remote sensing

MDPI

Article

JMLNet: Joint Multi-Label Learning Network for Weakly Supervised Semantic Segmentation in Aerial Images

Rongxin Guo [1,2,3,4], Xian Sun [1,2,3,4,*], Kaiqiang Chen [1,3], Xiao Zhou [1,5], Zhiyuan Yan [1,3], Wenhui Diao [1,3] and Menglong Yan [1,3]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China; guorongxin18@mails.ucas.ac.cn (R.G.); chenkaiqiang14@mails.ucas.ac.cn (K.C.); zhouxiao@aircas.ac.cn (X.Z.); yanzy@aircas.ac.cn (Z.Y.); diaowh@aircas.ac.cn (W.D.); yanml@aircas.ac.cn (M.Y.)

[2] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100190, China

[3] Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China

[4] University of Chinese Academy of Sciences, Beijing 100190, China

[5] Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China

* Correspondence: sunxian@mail.ie.ac.cn

Received: 5 August 2020; Accepted: 20 September 2020; Published: 27 September 2020

Abstract: Weakly supervised semantic segmentation in aerial images has attracted growing research attention due to the significant saving in annotation cost. Most of the current approaches are based on one specific pseudo label. These methods easily overfit the wrongly labeled pixels from noisy label and limit the performance and generalization of the segmentation model. To tackle these problems, we propose a novel joint multi-label learning network (JMLNet) to help the model learn common knowledge from multiple noisy labels and prevent the model from overfitting one specific label. Our combination strategy of multiple proposals is that we regard them all as ground truth and propose three new multi-label losses to use the multi-label guide segmentation model in the training process. JMLNet also contains two methods to generate high-quality proposals, which further improve the performance of the segmentation task. First we propose a detection-based GradCAM (GradCAM^D) to generate segmentation proposals from object detectors. Then we use GradCAM^D to adjust the GrabCut algorithm and generate segmentation proposals (GrabCut^C). We report the state-of-the-art results on the semantic segmentation task of iSAID and mapping challenge dataset when training with bounding boxes annotations.

Keywords: deep learning; image segmentation; weak supervision; aerial image; multi-label learning

1. Introducition

Semantic segmentation in aerial images is a significant task, which aims at classifying each pixel in the given aerial images. It is useful for city planning, weather service, and other applications of remote sensing. Recently, Fully Convolutional Network (FCN) [1] based methods [2–21] have made great progress in semantic segmentation. These works require pixel-level supervised data in the training process. However, it is rather expensive to create pixel-level semantic segmentation training sets. Pixel-level annotations cost about 15x more time [22] than bounding box annotations. Considering bounding boxes are cheaper, we can research semantic segmentation with bounding boxes supervision. Several weakly supervised segmentation methods [23–27] explore closing the gap between pixel-level supervision and bounding boxes supervision.These methods mainly refine segmentation proposals from bounding boxes supervision, then take these segmentation proposals as pixel-level supervision and train deep FCN model.These methods mainly use traditional proposals like CRF [25], MCG [28] and GrabCut [29]. CRF [25] has been broadly used in semantic segmentation. It tries to model the relationship between pixels and enforce the predictions of pixels that have similar visual appearances to be more consistent. MCG [28] is a unified approach for bottom-up hierarchical image segmentation and object proposal generation. GrabCut [29] is an image segmentation method based on graph cuts. It requires a bounding box around the object. GrabCut estimates the color distribution of the target object and background using a Gaussian mixture model. BoxSup [23] takes MCG [28] as initial segmentation proposals and updated the proposals in an iterative way. SDI [30] takes intersection of MCG [28] and GrabCut [29] as segmentation proposals. Song et al.[27] use dense CRF [25] as segmentation proposals. These methods all feed one specific proposal to segmentation model, which easily overfit the wrongly labeled pixels from noisy label and limit the performance and generalization of segmentation model. So it is a natural idea to tackle these problems by taking advantage of multiple proposals in the training process.

To train with multiple proposals, traditional combining methods take intersection [30] of two kinds of segmentation proposals as supervision to reduce the noise. Pixels out of intersection are ignored in training. These pixels usually take up mainly part of the box area in difficult situations, which reduces the semantic information and limits segmentation model performance. We propose a joint multi-label learning network(JMLNet) to address the issue. The overall pipeline of our JMLNet is in Figure 1. Different from simply using the intersection of two proposals or only use one specific proposal, we regard multiple proposals as multi-label and make all noisy proposals contribute in the training process. Specifically, we propose three multi-label losses for training, including multi-label average loss (MA-Loss), multi-label minimum loss (MM-Loss), and box-wise multi-label minimum loss (BMM-Loss). These loss functions help segmentation model learn common knowledge from multiple noisy labels and prevent the model from overfitting one specific label.

The quality of Proposals is vital to weakly supervised semantic segmentation. Previous approaches train the models with MCG, GrabCut, or CRF proposals based on box supervision. Lacking high-level semantic knowledge, these proposals are easy to confuse in complicated scenes. As shown in Figure 2c, GrabCut confuses *building* and *plane* because of similar color. Low quality of traditional proposals damages the performance of segmentation model. We address this problem by proposing GradCAM^D and GrabCut^C, which generate high-quality pixel-level proposals. First, GradCAM^D aims to generate visual explanations and proper proposals from object detectors. GradCAM^D generates reliable proposals because the detection networks learn precise semantic information, as shown in Figure 2d. Second, we use GradCAM^D to adjust GrabCut algorithm and generate proposals, which is denoted as GrabCut^C. C indicates *GradCAM*. GrabCut^C can be simply seen as *GradCAM* + *GrabCut*. As shown in Figure 2e, GrabCut^C proposals are both reliable in the distinguished semantic area and detailed in instance edge. Our method improves the segmentation proposals' quality, which further improves the segmentation performance of JMLNet.

We summarize our contributions as follows:

- We propose a novel joint multi-label learning network(JMLNet), which first regards multiple proposals as multi-label supervision to train weakly supervised semantic segmentation model. JMLNet learns common knowledge from multiple noisy labels and prevents the model from overfitting one specific label.
- GradCAM^D and GrabCut^C methods are proposed to generate high-quality segmentation proposals, which further improve the segmentation performance of JMLNet. These proposals perform both reliable in the distinguished semantic area and detailed in instance edge.

- We report the state-of-the-art results on semantic segmentation tasks of iSAID and mapping challenge dataset when training using bounding boxes supervision, reaching comparable quality with the fully supervised model.

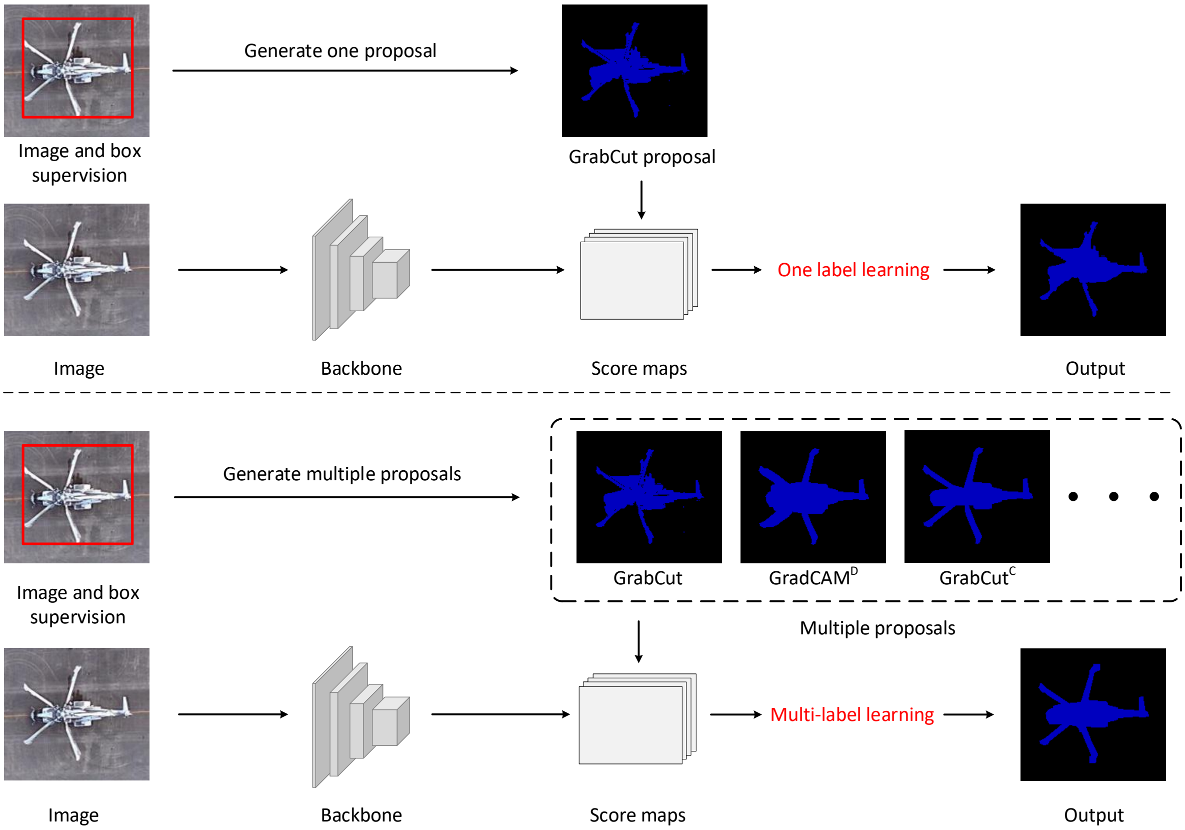

Figure 1. The overall pipeline of previous weakly supervised semantic segmentation methods (top) and our proposed JMLNet (bottom). Previous methods generate one specific proposal and use it in the training process. However, we first generate multiple proposals as multi-label supervision and use multi-label loss to train the segmentation model.

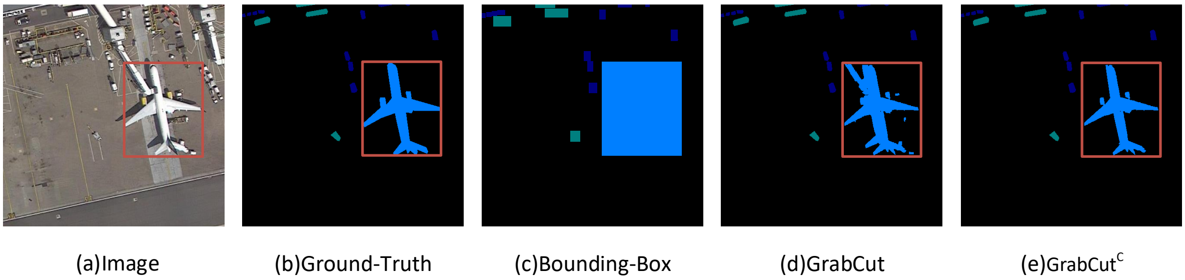

Figure 2. Segmentation proposals obtained from bounding box. (**a**) A training image. (**b**) Ground truth. (**c**) Rectangle proposals. (**d**) GrabCut [29] proposals. (**e**) We propose GrabCut^C proposals, which perform better than traditional proposals.

2. Related Work

We introduce the weakly supervised semantic segmentation methods of natural image and remote sensing image and aerial image, region proposal from box supervision, and learning semantic knowledge with noisy labels that are related to our work.

2.1. Weakly Supervised Semantic Segmentation of Natural Image

Weakly supervised semantic segmentation methods of natural image can be classified into four parts, including image labels methods [15,31–35], points labels methods [26], scribbles labels methods [36,37], and bounding boxes labels methods [23,27,30]. We mainly introduce bounding boxes labels methods in the following paragraph. BoxSup [23] takes MCG [28] as initial segmentation proposals and updated the proposals in an iterative way. SDI [30] takes intersection of MCG [28] and GrabCut [29] as segmentation proposals. Song et al.[27] propose an attention model to focus on the foreground regions.

2.2. Weakly Supervised Semantic Segmentation of Remote Sensing Image and Aerial Images

Weakly supervised semantic segmentation methods of remote sensing image and aerial images can be also classified into four parts, including image labels methods [4,38], points labels methods [39], scribbles labels methods [40], and bounding boxes labels methods [41]. WSF-NET [4] introduces a feature-fusion network to fuse different level feature of FCN [1] and increase the ability of feature representation. SPMF-Net [38] combines superpixel pooling to segmentation methods and use low level feature to get detail prediction. Wang et al. [39] use CAM [31] proposals as ground truth and train FCN [1] based model. Wu et al. [40] propose an adversarial architecture based model for segmentation. Rafique et al. [41] convert the bounding box into probabilistic masks and propose a boundary based loss function to restrict the edge of predict map to close to bounding box. We separate weakly supervised semantic segmentation as two aspects, including region proposal from box supervision and learning semantic knowledge with noisy labels.

2.3. Region Proposal from Box Supervision

Without proper pixel-level supervision, weakly supervised methods extract region proposal from box supervision. [25,28,29] are the most popular region proposal methods. BoxSup [23] takes MCG [28] as initial segmentation proposals and updated the proposals in an iterative way. SDI [30] takes intersection of MCG [28] and GrabCut [29] as segmentation proposals. Song et al. [27] use dense CRF [25] as segmentation proposals. These region proposal methods extract proposals from class-agnostic low-level features, which leads to generating confusing proposals in complicated scenes because of lacking high-level semantic information. To this end, we propose a GradCAM^D method to generate visual explanations from object detectors and proper proposals by setting the threshold. GradCAM^D generates reliable proposals because the detection network learns precise semantic information. Then we use GradCAM^D to adjust GrabCut algorithm and generate training labels, which performs both reliable in the distinguished semantic area and detailed in instance edge.

2.4. Learning Semantic Knowledge with Noisy Labels

Though we can use [25,28,29] to generate proposals within bounding boxes annotations, there are still so many noises compared with a full-supervised label. How to learn with noisy labels becomes a key problem of weakly supervised semantic segmentation. SDI [30] directly uses the intersection of two kinds of segmentation proposals to reduce the noise. Song et al.[27] use different filling rates as priors to help the model training. These methods all use one specific pseudo label. We first propose JMLNet to combine multiple noisy labels in the training process. JMLNet helps the model learn common knowledge from multiple noisy labels and prevent it from overfitting one specific label.

3. Our Method

3.1. Overview

In this section, we introduce the general pipeline of JMLNet. As shown in Figure 3, we collect multiple proposals like GrabCut, GradCAM^D, and GrabCut^C proposals as multi-label supervision and train the segmentation model with the proposed multi-label loss.

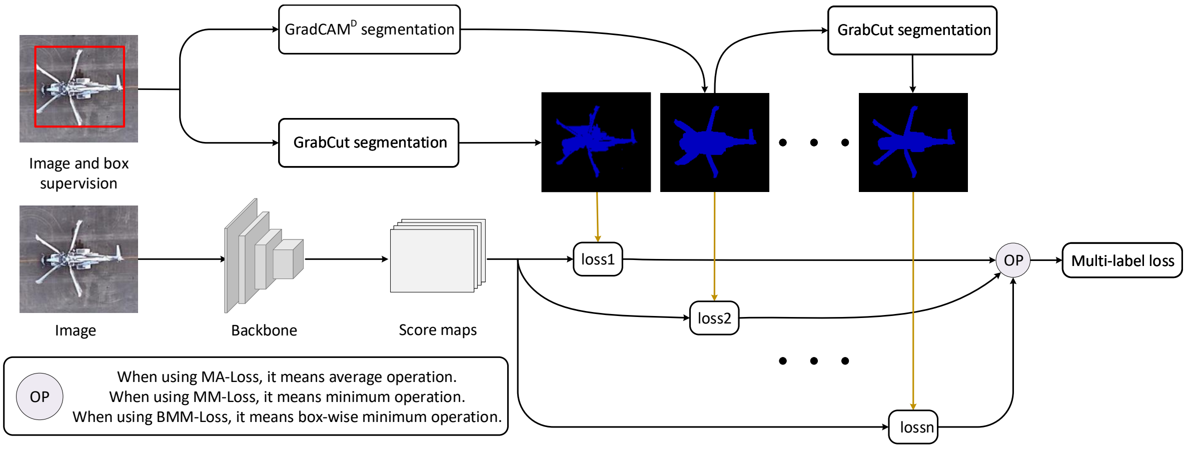
GradCAM^D segmentation
GrabCut segmentation
GrabCut segmentation
Image and box supervision
Image
Backbone
Score maps
loss1
loss2
lossn
OP
Multi-label loss
OP
When using MA-Loss, it means average operation.
When using MM-Loss, it means minimum operation.
When using BMM-Loss, it means box-wise minimum operation.

Figure 3. Overview of JMLNet. We generate multiple proposals as multi-label supervision and use multi-label loss to train the segmentation model.

Generating pseudo supervision. Except for popular segmentation proposals with bounding boxes labels, we generate GradCAM^D and GrabCut^C proposals as pseudo supervision. GradCAM^D is a detection-based GradCAM. The first step to generate GradCAM^D proposals is to train an object detector. We choose Faster R-CNN [42], a classical object detector, in our experiment. Then we calculate the GradCAM^D in feature map of Faster R-CNN and generate the pixel-level proposals. GradCAM^D is also used to adjust GrabCut algorithm and generate GrabCut^C proposals. All these proposals contribute in the training process.

Model training with multiple noisy labels. As shown in Figure 1, we choose popular Deeplab v3 [43] as semantic segmentation model. Note that we collect multiple proposals $\{CRF, GrabCut, GradCAM^D,\ GrabCut^C\}$ for a single input image, so we propose multi-label average-loss (MA-Loss), multi-label minimum loss (MM-Loss), and box-wise multi-label minimum loss (BMM-Loss) to help the model learn common knowledge from multiple noisy labels and prevent the model from overfitting one specific label.

3.2. Multi-Label Losses for Multiple Proposals

Most semantic segmentation methods use pixel-wise cross entropy loss as loss function:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c} \log p_{n,c} \tag{1}$$

where N is the number of pixels, C is the number of classes, $y \in \{0,1\}$ is the ground truth, and $p \in [0,1]$ is the estimated probability.

It is obvious that our pseudo proposals are all noisy within bounding boxes annotations and one specific proposal is hard to perform best in all image sets. Based on the analysis above, we propose three multi-label losses to help the model learn common knowledge from multiple noisy labels and prevent the model from overfitting one specific label. In practice, we propose multi-label average-loss (MA-Loss), multi-label minimum loss (MM-Loss), and box-wise multi-label minimum loss (BMM-Loss).

Dealing with multiple noisy labels, an intuitive idea is to calculate the average value of cross entropy losses for multiple proposals. We denote it as multi-label average-loss (MA-Loss):

$$\mathcal{L}_{MA}(p, \mathcal{Y}) = \frac{1}{Z}\sum_{z=1}^{Z}(\mathcal{L}_{CE}(p, y_z)) \tag{2}$$

where $\mathcal{Y}$ denotes pseudo labels set, Z is the number of proposals types.

Further, we calculate the cross entropy losses for multi proposals and take the minimum value in back propagation. We denote it as multi-label minimum loss (MM-Loss):

$$\mathcal{L}_{MM}(p, \mathcal{Y}) = \min_{z} \mathcal{L}_{CE}(p, y_z), z \in [1, Z] \tag{3}$$

In weakly supervised segmentation, a set of box-level labeled data $\mathcal{D} = \{(\boldsymbol{I}, \boldsymbol{B})\}$ are given, where $\boldsymbol{I}$ and $\boldsymbol{B}$ denote an image and box-level ground truth respectively. We know the pixels out of $\boldsymbol{B}$ are background class according to ground truth. So pixels in $\boldsymbol{B}$ are key problem for our case. We categorize image pixels into two sets $\mathcal{P}^+$ and $\mathcal{P}^-$ according to their coordinates position by

$$\mathcal{P}^+ = \{(i, j)|(i, j) \in \boldsymbol{B}\} \tag{4}$$

$$\mathcal{P}^- = \{(i, j)|(i, j) \notin \boldsymbol{B}\} \tag{5}$$

where (i, j) is coordinate.

We calculate the minimum value of cross entropy losses for multi proposals in $\mathcal{P}^+$ as follows:

$$\mathcal{L}^+(p, \mathcal{Y}) = \frac{1}{n^+}\sum_{(i,j)\in\mathcal{P}^+} \min_{z}(\sum_{c=1} -y^z_{ijc} \log p_{ijc}), z \in [1, Z] \tag{6}$$

where y^Z_{ijc} indicate estimated probability of different proposals and n^+ indicates pixel number of $\mathcal{P}^+$.

For all coordinates (i, j) in $\mathcal{P}^-$, $y_{i,j} = 0$. We use cross entropy loss in $\mathcal{P}^-$ as follows:

$$\mathcal{L}^-(p, \mathcal{Y}) = -\frac{1}{n^-}\sum_{(i,j)\in\mathcal{P}^-} \log p^b_{ij} \tag{7}$$

where p^b_{ij} indicates estimated probability of background and n^- indicates pixel number of $\mathcal{P}^-$.

The $\mathcal{L}^+$ and $\mathcal{L}^-$ make up box-wise multi-label minimum loss (BMM-Loss):

$$\mathcal{L}_{BMM}(p, \mathcal{Y}) = \mathcal{L}^+(p, \mathcal{Y}) + \mathcal{L}^-(p, \mathcal{Y}) \tag{8}$$

Our proposed MA-Loss, MM-Loss and BMM-Loss help the model learn common knowledge from multiple noisy labels and prevent the model from overfitting one specific label.

3.3. Pseudo Label Generation by Gradcam^D and Grabcut^C

The GradCAM^D of our approach is shown in Figure 4 and Algorithm 1. In order to obtain the GradCAM^D $\mathcal{D} \in \mathbb{R}^{u\times v}$ of width u and height v for target class, we first compute the gradient of target score s with respect to feature maps M_k, *i.e.*$\frac{\partial s}{\partial M_{ij}}$. $k \in [1, K]$ and K is the channel number of feature maps. These gradients flowing back obtain the weight α_k, which represents the weight of feature map M_k for target class.

$$\alpha_k = \frac{1}{uv}\sum_{i}\sum_{j}\frac{\partial s}{\partial M_{ij}} \tag{9}$$

We calculate a weighted combination of feature maps.

$$\mathcal{D} = \sum_{k=1}^{K} \alpha_k M_k \tag{10}$$

As shown in Figure 5, GradCAM^D explains why detector classifies a specific area as a specific class and cover instance region well. Based on the observation, we generate high GradCAM^D proposal and low GradCAM^D proposal by setting high and low thresholds to GradCAM^D, as shown in Figure 4 and Algorithm 1. Low GradCAM^D proposal $\mathcal{D}_\ell$ is closer to ground truth, and we can use it as the pseudo label to train segmentation model. High GradCAM^D proposal $\mathcal{D}_h$ can't cover all positive pixels of ground truth but contains less false-positive pixels.

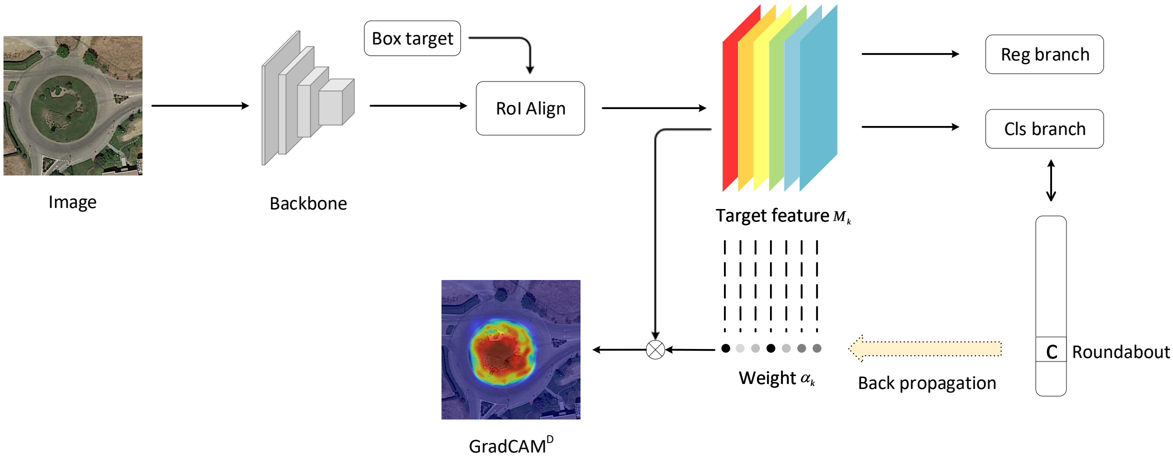

Figure 4. Overview of the GradCAM^D. We generate GradCAM^D using back propagation in the detector's classification branch. Best viewed in color.

Algorithm 1: Generation of Low GradCAM^D Proposals $\mathcal{D}_\ell$ and High GradCAM^D Proposals $\mathcal{D}_h$

```
Input: Image $I$; box supervision $B$; low GradCAM$^D$ threshold $\tau_\ell$; high GradCAM$^D$ threshold $\tau_h$.
Output: Low GradCAM$^D$ proposals $\mathcal{D}_\ell$; high GradCAM$^D$ proposals $\mathcal{D}_h$.
Feed the $I$ into the detector's backbone to produce feature $F^b$ ;
Feed the $F^b$ and $B$ into the RoIAlign to produce feature $F^{roi}$ ;
Feed the $F^{roi}$ into the detector's RCNN conv layer to produce feature $M_k$ ;
Feed the $M_k$ into the detector's classification branch to produce target score $s$ ;
Get weight $\alpha_k$ by Equation (9) ;
Get GradCAM$^D$ $\mathcal{D}$ by Equation (10) ;
for each value $p \in \mathcal{D}$ do
    if $p > \tau_\ell$ then
        $\mathcal{D}_\ell$.append($p$);
    end
    if $p > \tau_h$ then
        $\mathcal{D}_h$.append($p$);
    end
end
```

Different from generating visual explanations from classification network, like CAM [31] and GradCAM [32], GradCAM^D generates visual explanations from object detector. Box supervision is fully used, and the detector learns precise semantic information, which improves the proposal quality.

As shown in Figure 6, we take GradCAM^D as priors and categorize pixels into thre sets $\mathcal{D}^+$, $\mathcal{D}^-$ and $\mathcal{D}^u$ by

$$\mathcal{D}^+ = \{(i,j)|(i,j) \in \mathcal{D}_h\} \tag{11}$$

$$\mathcal{D}^- = \{(i,j)|(i,j) \notin \mathcal{D}_\ell\} \tag{12}$$

$$\mathcal{D}^u = \{(i,j)|(i,j) \in \mathcal{D}_\ell \cap (i,j) \notin \mathcal{D}_h\} \tag{13}$$

where (i,j) is coordinate. Pixels in $\mathcal{D}^+$ are fixed to the foreground, pixels in $\mathcal{D}^-$ are fixed to background and pixels in $\mathcal{D}^u$ are still uncertain. GrabCut updates proposals by taking these foreground and background information. The updated proposals are denoted as GrabCut^C proposals. As shown in Figure 2f, GrabCut^C generates proposals both reliable in the distinguished semantic area and detailed in instance edge.

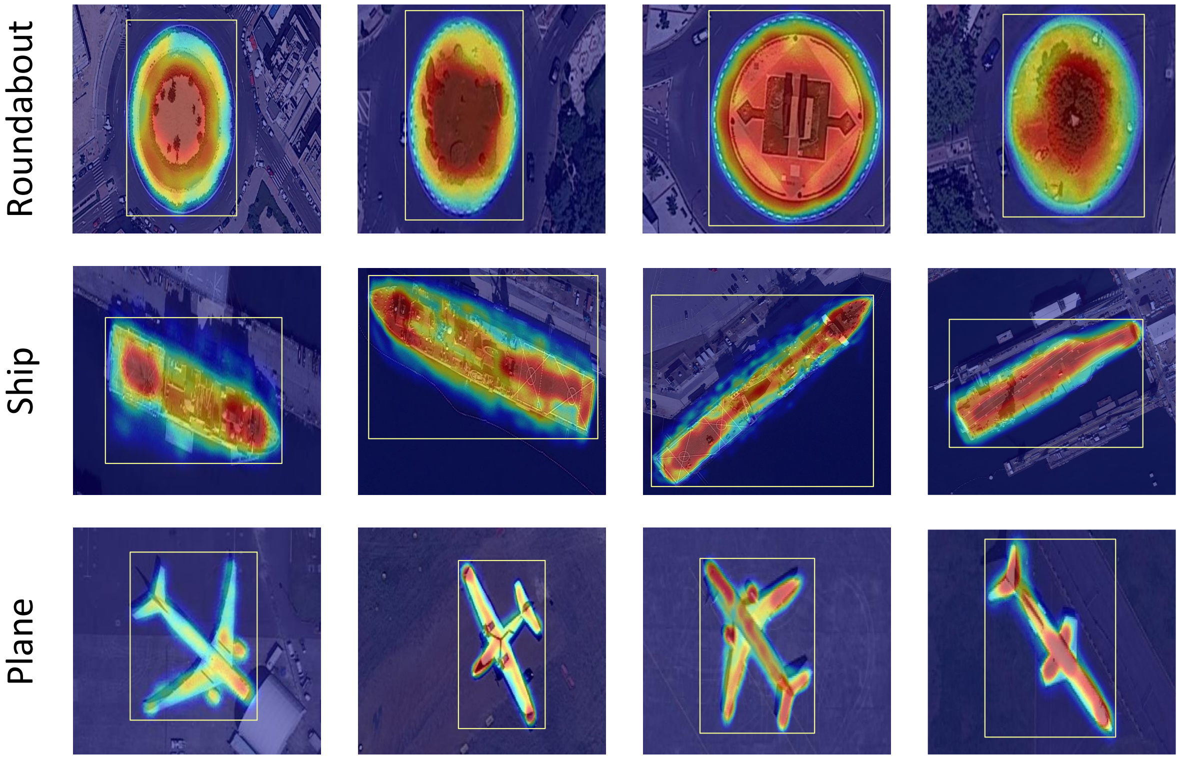

Figure 5. Visualization of the GradCAM^D. It shows why the detector classifies a specific area as a specific class and covers instance region well.

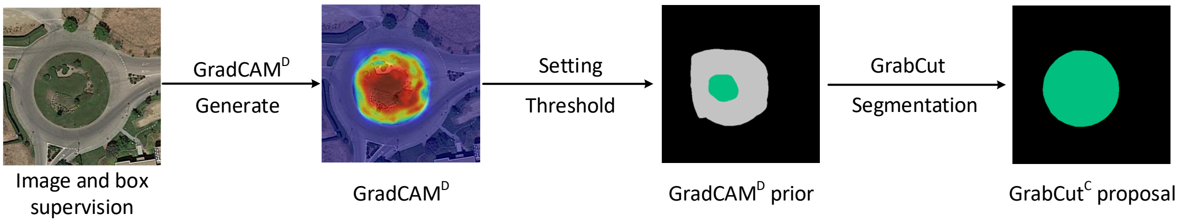

Figure 6. Overview of the GrabCut^C. We use GradCAM^D as prior to GrabCut and generate GrabCut^C. In GradCAM^D prior, green pixels represent foreground, black pixels represent background, and gray pixels represent uncertainty area. GrabCut takes this information as input and further refines proposal.

4. Experiments

In the experiments, we first introduce the experimental setup, then do ablation study of different super parameter, finally compare our method with the state-of-the-art methods.

4.1. Experimental Setup

In experimental setup, we introduce dataset, evaluation method and implementation details of our experiments.

Dataset: In our experiments, two aerial images dataset are used: iSAID [44] dataset and mapping challenge dataset [45]. We use iSAID [44] dataset, which is a further semantic labeled version for DOTA [46] dataset. It contains 15 classes of different objects and 1 background class. The spatial resolution of images ranges from 800 pixels to 13000 pixels, which exceed resolution of natural images by far. We train our method with 1,411 high-resolution images, eval with 458 high-resolution images. We use the mapping challenge dataset [45]. It contains 1 building class and 1 background class. We train our method with 280,741 images, eval with 60,317 images of size 300x300 pixels. We only exploit bounding boxes annotations when training. While the dataset contains labels for semantic segmentation, we only exploit box-level labels.

Evaluation: To evaluate the performance of our method and compare our results to other state-of-the-art methods, we calculate mean pixel Intersection-over-Union(mIoU), overall accuracy (OA), true positive rate (TPR) and true negative rate (TNR) as common practice [22,47]. IoU is defined as:

$$IoU = \frac{TP}{TP+FP+FN} \tag{14}$$

and mIoU is defined as:

$$mIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{TP}{TP+FP+FN} \tag{15}$$

and OA is defined as:

$$OA = \frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

and TPR is defined as:

$$TPR = \frac{TP}{TP+FN} \tag{17}$$

and TNR is defined as:

$$TNR = \frac{FP}{FP+TN} \tag{18}$$

where TP, FP, TN, FN are the number of true positives, false positives, true negatives and false negatives. C indicates the number of classes.

Implementation Details: For iSAID dataset, we crop the high-resolution images to 512 × 512 patches. We adopt the classical Deeplab v3 [43] model for our experiments, which takes widely used ResNet-50 [48] as backbone. Firstly, we train a detection model Faster-RCNN [42] with box-level labels of iSAID [44]. Using the proposed GradCAM^D and GrabCut^C methods, we generate pseudo segmentation proposals for train set. Secondly, we train the Deeplab v3 model with the GrabCut^C supervision for 50k iterations, further finetune it with proposed loss function for 10k iterations. We choose SGD as default optimizer. Mini-batch size is seted to 20. We set initial learning rate to 0.007 and multiply by $(1-\frac{step}{max_step})^{power}$ and $power$ is set to 0.9. We apply random horizontal flipping and random cropping to augment the diversity of dataset. We implement our method with the PyTorch [49] framework. For mapping challenge dataset, we follow the same basic setting as Rafique et al. [41] for fair comparison. We choose Adam optimizer with learning rate of $5e^{-4}$, β_1 = 0.9, and β_2 = 0.999. Mini-batch size is seted to 16. We train the network for 3 epochs.

4.2. Ablation Study

We conduct two types of ablation studies, including the analysis of the contribution of proposed loss functions and the performance of the proposal with different thresholds.

Proposals quality. We do experiments on different proposals and loss functions. As shown in Table 1, experimental results show that our proposed GradCAM^D and GrabCut^C proposals perform better than traditional proposals. We train the Deeplab v3 model with different proposals as pseudo labels, including rectangle proposals, CRF proposals, GrabCut proposals, our proposed GradCAM^D proposals and GrabCut^C proposals. As shown in Table 1, our proposed GradCAM^D and GrabCut^C proposals achieve 53.88% and 54.24% mIoU, outperforming all the compared methods. As shown in Figure 7, the main difference between GrabCut and our proposed GradCAM^D and GrabCut^C proposals is edge predictions. Using GrabCut as label, segmentation model will tend to do predictions based on low level features, including color and edge. In hard cases, low level features can not represent precise information of target features, which lead to wrong predictions. GradCAM^D and GrabCut^C proposals perform better because they are of high level features obtained from object detector. As shown in Table 2, we evaluate the effectiveness of GradCAM^D proposals on iSAID validation set. Our GradCAM^D can be seen as a detection-based GradCAM. So we make a comparison between GradCAM^D and standard GradCAM proposals within bounding box. Experimental results show that our proposed GradCAM^D outperforms standard GradCAM.

Table 1. Evaluating the effectiveness of JMLNet, including GradCAM^D proposals, GrabCut^C proposals and three novel loss functions on iSAID validation set. BOX: Rectangle proposals, CRF: CRF proposals, GrabCut: GrabCut proposals.

Loss	**Proposals**					**mIoU**
	BOX	**CRF**	**GrabCut**	**GradCAM^D**	**GrabCut^C**	
CE Loss	✓					46.20
		✓				51.27
			✓			53.12
				✓		53.88
					✓	54.24
MA-Loss	✓				✓	52.27
		✓			✓	52.64
			✓		✓	53.58
				✓	✓	54.45
			✓	✓	✓	54.61
		✓	✓	✓	✓	54.21
	✓	✓	✓	✓	✓	53.72
MM-Loss	✓				✓	52.45
		✓			✓	52.83
			✓		✓	54.25
				✓	✓	54.64
			✓	✓	✓	54.97
		✓	✓	✓	✓	54.63
	✓	✓	✓	✓	✓	53.94
BMM-Loss	✓				✓	53.22
		✓			✓	53.41
			✓		✓	54.67
				✓	✓	55.10
			✓	✓	✓	55.34
		✓	✓	✓	✓	54.85
	✓	✓	✓	✓	✓	54.05

Losses selection. As shown in Table 1, experimental results show that our proposed MA-Loss, MM-Loss and BMM-Loss all improve segmentation results, in which BMM-Loss performs best. We combine different proposals and use our proposed loss functions to train the Deeplab v3 model. As shown in Table 1, using a combination of different proposals and our proposed loss functions, we improve segmentation results significantly. In particular, combination of $\{GrabCut, GradCAM^D, GrabCut^C\}$ and BMM-Loss achieve the best performance, 55.34% mIoU. We analyze that the reason why BMM-Loss performs best is BMM-Loss considers the similarity between predictions and multiple proposals in pixel-wise within boxes. The other loss functions, MA-Loss and MM-Loss, only focus on loss of the whole image. Segmentation performance will not be improved by adding rectangle proposals and CRF proposals to $\{GrabCut,GradCAM^D,GrabCut^C\}$. We analyze that compared with $\{GrabCut,GradCAM^D,GrabCut^C\}$, rectangle proposals and CRF proposals are quite rough and introduce more wrongly labeled pixels. Low quality of proposals will hurt the performance of segmentation model. We deal with noisy label by automatically selecting the high quality label in training process. It partly solves the problem of noisy label but adding bad pseudo-labels will hurt our performance in practice. There are still many future works that can be done to handle the bad influence of noisy label.

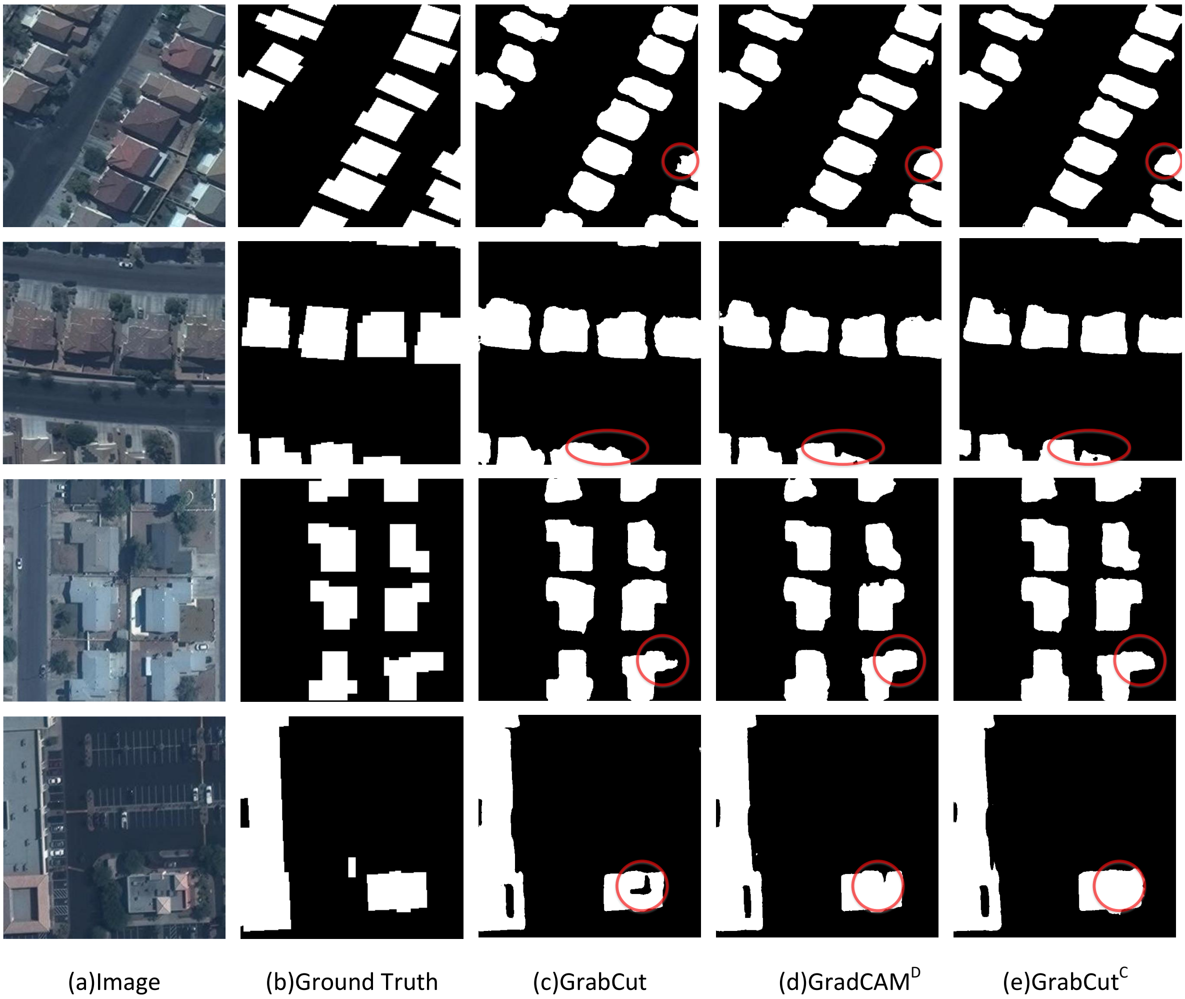

Figure 7. Examples of segmentation results of our method on mapping challenge dataset. (**a**) Image. (**b**) Ground Truth. (**c**) GrabCut. (**d**) GradCAM^D. (**e**) GrabCut^C. Red circle indicates the difference results.

Table 2. Evaluating the effectiveness of GradCAM^D proposals on iSAID validation set. GradCAM: Standard GradCAM proposals within bounding box.

	GradCAM	**GradCAM^D**
mIoU (%)	51.35	53.88

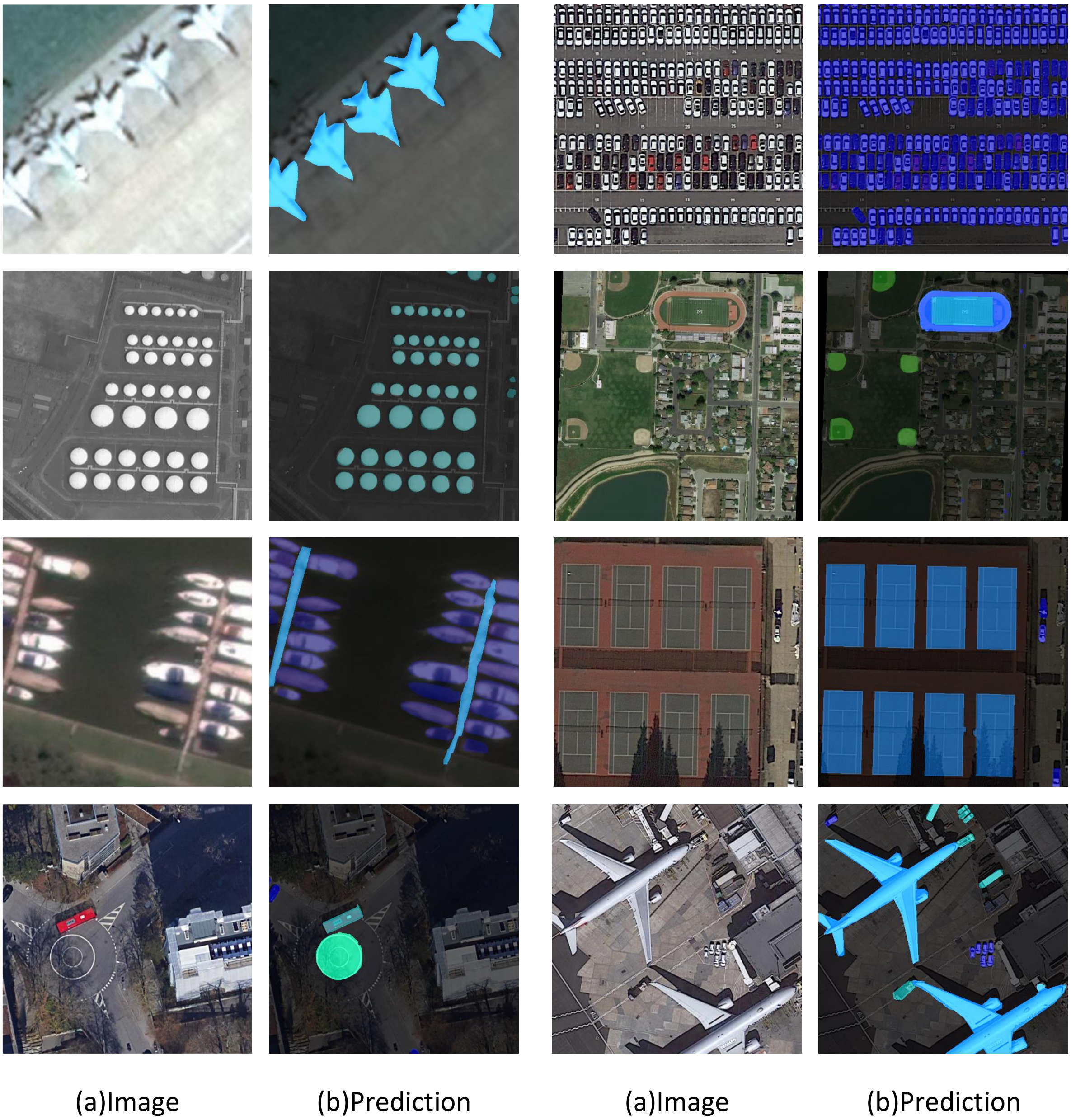

Figure 8. Examples of segmentation results of our method on iSAID. (**a**) Image. (**b**) Prediction.

Threshold τ_ℓ of low GradCAM^D proposals $\mathcal{D}_\ell$. Low GradCAM^D proposals $\mathcal{D}_\ell$ depends on one key hyper-parameter, threshold τ_ℓ. We use $\mathcal{D}_\ell$ as pseudo label to train segmentation model, which is vital to final performance. The threshold τ_ℓ balances the foreground and background pixels within boxes annotations. If τ_ℓ is set to 0, all pixels within boxes annotations are seen as proposals. As τ_ℓ increases, the area of proposals decreases and only the distinguished part of GradCAM^D remained in proposals. Table 3 shows the influence of threshold τ_ℓ. As τ_ℓ get higher, the area of foreground pixels get lower. Because foreground pixels usually take up most area within boxes annotations, so we find best τ_ℓ in small values. When $\tau_\ell = 0.15$, using $\mathcal{D}_\ell$ proposals as ground truth, we achieve the best performance. Table 1 indicate that $\mathcal{D}_\ell$ reachs 53.88% mIoU on iSAID validation set. We also fix $\tau_\ell = 0.15$ in generating GrabCut^C proposals.

Table 3. Influence of τ_ℓ. The hyper-parameter τ_ℓ balances the foreground and background pixels when generating low GradCAM^D proposals. τ_ℓ = 0 means all pixels within boxes annotations are seen as proposals.

τ_ℓ	0	0.05	0.1	0.15	0.2	0.25
mIoU (%)	44.20	51.34	53.46	53.88	53.76	53.20

Threshold τ_h of high GradCAM^D proposals $\mathcal{D}_h$. High GradCAM^D proposals $\mathcal{D}_h$ depends on threshold τ_h. We use $\mathcal{D}_h$ as foreground to adjust GrabCut algorithm and generate GrabCut^C. Table 4 shows the influence of threshold τ_h. When τ_h = 0.8, GrabCut^C achieves the best performance. Table 1 indicates that GrabCut^C reachs 54.24% mIoU in iSAID validation set.

Table 4. Influence of τ_h. The hyper-parameter τ_h influences quality of GrabCut^C. τ_h = 1 means no positive foreground for GrabCut proposals.

τ_h	0.5	0.6	0.7	0.8	0.9	1
mIoU (%)	53.30	53.58	54.10	54.24	54.23	53.67

4.3. Comparison with the State-Of-The-Art Methods

In the comparison with the state-of-the-art methods, we mainly choose SDI [30], Song et al. [27] and Rafique et al. [41].

Results of weakly-supervised semantic segmentation on iSAID dataset. As shown in Table 5, our method achieves 55.34% mIoU, 98.58% OA, 61.75% TPR and 99.63% TNR on iSAID validation set. Specific IOU for per category can be found in Table 6. Figure 8 shows the segmentation results of our method. Our method outperforms all compared weakly supervised semantic segmentation approaches. The results indicate that our proposed method is effective when learning common knowledge from multiple noisy labels.

Table 5. Weakly supervised results on iSAID validation set.

Supervision	Methods	mIoU (%)	OA (%)	TPR (%)	TNR (%)
Weak	SDI [30]	53.82	98.30	59.87	99.59
	Song et al. [27]	54.18	98.36	60.56	99.60
	Ours	55.34	98.58	61.75	99.63
Semi	SDI [30]	54.87	98.43	61.23	99.61
	Song et al. [27]	55.15	98.50	61.64	99.62
	Ours	56.76	98.62	63.25	99.64
Full	Deeplab v3 [50]	59.05	98.75	65.78	99.67

Table 6. Our segmentation results for per category on iSAID validation set, which are evaluated by mIoU (%). ST: Storage tank, BD: Baseball diamond, TC: Tennis court, BC: Basketball court, GTF: Ground field track, LV: Large vehicle, SV: Small vehicle, HC: Helicopter, SP: Swimmingpool, RA: Roundabout, SBF: Soccerballfield.

Supervision	Ship	ST	BD	TC	BC	GTF	Bridge	LV	SV	HC	SP	RA	SBF	Plane	Harbor	Mean
Weak	55.36	47.98	73.10	78.81	55.32	56.15	28.22	51.76	28.57	27.05	41.37	62.74	68.84	69.18	42.94	55.34
Semi	56.85	49.62	74.62	80.64	56.56	57.99	29.61	53.10	30.44	28.51	43.26	64.80	70.10	70.12	44.09	56.76
Full	59.74	50.49	76.98	84.21	57.92	59.57	32.88	54.80	33.75	31.29	44.74	66.03	72.13	75.84	45.68	59.05

Results of weakly-supervised semantic segmentation on mapping challenge dataset. We compare our proposed method with existing state-of-the-art weakly supervised semantic segmentation approaches on mapping challenge dataset. As shown in Table 7, our method achieves 75.65% mIoU on mapping challenge dataset validation set. Our method outperforms Rafique et al. [41], around 1.31% in mIoU, 0.64% in OA, 0.84% in TPR and 0.33% in TNR. Figure 9 shows the segmentation results of our method. The results indicate that our proposed method is effective in different datasets.

Results of semi-supervised semantic segmentation on iSAID dataset. We also do semi-supervised semantic segmentation experiments and compare to state-of-the-art approaches. In semi-supervised task, 141 pixel-level labels, 1/10 of the training sets, are added for training. As shown in Table 5, our proposed method outperforms all the compared methods and achieves 56.76% mIoU, 98.62% OA, 63.25% TPR and 99.64% TNR. Specific IoU for per category can be found in Table 6. The results indicate that our method is still effective in semi-supervised condition and the performance is very close to the fully supervised model.

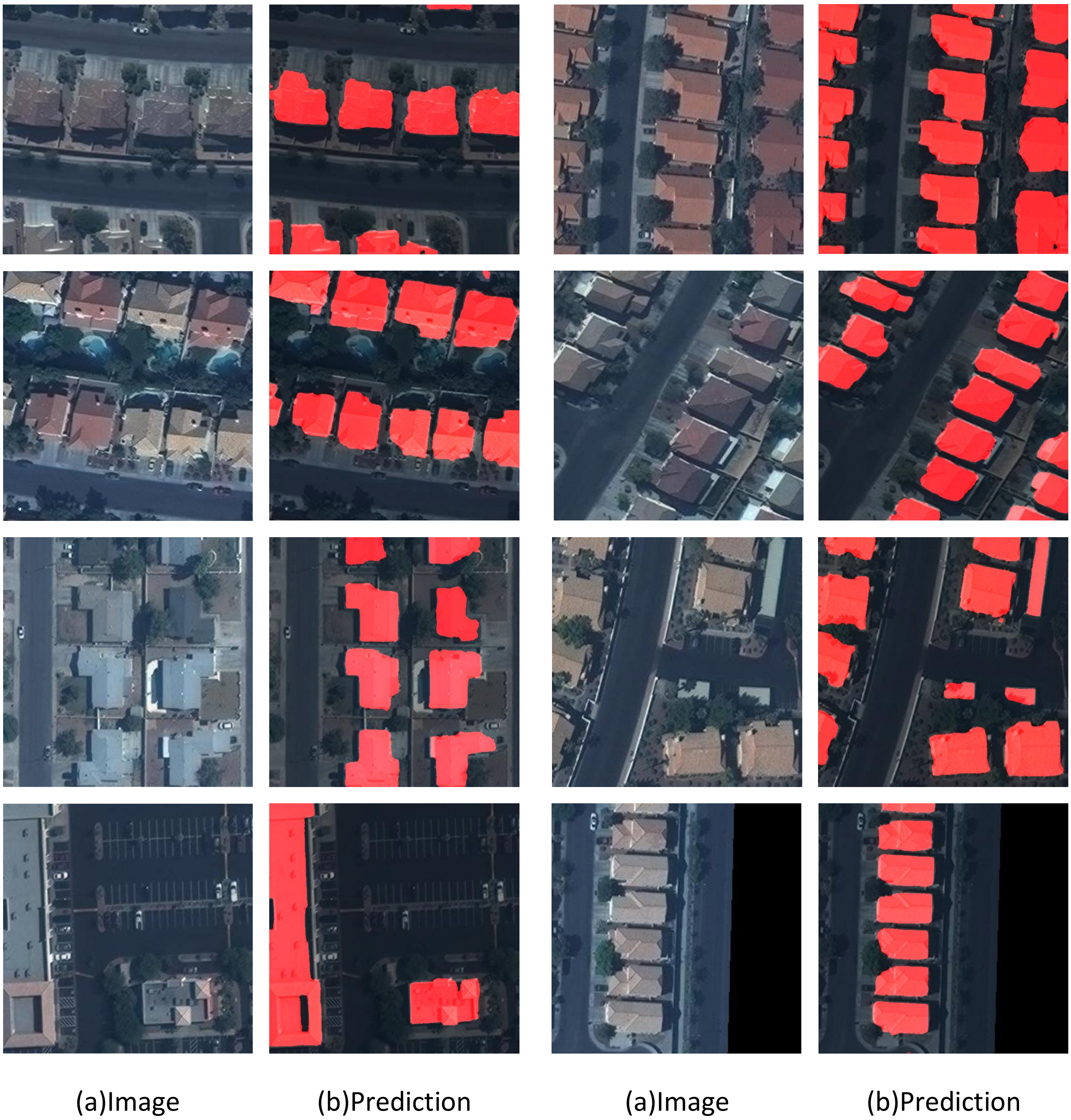

Figure 9. Examples of segmentation results of our method on mapping challenge dataset. (**a**) Image. (**b**) Prediction.

Table 7. Weakly supervised results on validation set of mapping challenge dataset. These methods are all box-based and only have bounding boxes annotations in the training process.

Methods	mIoU (%)	OA (%)	TPR (%)	TNR (%)
SDI [30]	73.57	85.67	88.76	87.70
Song et al. [27]	73.95	85.82	89.42	87.96
Rafique et al. [41]	74.34	86.31	90.10	88.24
Ours	75.65	86.95	90.94	88.57

5. Discussion

In this section, we further discuss: (1) The advantages of our method compared to the traditional methods, (2) the limits of our method and (3) potential improvement of the framework.

(1) The advantages of our method.

Learning strategy from noisy labels and the quality of proposals are two key problems of weakly supervised semantic segmentation in aerial images. We tackle these problems by taking advantage of multiple proposals in the training process and proposing two kinds of high quality proposals, GradCAM^D and GrabCut^C. The experimental results in Sections 4.2 and 4.3 prove that the proposed method can effectively improve the performance of weakly supervised semantic segmentation in aerial images.

(2) The limits of our method.

Our method needs bounding boxes annotations, which have two weaknesses in aerial images. On the one hand, bounding boxes annotations are slightly more expensive than image level annotations and points level annotations. On the other hand, bounding boxes annotations are not suitable for all semantic segmentation tasks in aerial images. For example, bounding boxes annotations represent airplanes, cars and buildings well but can not represent roads because roads are more similar to lines.

(3) Potential improvement of the framework.

As shown in Table 1, although our method improves segmentation results significantly by using combination of different proposals. The performance will not increase when adding all kinds of proposals. In particular, adding rectangle proposals or CRF proposals to $\{GrabCut, GradCAM^D, GrabCut^C\}$ will hurt the performance. We analyze that low quality of proposals will hurt the performance of segmentation model. In the ideal condition, we want our method can ignore most of the noise which is coming from noisy label. There are still many future works that can be done to handle the bad influence of noisy label.

Our combination strategy of multi-label is naive and can be improved by introducing more advanced statistical methods. Expectation-Maximization is elegant and we think it will contribute to experiments. We will try to realize it in future research.

6. Conclusions

In this paper, we propose a novel JMLNet, which first regards multiple proposals as multi-label supervision to train weakly supervised semantic segmentation model. JMLNet learns common knowledge from multiple noisy labels and prevents the model from overfitting one specific label. GradCAM^D and GrabCut^C methods are proposed to generate high-quality segmentation proposals, which further improve the segmentation performance of JMLNet. These proposals perform both reliable in the distinguished semantic area and detailed in instance edge. We report the state-of-the-art results on semantic segmentation tasks of iSAID and mapping challenge dataset when training using bounding boxes supervision, reaching comparable quality with the fully supervised model.

Author Contributions: Conceptualization, R.G., Investigation, R.G.; Formal analysis, R.G., Methodology, R.G.; Supervision, X.S., K.C., X.Z., Z.Y., W.D. and M.Y.; Visualization, R.G.; Writing—original draft, R.G.; Writing—review and editing, X.S., K.C., X.Z., Z.Y., W.D. and M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by the National Natural Science Foundation of China under Grants 41701508 and 61725105.

Acknowledgments: The authors would like to thank the anonymous reviewers for their very competent comments and helpful suggestions.

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

Reference

1. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
2. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
3. Zhang, W.; Huang, H.; Schmitz, M.; Sun, X.; Wang, H.; Mayer, H. Effective fusion of multi-modal remote sensing data in a fully convolutional network for semantic labeling. *Remote Sens.* **2018**, *10*, 52. [CrossRef]
4. Fu, K.; Lu, W.; Diao, W.; Yan, M.; Sun, H.; Zhang, Y.; Sun, X. WSF-NET: Weakly supervised feature-fusion network for binary segmentation in remote sensing image. *Remote Sens.* **2018**, *10*, 1970. [CrossRef]
5. Chai, Y.; Fu, K.; Sun, X.; Diao, W.; Yan, Z.; Feng, Y.; Wang, L. Compact Cloud Detection with Bidirectional Self-Attention Knowledge Distillation. *Remote Sens.* **2020**, *12*, 2770. [CrossRef]
6. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile, 11–18 December 2015.
7. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
8. Li, Y.; Qi, H.; Dai, J.; Ji, X.; Wei, Y. Fully convolutional instance-aware semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2359–2367.
9. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 325–341.
10. Fu, K.; Chang, Z.; Zhang, Y.; Xu, G.; Zhang, K.; Sun, X. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2020**, *161*, 294–308. [CrossRef]
11. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
12. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large kernel matters–improve semantic segmentation by global convolutional network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4353–4361.
13. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
14. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
15. Huang, Z.; Wang, X.; Wang, J.; Liu, W.; Wang, J. Weakly-supervised semantic segmentation network with deep seeded region growing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7014–7023.
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing And Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.

17. Wei, H.; Zhang, Y.; Wang, B.; Yang, Y.; Li, H.; Wang, H. X-LineNet: Detecting Aircraft in Remote Sensing Images by a Pair of Intersecting Line Segments. *IEEE Trans. Geosci. Remote. Sens.* **2020**. [CrossRef]
18. Sun, X.; Shi, A.; Huang, H.; Mayer, H. BAS^{4}Net: Boundary-Aware Semi-Supervised Semantic Segmentation Network for Very High Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **2020**, *13*, 5398–5413. [CrossRef]
19. Fu, K.; Chang, Z.; Zhang, Y.; Sun, X. Point-Based Estimator for Arbitrary-Oriented Object Detection in Aerial Images. *IEEE Trans. Geosci. Remote. Sens.* **2020**, 1–18. [CrossRef]
20. Yan, Z.; Yan, M.; Sun, H.; Fu, K.; Hong, J.; Sun, J.; Zhang, Y.; Sun, X. Cloud and cloud shadow detection using multilevel feature fused segmentation network. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1600–1604. [CrossRef]
21. Chen, K.; Fu, K.; Yan, M.; Gao, X.; Sun, X.; Wei, X. Semantic segmentation of aerial images with shuffling convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 173–177. [CrossRef]
22. 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, Zürich, Switzerland, 6–12 September 2014; pp. 740–755.
23. Dai, J.; He, K.; Sun, J. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1635–1643.
24. Papandreou, G.; Chen, L.C.; Murphy, K.P.; Yuille, A.L. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Las Condes, Chile, 11–18 December 2015; pp. 1742–1750.
25. Krähenbühl, P.; Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2011; pp. 109–117.
26. 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, Graz, Austria, 7–13 May 2016; pp. 549–565.
27. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3136–3145.
28. 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.* **2016**, *39*, 128–140. [CrossRef] [PubMed]
29. Rother, C.; Kolmogorov, V.; Blake, A. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM Transactions on Graphics (TOG)*; ACM: New York, NY, USA, 2004; Volume 23, pp. 309–314.
30. Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; Schiele, B. Simple does it: Weakly supervised instance and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 876–885.
31. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.
32. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
33. Kolesnikov, A.; Lampert, C.H. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2016; pp. 695–711.
34. Ahn, J.; Kwak, S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4981–4990.
35. Zhou, Y.; Zhu, Y.; Ye, Q.; Qiu, Q.; Jiao, J. Weakly supervised instance segmentation using class peak response. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3791–3800.

36. Lin, D.; Dai, J.; Jia, J.; He, K.; Sun, J. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3159–3167.
37. Tang, M.; Perazzi, F.; Djelouah, A.; Ben Ayed, I.; Schroers, C.; Boykov, Y. On regularized losses for weakly-supervised cnn segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 507–522.
38. Chen, J.; He, F.; Zhang, Y.; Sun, G.; Deng, M. SPMF-Net: Weakly Supervised Building Segmentation by Combining Superpixel Pooling and Multi-Scale Feature Fusion. *Remote Sens.* **2020**, *12*, 1049. [CrossRef]
39. Wang, S.; Chen, W.; Xie, S.M.; Azzari, G.; Lobell, D.B. Weakly supervised deep learning for segmentation of remote sensing imagery. *Remote Sens.* **2020**, *12*, 207. [CrossRef]
40. Wu, W.; Qi, H.; Rong, Z.; Liu, L.; Su, H. Scribble-Supervised Segmentation of Aerial Building Footprints Using Adversarial Learning. *IEEE Access* **2018**, *6*, 58898–58911. [CrossRef]
41. Rafique, M.U.; Jacobs, N. Weakly Supervised Building Segmentation from Aerial Images. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 3955–3958.
42. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2015; pp. 91–99.
43. Chen, L.C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.
44. 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 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 28–37.
45. Crowdai Mapping Challenge. 2018. Available online: https://www.crowdai.org/challenges/mapping-challenge (accessed on 28 March 2018).
46. 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 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
47. 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. Vis.* **2015**, *111*, 98–136. [CrossRef]
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
49. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2019; pp. 8024–8035.
50. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]