-th iteration and 1×10^{-4} at the 120k-th iteration. We use ADAM [61] as the optimizer. For data augmentation, we use the same settings as those in Mask R-CNN [4]. Both Mask R-CNN [4] and CenterMask [8] are optimized on a server with eight NVIDIA GeForce GTX 1080 Ti GPUs.

Methods for comparison. To verify the effectiveness of the proposed method, we provide results under different label settings as follow:

- Weak supervision. Considering there exists no weakly supervised method for instance segmentation in aerial images, we use bounding box labels as pseudo instance-level pixel-wise labels for training, so as to serve as weakly supervised results.
- Full supervision. With all pixel-wise labels available, the full supervision can easily reach the best result. For a comprehensive comparison, we provide the fully supervised results with different percentage of pixel-wise labels available (i.e., 5%, 10%, 20% and 100%).
- Weak and full supervision. For a fair comparison, we provide the results of weak and full supervision. Under this setting, the instance segmentation model is trained with both instance-level pixel-wise labels and bounding box labels, notice that bounding box labels are utilized to train the bounding box branch and classification branch only.

Overall performance. As illustrated in Tables 1 and 2, with 10% of pixel-wise labels and Mask R-CNN [4] serves as instance segmentation model, our hybrid supervision largely suppress the weak supervision (31.2 AP vs. 13.3 AP, 32.1 AP vs. 26.5 AP) and full supervision with 10% of pixel-wise labels available (32.1 AP vs. 22.3 AP). Compared with weak and full supervision setting that uses the same percentage of pixel-wise labels and bounding box labels, we achieve 3.5 AP higher result (32.1 AP vs. 28.6 AP), which is a strong evidence to prove the effectiveness of our proposed method. Furthermore, the AP_{50} result of our method is very close to full supervision with 100% pixel-wise labels available (55.0 AP_{50} vs. 56.4 AP_{50}, i.e., 97.5%). Even with the most strict metric AP, our hybrid supervision achieves the 94.7% of performance of full supervision (32.1 AP vs. 33.9 AP). It means that our method can achieve similar performance to 100% fully supervised method with only 10% of pixel-wise labels available.

Besides, we provide the results of Mask R-CNN [4] with 5% and 20% of pixel-wise labels. As shown in Table 1, our method still performs well with only 5% of pixel-wise labels, and when pixel-wise labels increase to 20%, our result is even closer to full supervision (33.3 AP vs. 33.9 AP). These results show that our hybrid supervision is adaptive to the percentage of pixel-wise labels.

Table 1. AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L results on iSAID dataset [11].

Supervision	Instance Segmentation Model	Number of Pixel-Wise Labels	Number of Horizontal Bounding Box Labels	Number of Oriented Bounding Box Labels	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Weak	Mask R-CNN [4]	-	1411 (100%)	-	13.3	34.7	7.2	8.5	14.8	18.5
		-	-	1411 (100%)	26.5	47.0	26.2	14.8	32.1	36.0
	CenterMask [8]	-	1411 (100%)	-	12.6	33.0	7.2	7.4	14.5	18.4
		-	-	1411 (100%)	26.1	45.8	25.6	12.5	31.9	39.6
Full	Mask R-CNN [4]	70 (5%)	-	-	15.7	25.3	17.0	5.9	19.9	22.6
		141 (10%)	-	-	22.5	38.8	22.5	10.6	28.7	30.2
		282 (20%)	-	-	26.5	43.8	33.6	13.0	33.6	34.8
		1411 (100%)	-	-	33.9	56.4	35.7	18.8	41.1	50.1
	CenterMask [8]	70 (5%)	-	-	15.2	24.9	16.1	5.2	20.5	22.8
		141 (10%)	-	-	23.3	39.2	23.9	10.3	30.4	28.9
		282 (20%)	-	-	26.8	44.5	28.0	11.7	35.1	37.3
		1411 (100%)	-	-	34.0	56.3	35.8	18.0	42.8	53.8
Weak+Full	Mask R-CNN [4]	141 (10%)	1270 (90%)	-	28.6	49.1	28.9	15.5	34.2	40.7
	CenterMask [8]	141 (10%)	1270 (90%)	-	28.3	48.8	28.1	14.5	33.9	40.5
Hybrid(Ours)	Mask R-CNN [4]	70 (5%)	-	1341 (95%)	28.7	51.1	27.8	16.4	34.1	35.2
		141 (10%)	1270 (90%)	-	31.2	53.6	31.9	17.5	37.5	38.9
		141 (10%)	-	1270 (90%)	32.1	55.0	32.4	18.5	39.0	42.2
		242 (20%)	-	1169 (80%)	33.3	55.5	34.6	18.7	40.5	47.4
	CenterMask [8]	70 (5%)	-	1341 (95%)	29.3	51.4	28.8	15.2	36.5	45.6
		141 (10%)	1270 (90%)	-	31.0	52.9	31.1	16.2	38.6	47.9
		141 (10%)	-	1270 (90%)	31.5	54.0	31.2	17.0	39.2	48.0
		242 (20%)	-	1169 (80%)	32.2	54.2	32.1	17.2	40.1	49.1

Table 2. AP results for each category on iSAID dataset [11]. The asterisk "*" indicates using horizontal bounding box labels instead of oriented bounding box labels.

Method	Ship	Storage Tank	Baseball Diamond	Tennis Court	Basketball Court	Ground Field Track	Bridge	Large Vehicle	Small Vehicle	Heli-Copter	Swimming Pool	Round-about	Soccer-Ball Field	Plane	Harbor	AP
Mask R-CNN [4]																
Weak *	7.23	26.37	30.41	24.08	17.52	14.73	3.24	4.59	6.27	0.20	10.72	29.35	23.37	0.12	0.99	13.28
Weak	31.08	30.93	43.59	74.02	33.20	22.11	18.98	33.31	12.60	0.43	24.51	30.20	38.46	0.43	4.27	26.54
Full (5%)	11.68	21.80	32.09	54.81	4.16	10.21	2.01	23.23	7.70	0.00	0.22	15.88	19.18	31.29	0.75	15.67
Full (10%)	28.10	21.86	36.32	65.88	10.55	13.17	12.19	26.33	8.07	4.32	18.20	15.68	22.31	39.95	13.91	22.46
Full (20%)	33.38	25.75	40.13	68.82	17.96	16.46	14.52	30.33	12.25	5.65	24.14	18.22	29.75	42.84	17.77	26.53
Full (100%)	39.70	34.64	46.73	75.46	34.61	29.83	17.34	34.81	13.10	6.09	29.33	28.07	42.24	49.68	26.84	33.90
Weak+Full * (10%)	30.32	31.98	41.87	69.29	31.12	23.79	17.45	31.39	11.61	2.33	24.81	25.06	34.54	38.57	14.10	28.55
Ours (5%)	33.39	32.71	49.37	73.72	24.36	19.82	14.68	30.87	12.14	3.00	29.97	24.25	35.62	32.79	13.15	28.67
Ours * (10%)	36.58	34.16	49.40	72.64	26.01	22.85	18.55	31.66	12.49	2.37	28.87	25.60	37.24	45.99	23.55	31.20
Ours (10%)	37.66	34.77	49.71	73.53	29.18	23.88	19.01	32.01	12.93	4.17	29.13	27.05	40.63	44.66	23.59	32.13
Ours (20%)	38.05	35.51	48.12	76.82	31.59	25.36	19.36	35.20	13.36	5.08	30.10	31.38	42.87	43.23	23.21	33.28
CenterMask [8]																
Weak *	5.15	25.84	31.95	24.32	18.66	12.09	2.70	3.44	3.74	0.27	10.20	26.51	23.63	0.09	0.80	12.63
Weak	23.41	29.50	44.53	75.41	42.79	18.82	18.25	29.68	9.28	0.69	24.58	29.66	40.14	0.43	4.18	26.09
Full (5%)	9.87	24.69	27.09	57.13	8.03	9.50	2.45	17.40	5.83	0.00	0.20	18.18	16.49	30.73	0.72	15.22
Full (10%)	22.83	25.86	38.53	68.82	16.49	13.33	13.87	23.22	6.87	3.44	20.08	17.83	21.50	40.56	16.55	23.32
Full (20%)	24.80	28.42	42.48	69.81	24.77	17.39	16.02	27.20	8.17	3.69	24.27	23.01	31.02	42.71	18.13	26.79
Full (100%)	30.23	35.19	51.53	77.16	42.36	24.50	19.97	32.01	10.06	5.29	30.34	32.61	44.89	47.93	26.30	34.02
Weak+Full * (10%)	23.66	28.45	45.63	71.95	34.47	19.39	18.76	27.64	8.23	2.25	25.20	26.17	38.06	38.28	17.57	28.38
Ours (5%)	26.73	33.24	49.34	74.52	38.13	18.87	16.68	30.31	9.11	3.23	27.02	33.50	40.71	27.98	10.52	29.33
Ours * (10%)	27.90	33.36	49.58	76.02	39.26	18.27	17.69	30.76	9.28	4.64	27.87	32.45	40.08	37.97	19.15	30.95
Ours (10%)	28.32	33.80	49.52	76.51	40.19	20.10	18.64	31.26	9.68	4.32	28.10	33.63	42.04	36.97	19.57	31.51
Ours (20%)	28.48	33.97	50.10	76.67	42.11	21.47	18.74	31.61	9.72	4.87	29.62	34.18	43.93	37.58	20.61	32.22

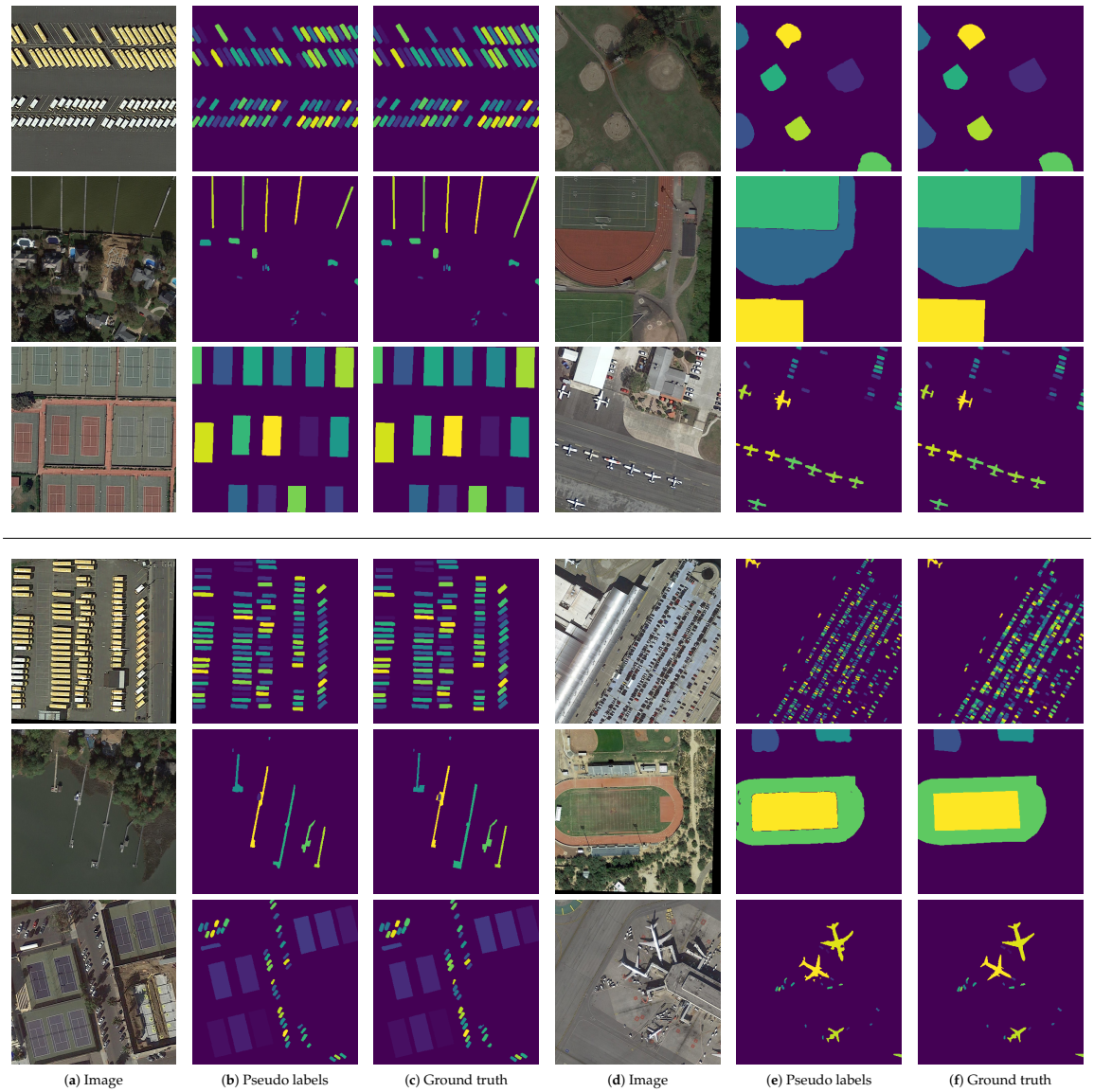
When we switch the instance segmentation model to CenterMask [4], our method with 5%, 10% and 20% pixel-wise labels also outperforms the weak supervision and achieves satisfactory performance compared with 100% full supervision, which shows the stability of our method.

Performance on small and dense objects. We further evaluate performance on small and dense objects, for it is a key challenge in aerial images. With only 10% of pixel-wise labels, our method with Mask R-CNN [4] can effectively save labeling cost and achieve similar

accuracy to 100% fully supervised result in *APs*, i.e., 18.5 *APs* vs. 18.8 *APs*, while the weakly supervised method can only achieve 14.8 *APs*. These results show that our method works well with small objects, and this conclusion keeps the same when we switch the instance segmentation model to CenterMask [8] (17.0 *APs* vs. 18.0 *APs*).

4.3. Qualitative Evaluation

Pseudo pixel-wise labels. We show some pseudo labels generated by our ancillary segmentation model when it is trained with 5%, 10% pixel-wise labels in Figure 7. It can be seen that we handle cluttered background well, and densely distributed small objects and arbitrary oriented objects are clearly identified in our pseudo labels. They are very similar to the ground truth in training set of iSAID dataset [11], which explains why our method can achieve similar performance with much less pixel-wise labels.



(**a**) Image (**b**) Pseudo labels (**c**) Ground truth (**d**) Image (**e**) Pseudo labels (**f**) Ground truth

Figure 7. Examples of pseudo pixel-wise labels generated by ancillary segmentation model when trained with 5% (upper three rows) and 10% (bottom three rows) of pixel-wise label.

Instance segmentation results. Figure 8 shows some representative instance segmentation results for comparison, and Figure 9 shows more examples. It can be seen that our approach is able to produce high quality instance segmentation results, even in challenging scenarios, and the visual quality is very close to 100% fully supervised method. In Figure 8, the

example of the first row shows that our method performs well in image with cluttered background, and is able to detect the vehicles in the shadow of buildings. The example of the second row may not be good enough in detecting small objects due to the limitation of primary instance segmentation model, it shows the recall rate of dense and small objects is basically the same with fully supervised setting. The example of the third row shows our method can segment object with arbitrary orientation well. The example of the fourth row shows that the mask prediction of our method outperforms the weakly supervised setting when segmenting objects with complex shape like planes. In short, the overall performance of our method is quite close to the fully supervised setting.

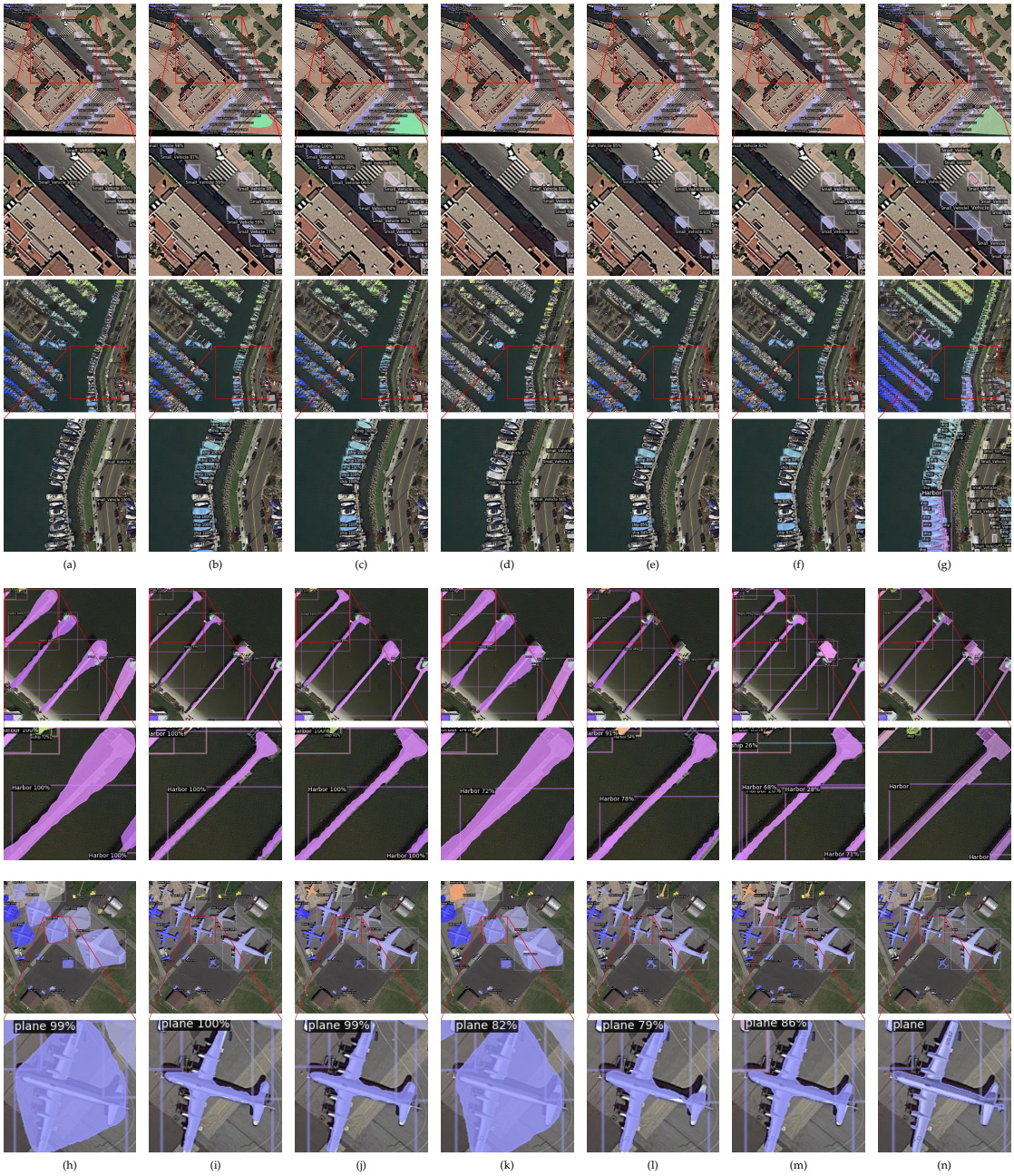
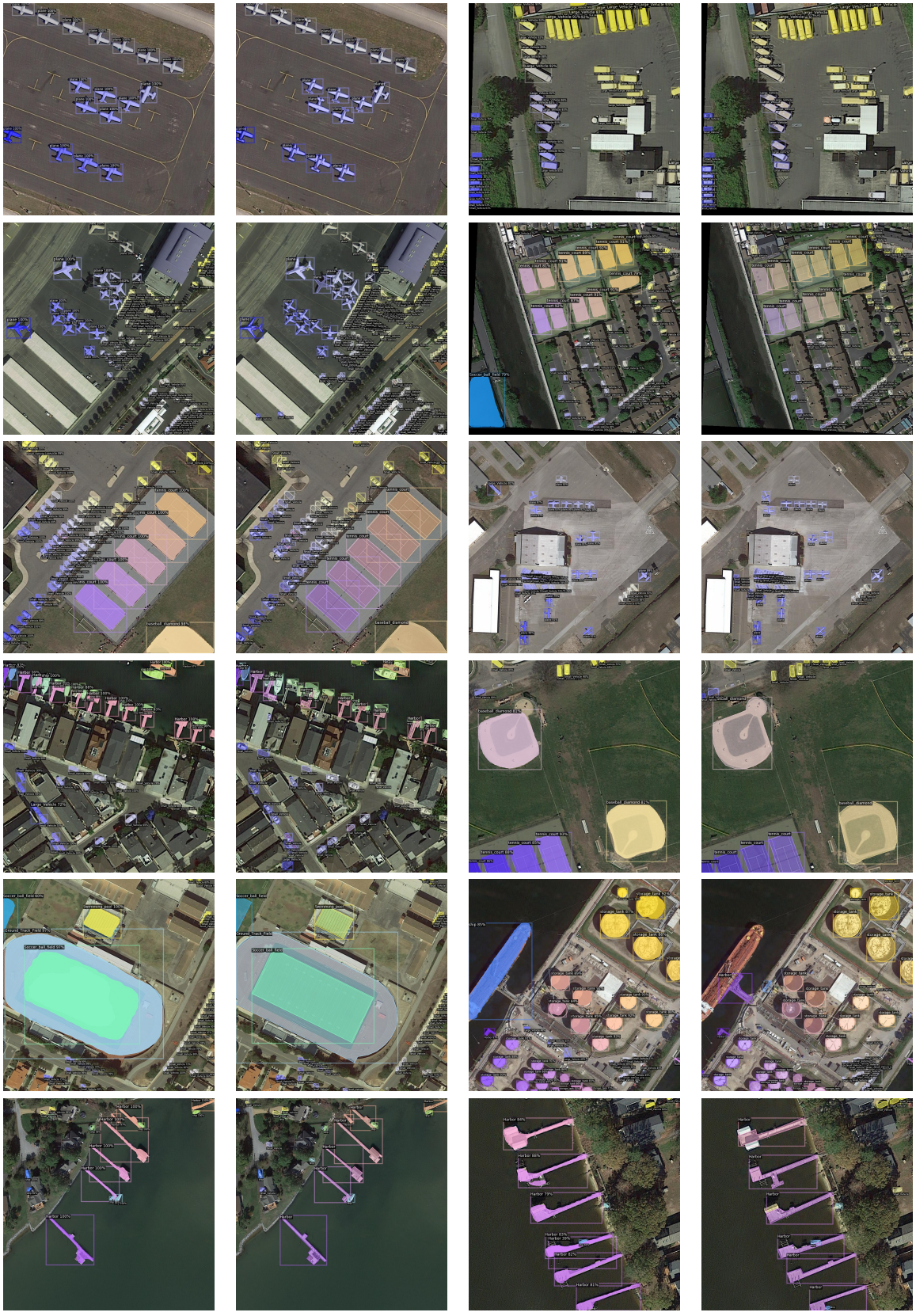


Figure 8. Some representative qualitative results on iSAID dataset [11]. (**a**) Full (10%, 22.5 *AP*) Mask R-CNN [4]; (**b**) Ours (10%, 32.1 *AP*) Mask R-CNN [4]; (**c**) Full (100%, 33.9 *AP*) Mask R-CNN [4]; (**d**) Full (10%, 23.3 *AP*) CenterMask [8]; (**e**) Ours (10%, 31.5 *AP*) CenterMask [8]; (**f**) Full (100%, 34.0 *AP*) CenterMask [8]; (**g**) Ground Truth; (**h**) Weak (26.5 *AP*) Mask R-CNN [4]; (**i**) Ours (10%, 32.1 *AP*) Mask R-CNN [4]; (**j**) Full (100%, 33.9 *AP*) Mask R-CNN [4]; (**k**) Weak (26.1 *AP*) CenterMask [8]; (**l**) Ours (10%, 31.5 *AP*) CenterMask [8]; (**m**) Full (100%, 34.0 *AP*) CenterMask [8]; (**n**) Ground Truth.



(**a**) (**b**) (**c**) (**d**)

Figure 9. Examples of the segmentation results with our method on iSAID dataset [11]. (**a**) Ours (10%, 32.1 *AP*) Mask R-CNN [4]; (**b**) Ground Truth; (**c**) Ours (10%, 31.5 *AP*) CenterMask [8]; (**d**) Ground Truth.

4.4. Ablation Study

The quality of pseudo instance-level pixel-wise labels generated by ancillary segmentation model significantly affects the final results. In this section, we conduct ablation study on the bounding box attention module, bounding box filter module and oriented bounding box labels to investigate their contribution on improving the quality of pseudo labels and effectiveness on solving the challenges in aerial images. We compare the pseudo labels with ground truth labels and evaluate their quality in the same metric used in main experiments, i.e., AP, AP_{50}, AP_{75}, AP_S, AP_M and AP_L. We use 10% of pixel-wise labels with 90% of bounding box labels for experiments and provide the results on validation set of iSAID dataset [11]. The quantitative results are shown in Table 3, to illustrate the quantitative results more clearly, we show the content of Table 3 in form of bar chart in Figure 10. And some examples of pseudo labels are shown in Figure 11 as qualitative results.

Baseline. As shown in Table 3, without bounding box attention modules, bounding box filter modules and bounding box labels, the pseudo labels generated by ancillary segmentation model have low accuracy. In Figure 11b we can see that there is severe noise in pseudo labels, e.g., false positives caused by a cluttered background, false positives of small objects and incomplete shapes of objects with arbitrary orientations.

Table 3. Quantitative results of pseudo pixel-wise labels generated by ancillary segmentation model when trained with 10% pixel-wise labeled images.

Bounding Box Attention Module	Bounding Box Filter Module	Oriented Bounding Box	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
×	×	×	12.9	23.9	11.8	6.3	15.4	14.2
√	×	×	38.1	70.4	36.6	32.9	41.3	36.3
√	√	×	41.2	77.2	38.8	34.5	47.3	42.6
√	√	√	44.2	79.1	44.5	35.1	51.7	44.9

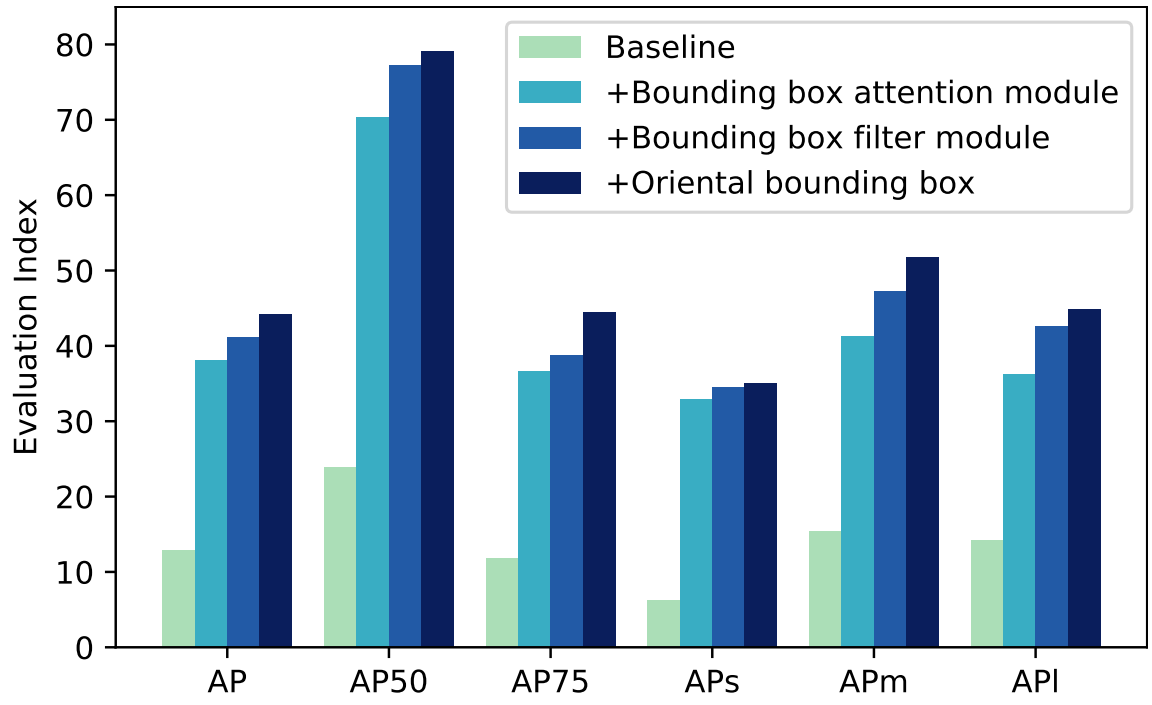
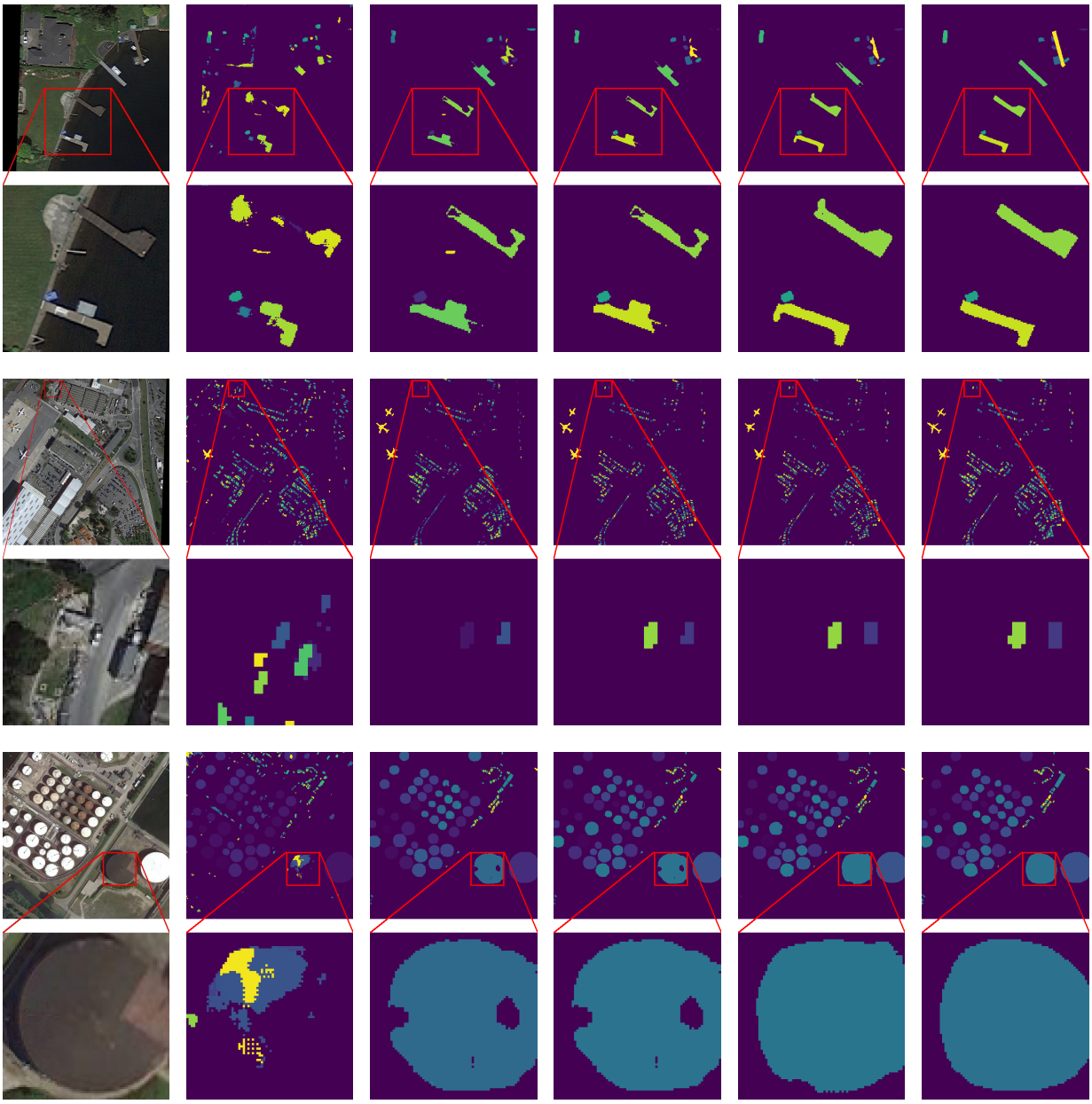


Figure 10. Visualized bar chart for the quantitative results of Table 3.



(**a**) Image (**b**) Baseline (**c**) +BBAM (**d**) +BBAM, BBFM (**e**) +BBAM, BBFM, OBB (**f**) Ground truth

Figure 11. Examples of pseudo pixel-wise labels generated by ancillary segmentation model. Different colors are used to distinguish different object instances in an image. We show the generated pseudo pixel-wise labels under different setting, it can be seen that bounding box attention modules (BBAM), bounding box filter modules (BBFM), and the usage of oriental bounding box (OBB) effectively improve the quality of pseudo labels.

Bounding box attention module. Although the structure of bounding box is simple, Table 3 and Figure 10 show that it significantly improves the overall results. Specifically, the result on AP_S increases from 6.3 to 32.9, which demonstrates its effectiveness on segmenting densely distributed small objects.

As for cluttered backgrounds, comparing the Figure 11b,c, we can see the false positives in background area are largely removed by the bounding box attention module. Furthermore, shapes of objects in pseudo labels are more complete. Both quantitative results and qualitative results show its effectiveness on addressing the challenge of cluttered background.

Bounding box filter module. As shown in Table 3 and Figure 10, the bounding box filter module further improves the overall result especially on AP_{50}, from 70.4 to 77.2. In

Figure 11, we can see that the false positives in background area in Figure 11c are removed by bounding box filter module in Figure 11d.

In short, these results show that the bounding box filter module improves performance of ancillary segmentation model by simply removing the false positives outside of bounding box. In this way, it largely reduces negative effects caused by cluttered background.

Horizontal bounding box vs. Oriented bounding box. Table 3 shows that the utilization of oriented bounding box improves result on AP_{75} from 38.8 to 44.5, which means oriented bounding box labels can help ancillary segmentation model to obtain more accurate segmentation results in aerial images. As shown in Figure 11e, after replacing horizontal bounding box labels with oriented bounding box labels, our ancillary segmentation model is able to generate more accurate pseudo labels for both small objects and objects with arbitrary orientations. This leads to about a 0.5–0.9 AP improvement for primary instance segmentation model as shown in Table 1, which verifies its effectiveness on solving the challenge of arbitrary orientations in aerial images. More importantly, our pseudo labels in Figure 11e are very close to ground truth pixel-wise labels, and this explains why the proposed method can achieve similar results to 100% fully supervised setting.

5. Conclusions

In this paper, we present a pipeline of hybrid supervision for instance segmentation in aerial images. We design an ancillary segmentation model to generate accurate pseudo pixels-wise labels from real-world aerial images which only needs a small portion of pixel-wise labels for training. It largely reduces the labeling cost while helping the primary instance segmentation model to achieve satisfactory performance. The proposed bounding box attention module can effectively suppress the noise from clutter background in aerial images, and improve the capability of segmenting small objects, addressing the key challenges of cluttered background and small objects. The proposed bounding box filter module removes the false positives caused by cluttered background and densely distributed objects, addressing the key challenge of cluttered background. Besides, we replace horizontal bounding box labels with oriented bounding box labels to further improve performance of ancillary segmentation model on generating high quality pseudo pixel-wise labels. On a recent large-scale instance segmentation dataset for aerial images, i.e., iSAID [11], we achieve comparable performance 32.1 AP to fully supervised setting 33.9 AP which is obviously higher than weakly supervised setting 26.5 AP.

In future, it is worth investigating how to design and further jointly optimize the primary instance segmentation model to achieve better performance for aerial images.

Author Contributions: Conceptualization, L.C. (Linwei Chen); methodology, L.C.; software, L.C.; validation, L.C., Y.F. (Ying Fu) and S.Y. (Shaodi You); formal analysis, L.C.; investigation, L.C.; resources, Y.F. and H.L. (Hongzhe Liu); data curation, L.C.; writing—original draft preparation, L.C.; writing—review and editing, L.C., Y.F. and S.Y.; visualization, L.C.; supervision, Y.F. and S.Y.; project administration, Y.F.; funding acquisition, Y.F., H.L. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grants No. 61936011, No. 61871039, No. 62006101, and the Collaborative Innovation Center for Visual Intelligence under Grant No. CYXC2011.

Data Availability Statement: Not applicable.

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

References

1. Wen, D.; Huang, X.; Zhang, L.; Benediktsson, J.A. A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 609–625. [CrossRef]
2. Volpi, M.; Tuia, D. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2**016, *55*, 881–893. [CrossRef]
3. Kopsiaftis, G.; Karantzalos, K. Vehicle detection and traffic density monitoring from very high resolution satellite video data. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 1881–1884.
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
5. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
6. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid task cascade for instance segmentation. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 4974–4983.
7. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT: Real-time instance segmentation. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 9157–9166.
8. Lee, Y.; Park, J. CenterMask: Real-Time Anchor-Free Instance Segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 13906–13915.
9. Chen, H.; Sun, K.; Tian, Z.; Shen, C.; Huang, Y.; Yan, Y. BlendMask: Top-down meets bottom-up for instance segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8573–8581.
10. Bearman, A.; Russakovsky, O.; Ferrari, V.; Li, F.-F. What's the point: Semantic segmentation with point supervision. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 549–565.
11. Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.S.; Bai, X. iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images. In Proceedings of the International Conference on Computer Vision Workshop, Seoul, Korea, 27–28 October 2019; pp. 28–37.
12. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision* **2010**, *88*, 303–338. [CrossRef]
13. Zhou, Y.; Zhu, Y.; Ye, Q.; Qiu, Q.; Jiao, J. Weakly supervised instance segmentation using class peak response. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3791–3800.
14. Ahn, J.; Cho, S.; Kwak, S. Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 2209–2218.
15. 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 International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 3345–3354.
16. Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; Schiele, B. Simple does it: Weakly supervised instance and semantic segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 876–885.
17. Li, Q.; Arnab, A.; Torr, P.H. Weakly-and semi-supervised panoptic segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 102–118.
18. 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 Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 6586–6597.
19. Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut" interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.* **2004**, *23*, 309–314. [CrossRef]
20. Arbeláez, P.; Pont-Tuset, J.; Barron, J.T.; Marques, F.; Malik, J. Multiscale combinatorial grouping. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 328–335.
21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3431–3440.
22. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 801–818.
23. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
24. Wu, T.; Tang, S.; Zhang, R.; Cao, J.; Li, J. Tree-structured kronecker convolutional network for semantic segmentation. In Proceedings of the International Conference on Multimedia and Expo, Shanghai, China, 8–12 July 2019; pp. 940–945.
25. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2881–2890.
26. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1925–1934.

27. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 3146–3154.
28. Hung, W.C.; Tsai, Y.H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; Yang, M.H. Scene parsing with global context embedding. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2631–2639.
29. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7151–7160.
30. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. Exfuse: Enhancing feature fusion for semantic segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 269–284.
31. Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; Wang, X. Mask scoring R-CNN. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 6409–6418.
32. Ying, H.; Huang, Z.; Liu, S.; Shao, T.; Zhou, K. EmbedMask: Embedding Coupling for One-stage Instance Segmentation. *arXiv* **2019**, arXiv:1912.01954.
33. 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.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
34. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
35. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
36. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
37. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
38. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 9627–9636.
39. Sherrah, J. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv* **2016**, arXiv:1606.02585.
40. Ghosh, A.; Ehrlich, M.; Shah, S.; Davis, L.S.; Chellappa, R. Stacked U-Nets for Ground Material Segmentation in Remote Sensing Imagery. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 257–261.
41. Hamaguchi, R.; Fujita, A.; Nemoto, K.; Imaizumi, T.; Hikosaka, S. Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, CA, USA, 12–15 March 2018; pp. 1442–1450.
42. Mou, L.; Zhu, X.X. Vehicle instance segmentation from aerial image and video using a multitask learning residual fully convolutional network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 6699–6711. [CrossRef]
43. Feng, Y.; Diao, W.; Zhang, Y.; Li, H.; Chang, Z.; Yan, M.; Sun, X.; Gao, X. Ship Instance Segmentation from Remote Sensing Images Using Sequence Local Context Module. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 1025–1028.
44. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
45. Lam, D.; Kuzma, R.; McGee, K.; Dooley, S.; Laielli, M.; Klaric, M.; Bulatov, Y.; McCord, B. xview: Objects in context in overhead imagery. *arXiv* **2018**, arXiv:1802.07856.
46. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? The inria aerial image labeling benchmark. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229.
47. Goldberg, H.; Brown, M.; Wang, S. A benchmark for building footprint classification using orthorectified rgb imagery and digital surface models from commercial satellites. In Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop, Washington, DC, USA, 10–12 October 2017; pp. 1–7.
48. Weir, N.; Lindenbaum, D.; Bastidas, A.; Van Etten, A.; McPherson, S.; Shermeyer, J.; Kumar, V.; Tang, H. SpaceNet MVOI: A Multi-View Overhead Imagery Dataset Supplementary Material. *arXiv* **2019**, arXiv:1903.12239.
49. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision* **2015**, *111*, 98–136. [CrossRef]
50. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
51. 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, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.

52. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision* **2015**, *115*, 211–252. [CrossRef]
53. Bellver, M.; Salvador, A.; Torres, J.; Giro-i-Nieto, X. Budget-aware Semi-Supervised Semantic and Instance Segmentation. *arXiv* **2019**, arXiv:1905.05880.
54. Wei, Y.; Xiao, H.; Shi, H.; Jie, Z.; Feng, J.; Huang, T.S. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7268–7277.
55. Ibrahim, M.S.; Vahdat, A.; Macready, W.G. Semi-Supervised Semantic Image Segmentation with Self-correcting Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 12715–12725.
56. Neven, D.; Brabandere, B.D.; Proesmans, M.; Gool, L.V. Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 8837–8845.
57. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
58. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational region cnn for orientation robust scene text detection. *arXiv* **2017**, arXiv:1706.09579.
59. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 8232–8241.
60. Berman, M.; Rannen Triki, A.; Blaschko, M.B. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4413–4421.
61. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the PInternational Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014; pp. 1–15.
62. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.