remote sensing

MDPI

Article

Deep Learning Network Intensification for Preventing Noisy-Labeled Samples for Remote Sensing Classification

Chuang Lin [1,2], Shanxin Guo [1,3,*], Jinsong Chen [1,3], Luyi Sun [1,3], Xiaorou Zheng [1,2], Yan Yang [4] and Yingfei Xiong [1,2]

[1] Center for Geo-Spatial Information, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; chuang.lin1@siat.ac.cn (C.L.); js.chen@siat.ac.cn (J.C.); ly.sun@siat.ac.cn (L.S.); xiaorou.zheng@siat.ac.cn (X.Z.); yf.xiong@siat.ac.cn (Y.X.)
[2] University of Chinese Academy of Sciences, Beijing 101407, China
[3] Shenzhen Engineering Laboratory of Ocean Environmental Big Data Analysis and Application, Shenzhen 518055, China
[4] Big Data Center of Geospatial and Natural Resources of Qinghai Province, Xining 810000, China; lolotus@outlook.com
* Correspondence: sx.guo@siat.ac.cn; Tel.: +86-755-8639-2331

Abstract: The deep-learning-network performance depends on the accuracy of the training samples. The training samples are commonly labeled by human visual investigation or inherited from historical land-cover or land-use maps, which usually contain label noise, depending on subjective knowledge and the time of the historical map. Helping the network to distinguish noisy labels during the training process is a prerequisite for applying the model for training across time and locations. This study proposes an antinoise framework, the Weight Loss Network (WLN), to achieve this goal. The WLN contains three main parts: (1) the segmentation subnetwork, which any state-of-the-art segmentation network can replace; (2) the attention subnetwork (λ); and (3) the class-balance coefficient (α). Four types of label noise (an insufficient label, redundant label, missing label and incorrect label) were simulated by dilate and erode processing to test the network's antinoise ability. The segmentation task was set to extract buildings from the Inria Aerial Image Labeling Dataset, which includes Austin, Chicago, Kitsap County, Western Tyrol and Vienna. The network's performance was evaluated by comparing it with the original U-Net model by adding noisy training samples with different noise rates and noise levels. The result shows that the proposed antinoise framework (WLN) can maintain high accuracy, while the accuracy of the U-Net model dropped. Specifically, after adding 50% of dilated-label samples at noise level 3, the U-Net model's accuracy dropped by 12.7% for OA, 20.7% for the Mean Intersection over Union (MIOU) and 13.8% for Kappa scores. By contrast, the accuracy of the WLN dropped by 0.2% for OA, 0.3% for the MIOU and 0.8% for Kappa scores. For eroded-label samples at the same level, the accuracy of the U-Net model dropped by 8.4% for OA, 24.2% for the MIOU and 43.3% for Kappa scores, while the accuracy of the WLN dropped by 4.5% for OA, 4.7% for the MIOU and 0.5% for Kappa scores. This result shows that the antinoise framework proposed in this paper can help current segmentation models to avoid the impact of noisy training labels and has the potential to be trained by a larger remote sensing image set regardless of the inner label error.

Keywords: noisy labels; deep learning; building extraction; attention network

Citation: Lin, C.; Guo, S.; Chen, J.; Sun, L.; Zheng, X.; Yang, Y.; Xiong, Y. Deep Learning Network Intensification for Preventing Noisy-Labeled Samples for Remote Sensing Classification. *Remote Sens.* **2021**, *13*, 1689. https://doi.org/10.3390/rs13091689

Academic Editor: Wai Chi Fang

Received: 30 March 2021
Accepted: 25 April 2021
Published: 27 April 2021

1. Introduction

With the ability to unify features at different image levels, deep-learning networks have achieved great success in the remote sensing field. However, a deep-learning network's performance mainly depends on (1) the size and variety of the training data and (2) the accuracy of the training labels. Compared to the former factor, errors in training labels are usually hard to identify and correct. Unlike the nature-image dataset in the computer-science field, noisy labels are more likely to occur in remote-sensed image

datasets. First, the main reasons are that the land-cover and land-use type characteristics may vary depending on different times, locations and sensors. Identifying accurate labels requires specific background knowledge [1]. Second, inconsistency of labels occurs when experts from different backgrounds make labels [2]. Third, the historic land-cover/land-use maps are usually used in automatic-labeling processing. The unreliability of labels can be caused by the time mismatch between these historical maps and the images. Designing a framework to intensify deep-learning networks to reduce the impact of these erroneous samples is a much greater challenge.

The fundamental reason why the noisy-labeled samples can quickly impact the current deep networks is that, in a deep learning network, the network parameters are updated by samples with a high value in the loss function. Unfortunately, high-loss values can both be caused by correct samples with variate features and noisy-labeled samples. In the original network, there is no mechanism to distinguish these two types of samples. As a result, when the number of noisy-labeled samples increases, the network parameters are updated in the wrong direction by learning these incorrectly labeled samples. Therefore, adding a structure to help the deep-learning network distinguish noisy-labeled samples from correct samples is key to increasing the network's robustness.

Recently, some attempts have been made to overcome this difficulty to distinguish the noisy-labeled samples. The strategy can be classified into two categories: (1) methods focused on labels and (2) methods focused on the loss function.

In the first category, there are three main types of methods: (1) the data-selection model, (2) the multiple-networks model and (3) the noise-transition-matrix model.

The first type is the data-selection model, which is to clean up the noisy labels by creating a mechanism. Learning the training samples step by step to clean up samples with a low-to-high loss value is one of the typical methods employed [3]. These methods assume that the noisy-labeled samples usually show extremely high loss values compared to the samples with correct labels. Based on this assumption, the network can be first trained by the samples with a low-loss value; then the high-loss samples can be added step by step to the network and the network's performance monitored at the same time. When the overall network loss is acceptable, the remaining high-loss samples are considered the errors. This assumption can hold when the variety of the samples is relatively low. However, a high-data variation over time or a satellite platform makes this strategy inappropriate. MORPH [4] introduces the concept of deep-learning-network memory samples, which updates the initial set of samples containing noisy labels to the sample set of clean samples by self-transition learning. The data selection can also be improved by the higher-order topological information of the data. Based on this idea, the TopoFilter [5] is proposed as a new noise-label-filtering method to detect clean-labeled samples. Semisupervised methods can also help to select the clean data. DivideMix [6] combines semisupervised learning with label-noise learning by leveraging two-component and one-dimensional Gaussian mixture models to distinguish the clean samples from the noisy samples to train the model with the semisupervised methods. RoCL [7] adopts a two-stage learning strategy to supervise the training of selected clean samples. Then semisupervised learning is performed to generate pseudo-labels by relabeling the noisy samples with the trained network.

Another clean-up strategy can also be created by two networks with different structures [8]. The idea behind this strategy is that a different structure shows a different capacity to avoid different types of noise. They can cooperate to improve the ability to reduce the influence of noisy-labeled samples. In this method, the key task is to create specific rules to distinguish the erroneous samples. Once the rules fail in one step during the training section, the network may refuse to learn from the remaining correct samples.

The noise-transition matrix is the third type of method in this category. In these models, a transition matrix or layer is added to the end of the network to provide a chance for each sample to transfer its label to others [9–11]. The matrix tries to correct the noisy data by quantifying the probability from one class to another. Most of the existing noise-transfer-matrix methods rely on estimating the noisy-class posterior. Still, the estimation

error in the noisy-class posterior leads to poor estimates of the transfer matrix. The accuracy of this transfer matrix depends on the estimation of how the noise data are distributed in different categories. Dual T [12] solves this problem by using a partitioning paradigm to decompose the matrix into two estimated matrices for multiplication to avoid directly estimating the noisy class posterior. However, in remote-sensing-classification problems, sometimes obtaining these kinds of matrices in practice is still a challenge.

The second category method focuses on how to decrease the impact of noisy data by reducing the loss values during the training process. One way to achieve this goal is to modify the loss function. The basic idea is to design a robust loss function that suffers less impact from the erroneous samples. After comparing the mean absolute error (MAE), mean square error (MSE) and categorical cross-entropy (CCE) loss functions, Ghosh et al. (2017) found that the MAE loss function is more robust to noise in multiple classification tasks [13]. Based on this finding, Zhang and Sabuncu (2018) provided a generalized cross-entropy (GCE) loss function, which combined the MAE function with CCE to achieve higher performance when erroneous data were included [14]. Later, inspired by the symmetric KL-divergence, symmetric cross-entropy learning (SL) [15] was proposed, boosting cross-entropy (CE) symmetrically with a noise-robust counterpart, reverse cross-entropy (RCE). The SL approach simultaneously addresses both the insufficient learning and overfitting problem of CE in the presence of noisy data. The main drawback of these models is that the different types of noise in data labels are difficult to remove by one universal-loss function, which causes the low flexibility of the models. Moreover, the more complex the loss function, the more time is consumed during the training stage [14].

This category's other approach is to assign each training sample with a consistently updated weight during each learning iteration. The weighting process can help the model avoid noisy samples to achieve an accuracy equaling the mode trained with clean data [16]. The critical issue in this category is how to calculate the weights for different samples. Many structures have been created to achieve this goal. For example, the multilayer perceptron (MLP) is used to estimate the weighting function [17] and for the implicit calculation of the weighting factor [18,19]. All these models are designed and applied in image-sense classification tasks. To the best of the authors' knowledge, there are no attempts to achieve this goal in the image-pixel classification task, especially in remote-sensing land-cover classification.

This article proposes a general network framework to prevent noisy-labeled data in remote-sensing image-classification tasks: the Weighted Loss Network (WLN). The WLN follows the idea of the loss-value weighting strategy in the second category mentioned above. We designed a sample weight-adjustment scheme in combination with the attention mechanism to evaluate each sample's importance through the learning process. The experiments were performed on the remote-sensing-image public data set of the Inria Aerial Image Labeling Dataset and compared with the original classification-network method. This paper's main contributions are summarized as follows: (1) we propose a general network framework that is robust for noisy-labeled training samples for remote-sensing image classification; (2) four commonly occurring types of label noise (an insufficient label, redundant label, missing label and incorrect label) are considered. The result shows the provided algorithm can maintain high accuracy and good generalization performance under different noise types.

2. Datasets and Experiment Design

2.1. Dataset

In this paper, the Inria Aerial Image Labeling Dataset was used [20]. This dataset was designed to address the automatic labeling of aerial images at the pixel level. The Inria dataset has an image resolution of 30 cm and labels two types of information: building categories and nonbuilding categories. Furthermore, these images cover different urban areas; not only big cities with dense buildings, but also small towns with few buildings. The dataset is available for download from the web page https://project.inria.fr/aerialimagelabeling/ (accessed on 6 June 2020).

The Inria training dataset contains 180 images of 5000 × 5000 size, covering five regions—Austin, Chicago, Kitsap County, Western Tyrol and Vienna—with a total area of 405 km^2. The labels consist of 180 single-channel images, where 255 indicates the building category and 0 indicates the nonbuilding category, and the specific images are shown in Figure 1. Since this dataset is used for a competition, the labels of the test set are not available, so in this study, the original training set was divided into three parts according to the ratio of 8:1:1 for the training set, the validation set and the test set, which were independent of each other and did not overlap.

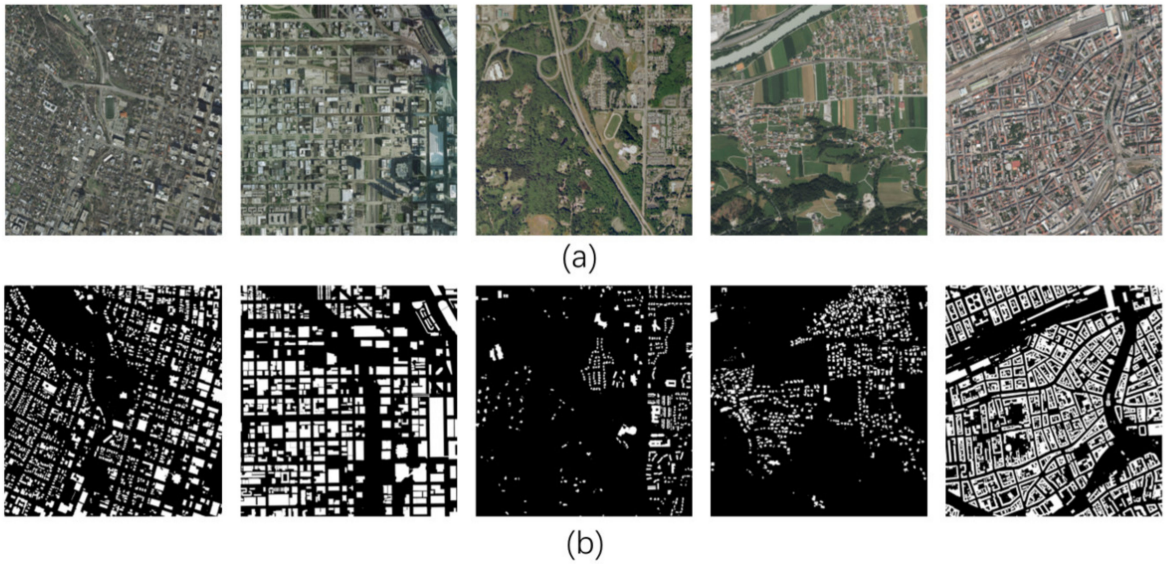

Figure 1. Examples from the Inria dataset, showing five sample image patches (**a**) and the corresponding ground truth (**b**) for Austin, Chicago, Kitsap County, Western Tyrol and Vienna.

2.2. Experiment Design

Erroneous samples with label noise are defined as anything that obscures the relationship between an instance's features and its class [21]. In pixel-level remote-sensing land-cover labeling, there are four typical errors: (1) an insufficient label, which means some parts of the true label are missing; (2) a redundant label, in which the labeled area is larger than the actual area; (3) a missing label, where the entire label of the object is missing; and (4) an incorrect label, which labels the object with the incorrect category.

To simulate these four types of erroneous samples, dilate and erode processing was used in this study. As shown in Figure 2, the first row represents the expansion processing, which simulates the redundant label (as showing in Figure 2c,d), which expanded the actual label by 9 × 9 and 17 × 17 kernels, respectively). When the kernel size increases to 25 × 25, the label contains many incorrect pixels and can be identified as an incorrectly labeled sample. Similarly, erosion processing was used to generate the insufficient label (shown in Figure 2f,g, which was processed by kernel sizes of 9 and 17, respectively) and the missing label was eroded by a 25 × 25 kernel size. (Figure 2h). To clarify the different errors, in this study, we set up three noise levels. We denoted the noise of 9 × 9 kernels as level-1 noise, 17 × 17 as kernels level-2 noise and 25 × 25 kernels as level-3 noise.

Five noise rates were set in this study to test how the network performed under different numbers of erroneous samples. We contaminated the training dataset with 0%, 25%, 35%, 45% and 50% of erroneous samples randomly. In this study, we used PyTorch as the deep-learning framework (https://pytorch.org/, accessed on 5 March 2021) in our experiments. The input image was 256 × 256 pixels. The learning rate was set as 0.0001, and the batch size was fixed to 16 in our experiments.

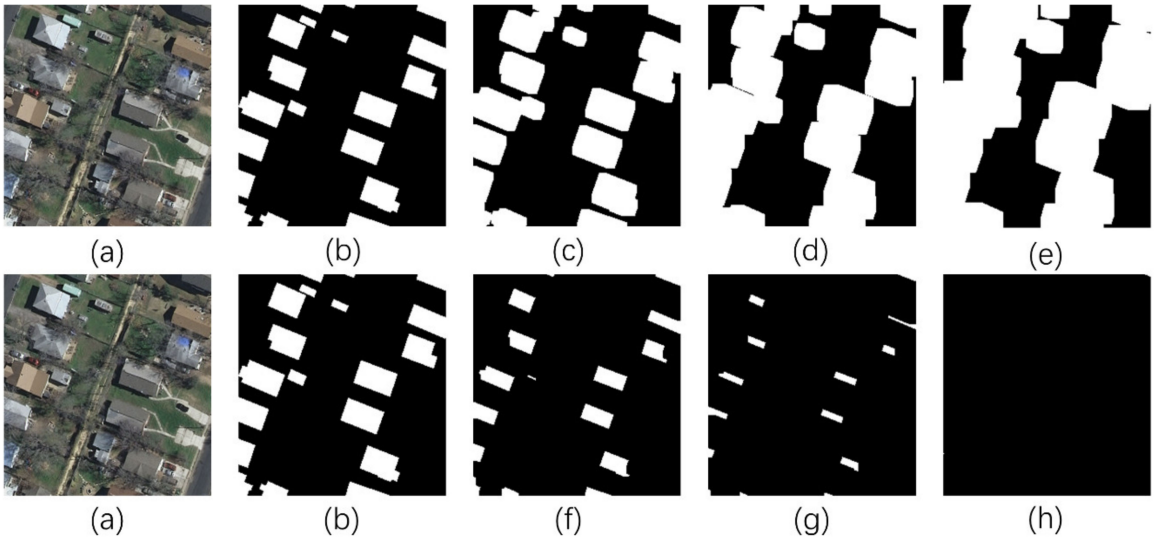

Figure 2. The erroneous-sample generation with the expansion and erosion processing. The first row is expansion results that represent (**a**) the image, (**b**) the ground-truth, (**c**) a redundant label with kernel size 9, (**d**) a redundant label processed by kernel size 17, and (**e**) the inc label processed by kernel size 25. The erosion processing is showing in the second row: (**f**) insufficient label with kernel 9, (**g**) insufficient label by kernel 17, and (**h**) missed label processed by kernel size 25.

3. Methodology

3.1. Study Workflow

In this paper, we designed a new end-to-end antinoise algorithm, the Weighted Loss Network (WLN), to assign weights to all training samples' loss values and iteratively update these weights during the training process by improving the cross-entropy loss. Figure 3 shows an overview of the workflow of this study.

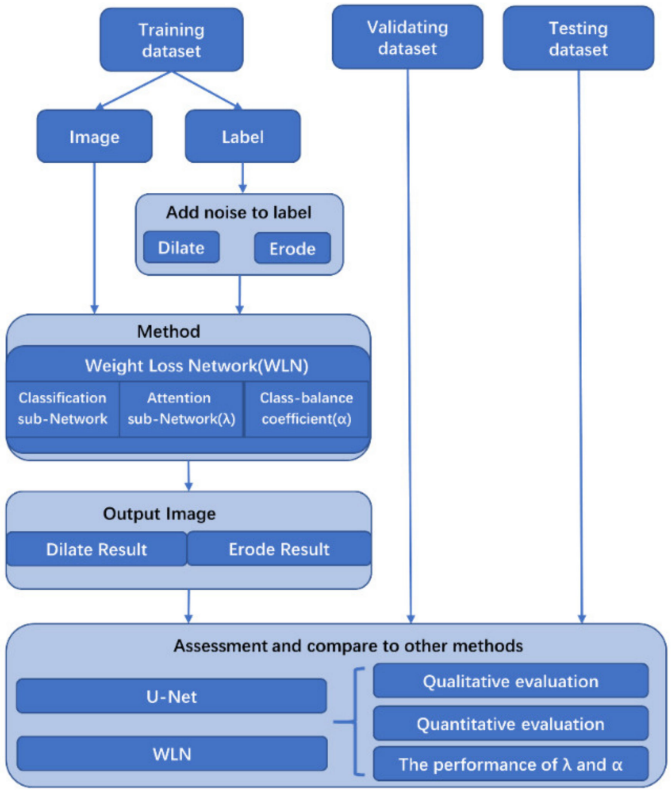

Figure 3. An overview of the workflow of this study. WLN: Weighted Loss Network.

As shown in Figure 3, during training the noisy-labeled samples were added under 0%, 25%, 35%, 45% and 50% noise rates with three noise levels (kernel sizes of 9, 17 and 25) of dilation or erosion processing. The Weighted Loss Network is a universal antinoise framework with three components: (1) a classifier network (we use U-Net as the classifier in this study), (2) an attention subnetwork (λ), and (3) a class-balance part (α). The classifier network can be changed for other deep-learning-based networks depending on the different applications. To evaluate the performance of the WLN and the λ and α part, the original segmentation network (U-Net) and U-Net with the Attention Subnetwork (λ) were qualitatively and quantitatively compared.

3.2. Modified Cross-Entropy Loss Function

The cross-entropy measures the predicted probability distribution $p(k|x)$ in comparison to the actual probability distribution $q(k|x)$, where $x \in \mathbb{R}^d$ denotes the d-dimensional input space and $k \in \{1, \ldots, K\}$ represents the K labels. The cross-entropy loss, serving as a loss function, is commonly used in deep-learning models, mainly for classification tasks, and the activation function is used in the output layer to model probabilities (Sigmoid) or distributions (Softmax). Unlike the squared-error cost function, which is commonly used in regression tasks, the cross-entropy loss overcomes the vanishing gradient problem when a neuron has a value of activation function (sigmoid function) close to 0 or 1 and avoids a reduction of speed. The cross-entropy is defined as

$$\mathcal{L}_{ce} = -\sum_{k=1}^{K} q(k|x) log(p(k|x)) \tag{1}$$

For the binary classification problem, the value of the label is 0 or 1. Therefore, the cross-entropy loss of n pixels in m images in a batch of training can be defined as

$$l_{ce} = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{n}\sum_{j=1}^{n} -\left[y_{ij} * log(p_{ij}) + (1 - y_{ij}) * log(1 - p_{ij})\right]\right) \tag{2}$$

where $y_{ij} \in \{0, 1\}$ is the group-truth label of the jth pixel of the ith label and $p_{ij} \in [0, 1]$ is the probability of the jth pixel of the ith label.

However, when noise contaminates labels, the cross-entropy loss will show an insufficient learning problem, especially for complex learning categories with a high feature diversity [15]. This problem's key cause is that every training sample is treated as equally contributing to the total loss. The noisy-labeled samples cause a high loss value, which will cause the weight to be updated in the wrong direction. The intuitive idea is to design a mechanism to evaluate each sample with different weights and reduce the noisy-labeled samples' influence on the loss.

Two components were added to the original cross-entropy function to achieve this goal. The first one was the weight λ for each sample, calculated by the attention subnetwork, to distinguish the correct samples from the erroneous samples. The second component was the class-balance coefficient α, which balanced the number of pixels in each sample to inhibit the situation in which one class dominated other classes during the training. The details of these two components are further explained in the following sections.

After adding the two components to the original cross-entropy loss function, the new function is defined as

$$l_{mce} = \frac{1}{m}\sum_{i=1}^{m}\left(\lambda_i * \frac{1}{n}\sum_{j=1}^{n} \alpha_{ij} * \left(-\left[y_{ij} * log(p_{ij}) + (1 - y_{ij}) * log(1 - p_{ij})\right]\right)\right) \tag{3}$$

where l_{mce} represents the modified cross-entropy loss, λ_i is the weight of the ith sample loss in a batch of m training samples, and $\sum_{i=1}^{m} \lambda_i = 1$. α_{ij} is the parameter of the class balance of the jth pixel of the ith class.

3.3. Attention Subnetwork (λ)

The attention network is designed to give the algorithm the ability to capture the most efficient feature to distinguish objects or classes in the same manner as a human. First, we proposed building global dependencies of inputs and applying them to speech translation [22,23]. As the goal of the attention network is to find the important features of the input, in this study, we adapt this characteristic to determine the weights of each sample in the training process. After reweighting the samples, the segmentation module backpropagates based on the reweighted loss. It focuses more on those with a higher weight, while the attention network backpropagates on the same loss. Figure 4 illustrates the overall structure of the WLN.

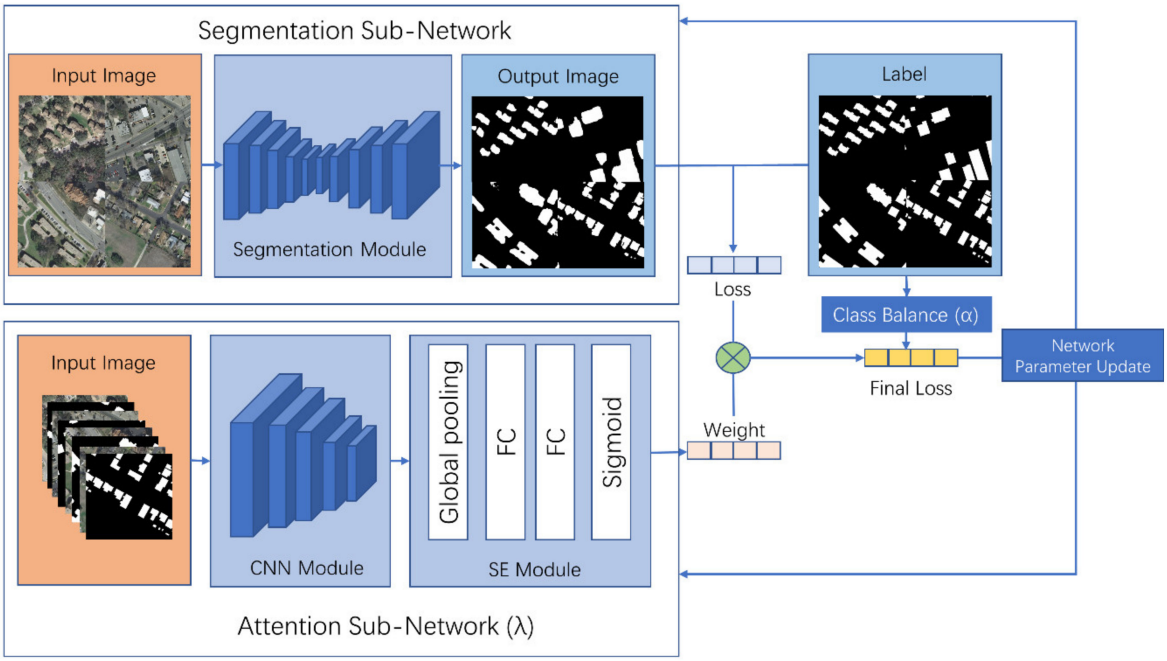

Figure 4. The overall structure of the proposed Weighted Loss Network (WLN).

The overall structure of the WLN is shown in Figure 4 with (1) the segmentation or classifier subnetwork, (2) the attention subnetwork (Figure 5) and (3) the class-balance coefficient. The segmentation module is the deep-learning network for classification. The attention subnetwork is a convolutional neural network (CNN) structure network with an attention module that takes the image's concatenation and its labels as an input and runs in parallel with the segmentation subnetwork. The attention subnetwork generates a weight (λ) for each input sample to reweight the samples. The class-balance coefficient (α) solves the imbalance of the label class to avoid insufficient learning for classes with few samples. The final cross-entropy loss (Equation (3)), combined with the contribution of λ and α, is then summed up and backpropagated to update the segmentation and attention subnetwork in combination.

In this study, we selected U-Net's structure [24] as the segmentation subnetwork, which other segmentation models can also replace. The attention module's implementation form adopts the squeeze-and-excitation (SE) module [25], which is shown in Figure 6. The SE module was proposed to consider the relationships among channels in the feature map and provides different weights for each channel. It automatically acquires the importance of each feature channel by learning and then strengthens useful features while suppressing those not useful to the current task according to each channel's importance. The squeeze and excitement are two critical operations in the SE module. These two operations help the SE model to capture channel-wise dependencies and significantly reduce the number of parameters and calculations [25].

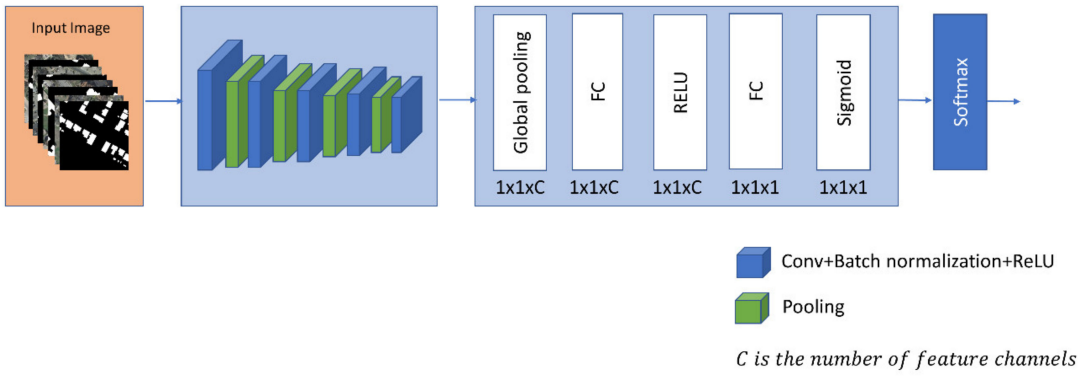

Figure 5. The structure of the attention subnetwork.

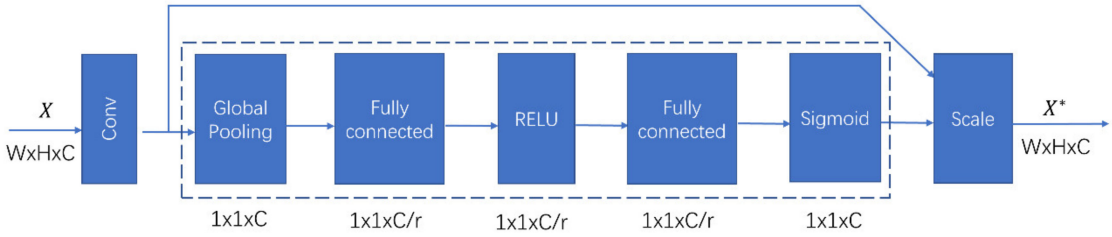

Figure 6. The SE module architecture [25].

In this study, we adopt the SE module as an attention module to obtain the importance between channels and evaluate the importance of each image-training sample. The attention module takes the CNN network output as an input, and we obtain the global-feature information representing the training sample through the global-average pooling layer. Two fully connected layers are used to fuse the characteristic information of this batch of samples. Finally, the output is normalized by a sigmoid-activation function and taken as the weight for each training-sample batch. This output can provide different weights for each training-sample loss by multiplication, and the training-sample information is either strengthened or suppressed.

3.4. Class-Balance Component (α)

Generally, when the network can choose which sample to learn, it first tends to choose those easy-learning samples with a low feature diversity of labels as the important samples and then choose the samples that are difficult to learn. Thus, the diversity of samples is the key to avoiding model overfitting. However, in practice, for some label errors, such as insufficient labels and missing labels, there may be many more pixels of the background class than the target class. This imbalance reduces the target class's feature diversity and makes the whole network focus on learning the background pixels. This imbalance of classes may cause two problems: (1) the target class's training is insufficient for those classes with few representative pixels; (2) the recall will decrease.

To overcome this problem, we created a class-balance coefficient α in the cross-entropy to balance the classes in each sample. The definition of α is shown as follows:

$$\alpha_i = \frac{l_0}{l_i} \tag{4}$$

where l_0 is the sum of the number of pixels of the background class in the label and l_i is the sum of the number of pixels of the ith $(i = 0, 1, 2, \ldots, C)$ class in the label. As we can see, the class-balance component (α) of the background always remains 1.

Figure 7 illustrates two simple examples of the class-balance component (α). The upper line shows the case of the insufficient label where the background (labeled as 0) is much better represented than the target class (labeled as 1). In this case, before adding the α weight, the contributions of the target class and the background class to this image's loss value are 4/16 and 12/16, respectively. After adding the α weight, these two classes are balanced, equally contributing to the image loss value at 12/16. The same results are shown in the case of the second line of the redundant label. As we can see, the key function of the class-balance component (α) is to take one class as the reference (in most cases, it is the background class) and then adjust the other classes to make them equally important to the reference class. This will help the network to learn each class equally and avoid the prementioned problems caused by an imbalance between the different classes.

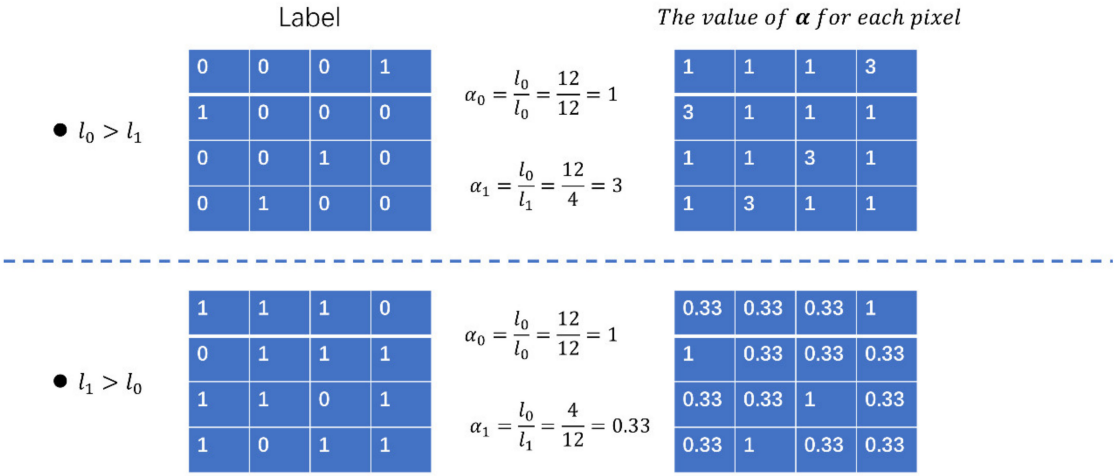

Figure 7. Two simple examples of the class-balance component (α).

3.5. Assessment

The accuracy evaluation metrics in this paper include (1) the overall accuracy, (2) the Cohen's Kappa coefficient and (3) the Mean Intersection over Union (MIOU).

The overall accuracy is defined as the number of correctly classified pixels over the total number of pixels. It is intuitive and straightforward but may fail to assess the performance thoroughly when the number of samples for different classes varies significantly.

The Cohen's Kappa coefficient is more robust, as it considers the probability of agreements occurring randomly. Let p_0 be the probability of correctly classified pixels, and p_e be the expected probability of agreement when the classifier assigns class labels by chance; then, Cohen's Kappa coefficient is defined as Equation (7):

$$K = \frac{p_0 - p_e}{1 - p_e} \tag{5}$$

Usually, we characterize $K < 0$ as no agreement, $[0, 0.20]$ as poor agreement, $[0.20, 0.40]$ as fair agreement, $[0.40, 0.60]$ as moderate agreement, $[0.60, 0.80]$ as good agreement, and $[0.80, 1]$ as almost perfect agreement.

MIOU is an important index to measure the segmentation accuracy in the field of computer-vision image segmentation. It represents the ratio of the intersection and union between the real value and the predicted value. MIOU is defined as Equation (8):

$$MIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{6}$$

where $k+1$ is the number of classifications (including background classes). *TP*, *FP*, and *FN* correspond to the number of true positive, false positive and false negative pixels for that class, respectively. An MIOU reaches its best value at 1 and its worst at 0.

4. Results

4.1. Performance with Dilated Samples

In this study, two-sample processing was used to simulate the four types of label noise. As mentioned before, a redundant label and an incorrect label were simulated by the dilating process. The eroding process produced the insufficient label and the missing label. To simplify these four types of label noise, we use the phrase "dilated samples" to represent the redundant label and incorrect label and "eroded samples" for the insufficient and missing label.

We trained the network on the training set with different noise levels of dilated labels and tested its performance on the clean labels. Table 1 and Figure 8 show the assessment of each method on the independent testing datasets. From the accuracy graph, both U-Net and WLN can be seen to maintain high accuracy for clean datasets. With the increasing noise rate, the network accuracy does not drop immediately. Surprisingly, the deep learning network's natural noise resistance is shown by the performance of the original U-Net model under Level 1 noise. When the noise rate and the noise level are relatively low, the U-Net method can still maintain high accuracy, which may be because the convolutional-neural network has specific noise resistance. However, with a noise rate of 50%, the accuracy of U-Net drops significantly, and at noise level 2, OA decreases by 4.4%, MIOU by 7.7%, and Kappa by 2.0%; at noise level 3, OA decreases by 12.7%, MIOU by 20.7%, and Kappa by 13.8%. In contrast, WLN can still maintain high accuracy even when the noise rate increases; with a noise rate of 50%, OA decreases by 1.2%, MIOU decreases by 2.1%, and Kappa decreases by 1.1% at noise level 2; at noise level 3, OA decreases by 0.2%, MIOU decreases by 0.3%, and Kappa decreases by 0.8%.

Table 1. Comparison of the performance of WLN and U-Net with dilated samples. MIOU: Mean Intersection over Union.

			Noise Rate (Changing Rate Compared to No Noise)				
Noise Level	**Method**		**No Noise**	**25%**	**35%**	**45%**	**50%**
Noise Level 1	U-Net	OA	0.907	0.903 (−0.4%)	**0.905** (−0.2%)	**0.897** (−1.1%)	0.896 (−1.2%)
		MIOU	0.796	0.790 (−0.8%)	**0.793** (−0.4%)	**0.780** (−2.0%)	0.779 (−2.1%)
		Kappa	0.783	0.780 (−0.4%)	0.783 (0.0%)	**0.790** (+0.9%)	**0.789** (+0.8%)
	WLN	OA	0.905	0.901 (−0.4%)	0.889 (−1.8%)	0.882 (−2.5%)	0.900 (−0.6%)
		MIOU	0.795	0.789 (−0.8%)	0.769 (−3.3%)	0.758 (−4.7%)	0.785 (−1.3%)
		Kappa	**0.798**	**0.801** (+0.4%)	**0.79** (−1.0%)	0.777 (−2.6%)	0.766 (−4.0%)
Noise Level 2	U-Net	OA	0.907	0.908 (+0.1%)	0.902 (−0.6%)	0.881 (−2.9%)	0.867 (−4.4%)
		MIOU	0.796	**0.799** (+0.4%)	**0.790** (−0.8%)	0.756 (−5.0%)	0.735 (−7.7%)
		Kappa	0.783	0.790 (+0.9%)	0.793 (+1.3%)	0.782 (−0.1%)	0.767 (−2.0%)
	WLN	OA	0.905	0.900 (−0.6%)	0.894 (−1.2%)	**0.896** (−1.0%)	**0.894** (−1.2%)
		MIOU	0.795	0.788 (−0.9%)	0.778 (−2.1%)	**0.783** (−1.5%)	**0.778** (−2.1%)
		Kappa	**0.798**	0.799 (+0.1%)	0.789 (−1.1%)	0.790 (−1.0%)	**0.789** (−1.1%)
Noise Level 3	U-Net	OA	0.907	0.909 (+0.2%)	**0.907** (0.0%)	0.880 (−3.0%)	0.792 (−12.7%)
		MIOU	0.796	0.801 (+0.6%)	**0.798** (+0.3%)	0.754 (−5.3%)	0.631 (−20.7%)
		Kappa	0.783	0.787 (+0.5%)	0.797 (+1.8%)	0.784 (+0.1%)	0.675 (−13.8%)
	WLN	OA	0.905	0.903 (−0.2%)	0.890 (−1.7%)	**0.907** (+0.2%)	**0.903** (−0.2%)
		MIOU	0.795	0.792 (−0.4%)	0.774 (−2.6%)	**0.800** (+0.6%)	**0.793** (−0.3%)
		Kappa	**0.798**	**0.798** (0.0%)	0.788 (−1.3%)	**0.792** (−0.8%)	**0.792** (−0.8%)

Bold represents a value more significant than 1% compared to the other method.

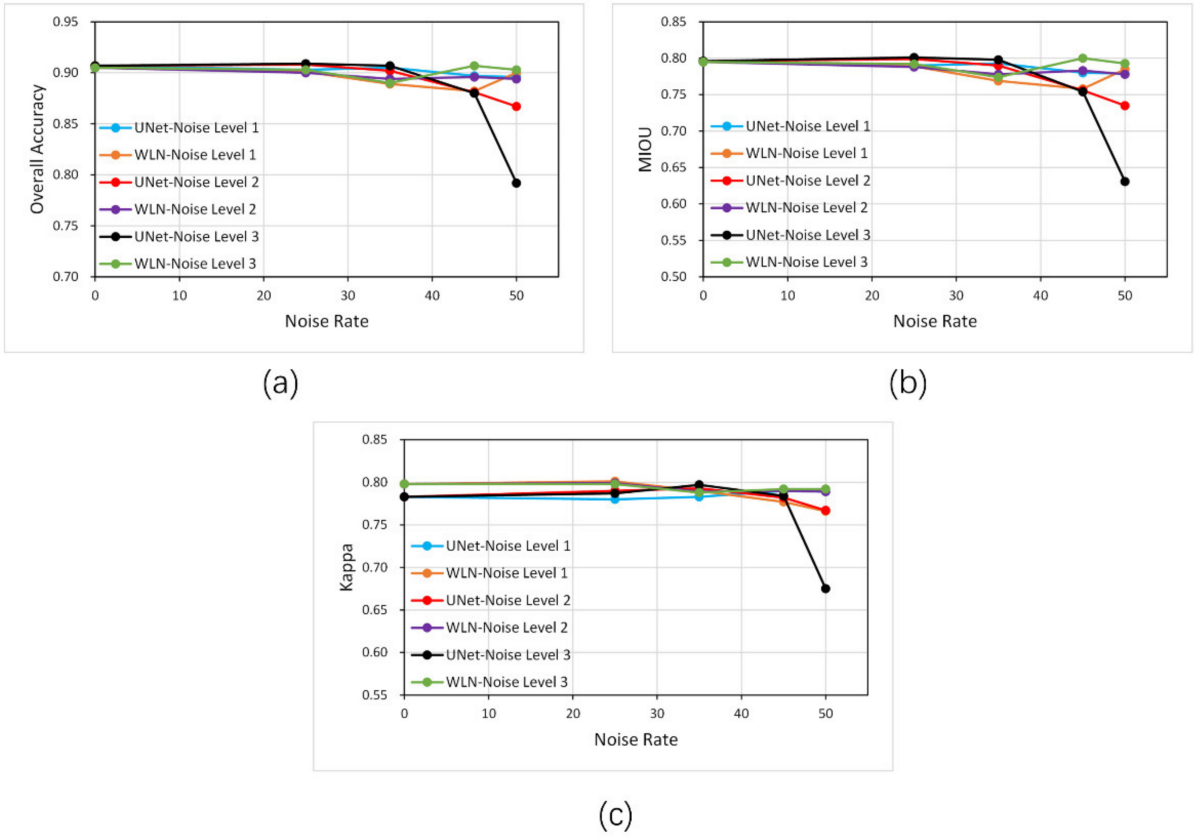

Figure 8. Accuracy plots of different dilated- noise rates and levels on the test set. (**a**) Overall accuracy, (**b**) MIOU and (**c**) Kappa.

4.2. Performance with Eroded Samples

Table 2 and Figure 9 show the assessment of U-Net and the WLN with eroded samples. It can be seen from the accuracy curves that the accuracy of the network changes with the eroded noise, similar to the accuracy changes in the dilated noise. The U-Net network shows noise resistance when the noise rate and the noise level are low due to its noise immunity. As the noise rate increases (at noise 3 with a 50% noise rate), the OA accuracy of the U-Net network on the test set decreases by 8.4%, MIOU decreases by 24.2%, and Kappa decreases by 43.3% in this process, which is a considerable variation. In contrast, WLN has better stability: OA only changes by 4.5%, MIOU changes by 4.7% and Kappa changes by 0.5% in this process. This result shows that the proposed WLN has higher noise resistance capability, especially when the noise rate and level are high.

Table 2. Comparison of the performance of WLN and U-Net with eroded samples.

			Noise Rate (Changing Rate Compared to No Noise)				
Noise Level	**Method**		**No Noise**	**25%**	**35%**	**45%**	**50%**
Noise Level 1	U-Net	OA	0.907	0.907 (0.0%)	0.907 (0.0%)	0.892 (−1.7%)	0.894 (−1.4%)
		MIOU	0.796	0.789 (−0.9%)	0.788 (−1.0%)	0.777 (−2.4%)	0.756 (−5.0%)
		Kappa	0.783	0.746 (−4.7%)	0.742 (−5.2%)	0.727 (−7.2%)	0.686 (−12.4%)
	WLN	OA	0.905	0.906 (+0.1%)	0.907 (+0.2%)	**0.911** (+0.7%)	**0.908** (+0.3%)
		MIOU	0.795	0.797 (+0.3%)	0.799 (+0.5%)	**0.804** (+1.1%)	**0.798** (+0.4%)
		Kappa	**0.798**	**0.803** (+0.6%)	**0.806** (+1.0%)	**0.791** (−0.9%)	**0.786** (−1.5%)

Table 2. *Conts.*

Noise Level	Method		Noise Rate (Changing Rate Compared to No Noise)				
			No Noise	25%	35%	45%	50%
Noise Level 2	U-Net	OA	0.907	0.910 (+0.3%)	0.906 (−0.1%)	0.901 (−0.7%)	0.868 (−4.3%)
		MIOU	0.796	0.798 (+0.3%)	0.790 (−0.8%)	0.772 (−3.0%)	0.691 (−13.2%)
		Kappa	0.783	0.771 (−1.5%)	0.757 (−3.3%)	0.711 (−9.2%)	0.587 (−25.0%)
	WLN	OA	0.905	0.903 (−0.2%)	0.903 (−0.2%)	0.900 (−0.6%)	**0.902** (−0.3%)
		MIOU	0.795	0.793 (−0.3%)	0.791 (−0.5%)	**0.787** (−1.0%)	**0.785** (−1.3%)
		Kappa	**0.798**	**0.805** (+0.9%)	**0.801** (+0.4%)	**0.794** (−0.5%)	**0.790** (−1.0%)
Noise Level 3	U-Net	OA	0.907	0.910 (+0.3%)	**0.908** (+0.1%)	0.899 (−0.9%)	0.831 (−8.4%)
		MIOU	0.796	**0.798** (+0.3%)	**0.792** (−0.5%)	0.768 (−3.5%)	0.603 (−24.2%)
		Kappa	0.783	0.773 (−1.3%)	0.746 (−4.7%)	0.703 (−10.2%)	0.444 (−43.3%)
	WLN	OA	0.905	0.903 (−0.2%)	0.890 (−1.7%)	0.892 (−1.4%)	**0.891** (−1.5%)
		MIOU	0.795	0.771 (−3.0%)	0.779 (−2.0%)	0.760 (−4.4%)	**0.758** (−4.7%)
		Kappa	**0.798**	**0.797** (−0.1%)	**0.784** (−1.8%)	**0.794** (−0.5%)	**0.794** (−0.5%)

Bold shows that the value is more significant than 1% compared to the other method.

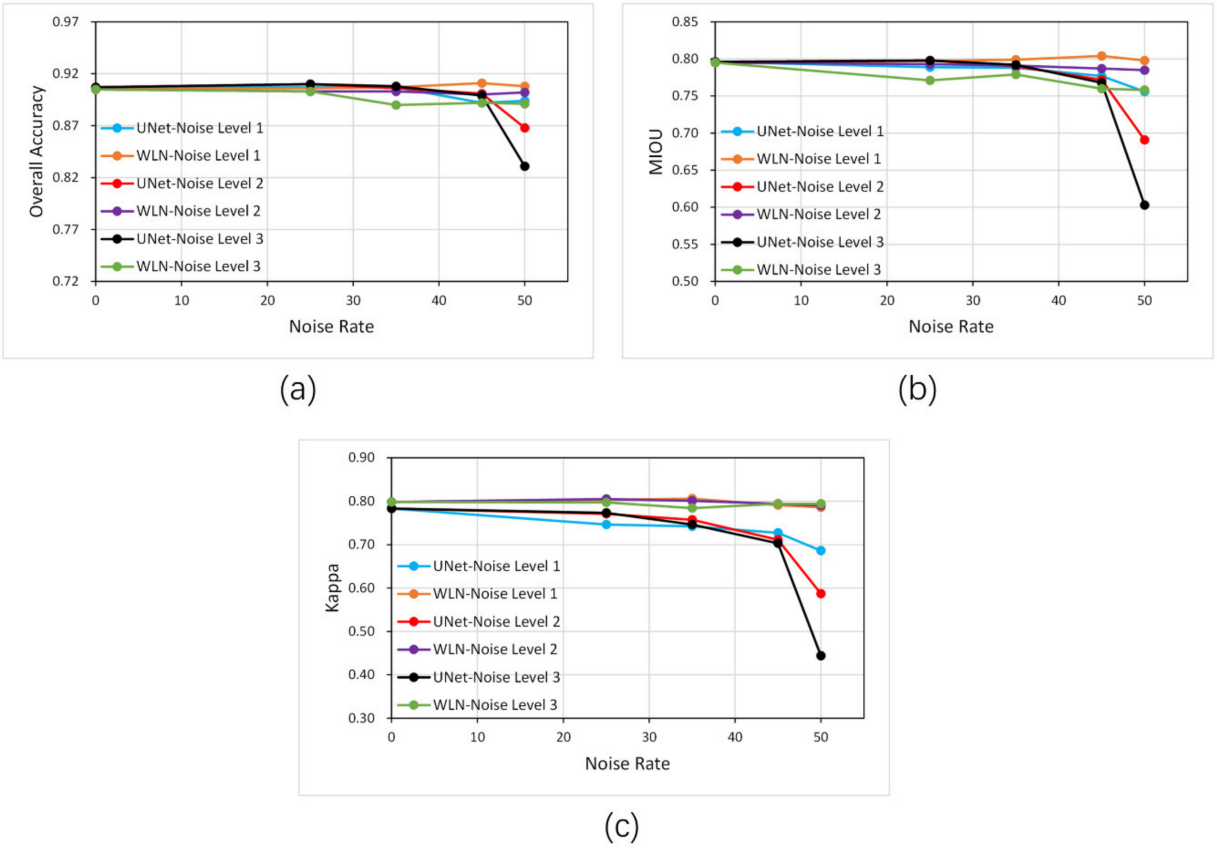

Figure 9. Accuracy plots of different eroded noise levels on the test set. (**a**) the overall accuracy, (**b**) MIOU and (**c**) Kappa.

4.3. Visual Assessment of the WLN with U-Net

In this subsection, we present the WLN and U-Net assessment to evaluate the extracted details with dilated and eroded samples utilizing visual interpretation.

4.3.1. Dilated-Noise Samples

Figure 10 shows the two methods' building extraction results for different noise levels with a 50% noise rate with dilated noise. The first and second rows show the extraction results for noise level 1, the third and fourth rows show the extraction results for noise level 2, and the fifth and sixth rows show the extraction results for noise level 3. From the figure, we can see that when the noise level is low, such as at level 1, WLN and U-Net's extraction results are similar. However, as shown in the red box in the figure, U-Net classified some nonbuilding pixels as buildings due to the influence of the dilated noise. When the noise level is high, such as at noise level 3, U-Net is affected by the dilated noise and can only extract the rough outline for buildings, incorrectly marking the streets as buildings. In contrast, WLN is less affected by noise and can identify buildings and the boundaries between buildings.

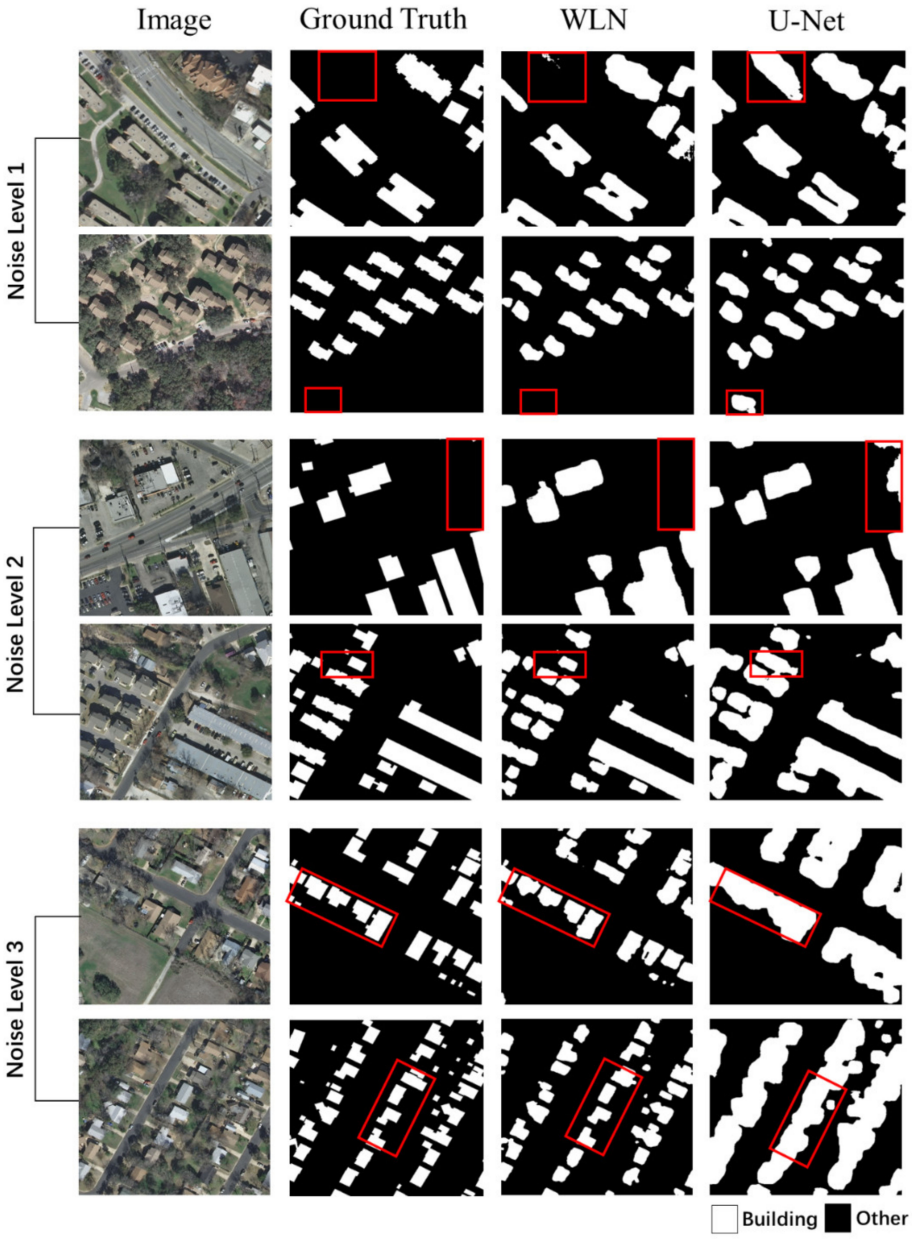

Figure 10. Extraction results of WLN and U-Net under different noise levels with a 50% noise rate of dilated noise.

4.3.2. Erode

Figure 11 shows the WLN and U-Net results with eroded noise with different noise levels at a 50% noise rate. The first and second rows show the extraction results of noise level 1, the third and fourth rows show the extraction results of noise level 2, and the fifth and sixth rows show the extraction results of noise level 3. From the figure, we can see that when the noise level is relatively low, such as at noise level 1, WLN and U-Net's extraction results are relatively complete. However, U-Net classifies some building pixels as nonbuildings due to the influence of eroding noise, as shown in the red box in the figure. With the increase in the noise level, this influence becomes more significant; for example, at noise level 3, U-Net can no longer classify buildings. At the same time, WLN can still obtain good classification results even at a high noise level.

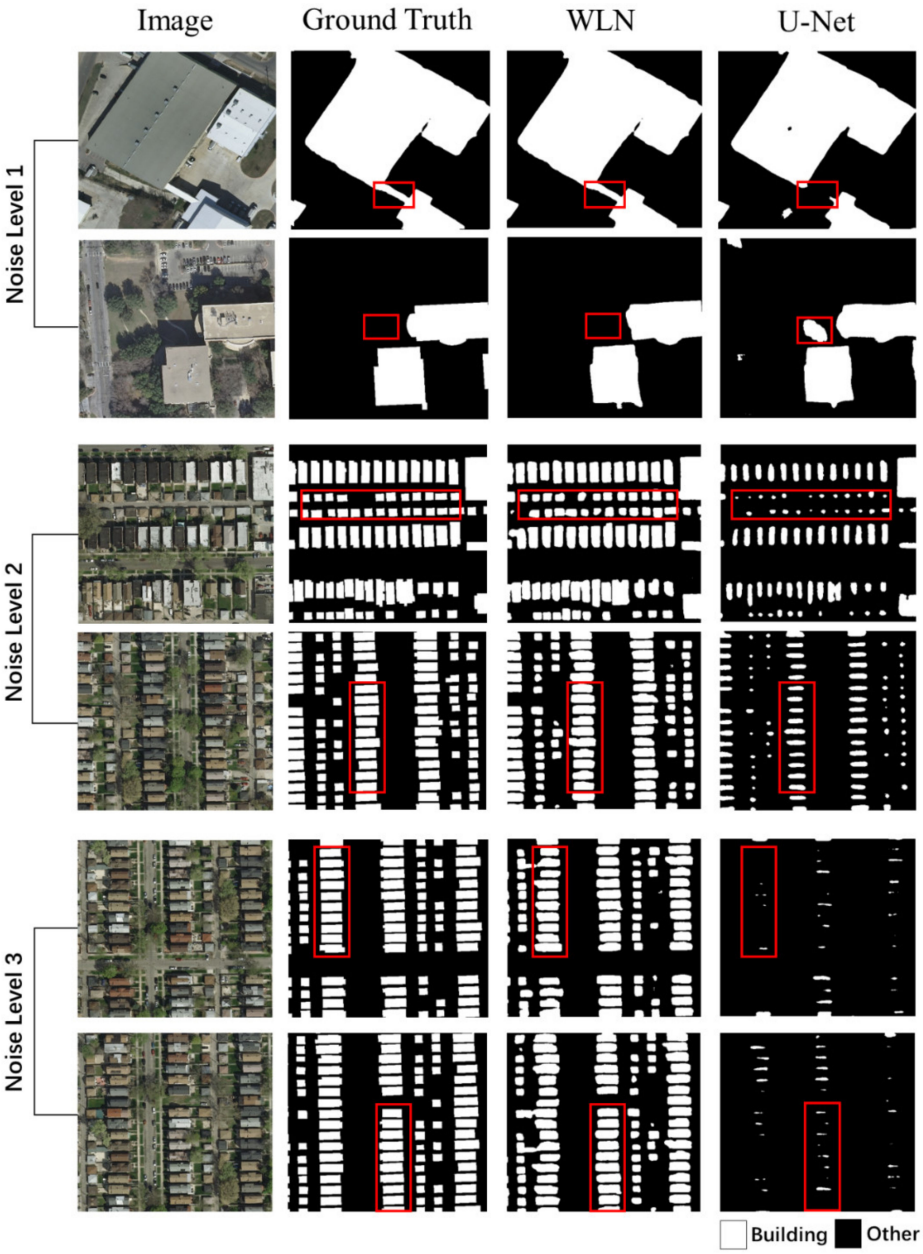

Figure 11. Extraction results of WLN and U-Net under different noise levels with a 50% noise rate of eroded noise.

5. Discussion

Noisy labels indirectly affect the direction of the process of updating the parameters of the network by influencing the convolutional-neural-network's loss value. Therefore, to improve the noise-immunity performance of the network, the key lies in the avoidance of the influence of noise on the network. In this study, the impact of noisy labels on the network is avoided by weighting the training samples, which is done to reduce the weight of loss values calculated from the erroneous samples. To achieve this goal, two parameters—the loss weight λ, and the class-balance coefficient α—were designed to add to the original cross-entropy loss function. In this section, we discuss the impact of λ and α on the noise immunity of the network.

5.1. The Local Optimum Problem When We Allow the Network to Choose What to Learn

The attention subnetwork (λ) takes an image with a label as input and generates the loss weight λ for each training sample through the CNN module and the SE module. However, we found a critical limitation if it only introduced the attention subnetwork to the framework. Figure 12 shows the average weights of clean-label samples and noisy-label samples during the training processing with a 50% noise rate. Figure 12a shows the average weights of the dilated noise, and Figure 12b shows the weights for the eroded noise.

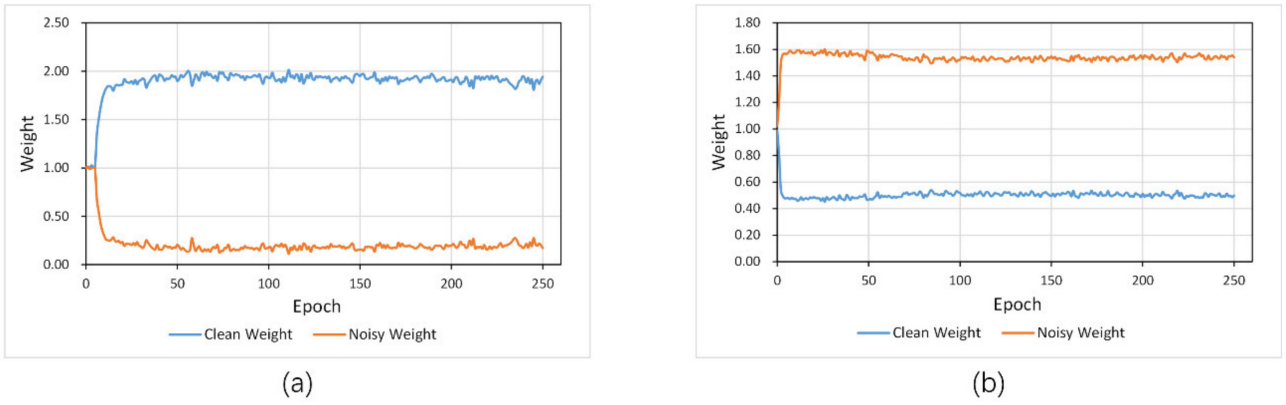

Figure 12. Average weights of clean-label samples and noisy-label samples during the training of the network at a noise rate of 50%. (**a**) Dilated noise, (**b**) eroded noise.

As we can see, for the dilated noise, the network cannot separate the clean samples from the noisy samples at the very beginning of training. Nevertheless, as the training batches increase, the clean samples' loss weights increase significantly and the loss weights of the noisy samples considerably decrease. This means that the attention subnetwork (λ) can distinguish the clean samples from the dilated noisy samples.

However, the opposite result is found when we examine the average weight for the eroded noise (Figure 12b). The attention subnetwork incorrectly considers the noisy samples as being more essential samples, which are thus assigned with higher weights.

The reason behind this might be more universal. When we give a network the power to choose which training samples are necessary to learn, it always tends to learn samples with low feature variation. From this perspective, in the experiment of adding dilated noise, the clean samples have a lower feature variation than the noisy samples, so the network gives them higher weights. The same phenomenon was found in the case of adding eroded noise. The noise has few pixels with a lower feature variation that are easy to learn. As a result, those samples are assigned with higher weights.

This local optimum problem is similar to how the human brain works. A child also intends to learn a new object with a low feature variation at the very beginning. The training samples with more complex features will more likely be ignored. The difference between humans and the deep-learning network is that we can adjust and update our loss function (if we have) after learning the same object under different scenarios. However,

for a deep-learning network, the loss function is permanently fixed for a specific task. Therefore, two possible pathways may help to overcome this problem: (1) enforcing the learning process, enlarging the training sample size, or applying modules from the enforced learning field; (2) designing a mechanism for updating the loss function.

5.2. Class-Balance Coefficient α Helps to Overcome the Local Optimum Problem

Because a low-feature variation sample always shows an imbalance between the background and the foreground classes, we added the class balance coefficient α to increase the loss percentage of the foreground class when calculating the loss to enforce network learning from the limited samples.

Figure 13 shows the accuracy of U-Net network (left column), U-Net with an attention subnetwork (λ) (middle column) and the WLN (right column) with different noise levels and noise rates with dilated (first row) and eroded (second row) noise during training. From the graph, we can see that when the noise rate is relatively low, all three methods can maintain high accuracy and the difference is not significant. Furthermore, when the noise rate is 50%, both the U-Net with λ and WLN methods can maintain high accuracy in the dilated-noise scenario. In contrast, the U-Net method's accuracy curve decreases significantly. In the eroded noise, because the attention subnetwork cannot assign correct weights to clean and noisy samples, the method of only adding λ will over-fit the noisy label samples, making the accuracy even lower than the original U-Net network. However, the WLN with the class balance coefficient (α) can solve this problem and maintain high accuracy. This result demonstrates that the class-balance coefficient (α) helps to overcome the local optimum problem.

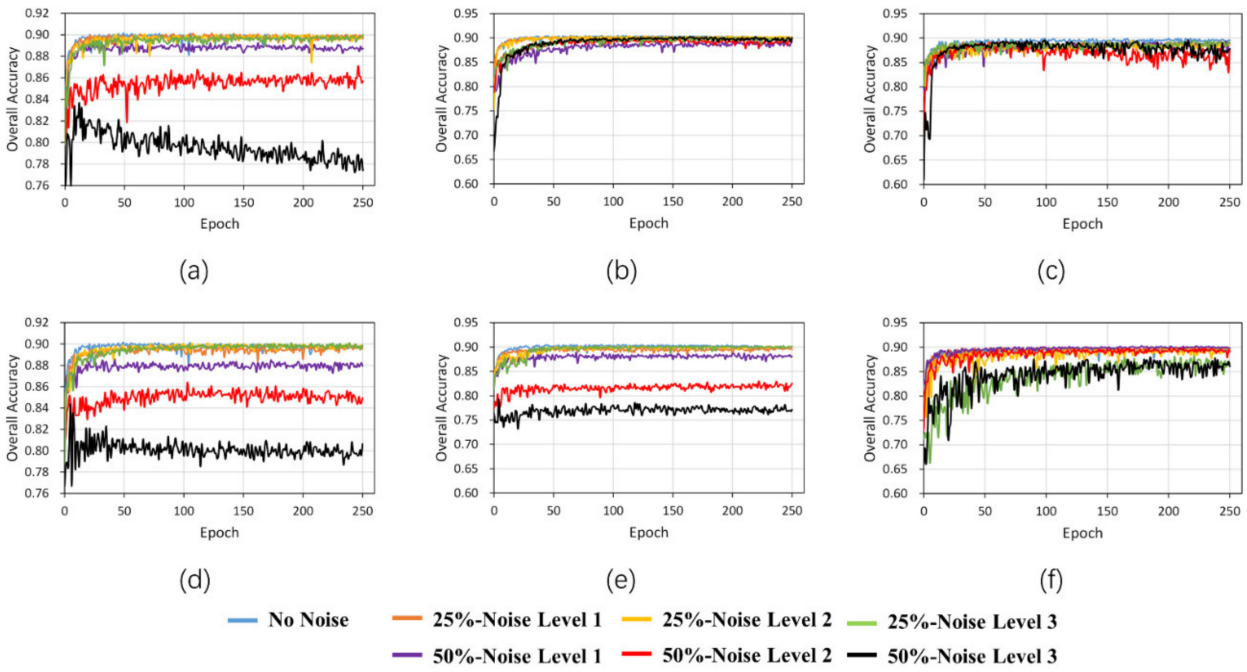

Figure 13. Accuracy of the U-Net network (left column (**a**,**f**)), U-Net with attention subnetwork (λ) (middle column (**b**,**e**)) and the WLN (right column (**c**,**f**)) with different noise levels and noise rates under dilated (first row (**a**–**c**)) and eroded (second row (**d**–**f**)) noise during training.

In this study, the class-balance coefficient (α) designed only addresses the class-imbalance problem for each training sample. It does not address the overall imbalance problem of the entire dataset. In some circumstances, the imbalance problem may still exist between the different classes and affect the network performance. This limitation can be improved in the future.

6. Conclusions

Errors in training labels are usually hard to identify and correct, especially in remote sensing datasets across different times and locations. In this paper, we propose a general antinoise network framework, the WLN, based on the idea of weighting the loss for each training sample to minimize the impact of erroneous samples for remote-sensing image classification. The framework consists of two networks: the segmentation subnetwork and attention subnetwork. The segmentation subnetwork classifies the image pixel per pixel and calculates the output results with the labels to obtain the initial loss during the training process. The attention subnetwork generates the weight loss of a batch sample and combines it with the class-balance coefficient to prevent a class imbalance for each training sample. These three parts combine to get the final loss and backpropagate the two subnetworks to update the network parameters.

Four types of label noise (an insufficient label, redundant label, missing label and incorrect label) were simulated by dilate and erode processing to test the network's antinoise ability. After comparing the performance of the proposed WLN to the original U-Net model when extracting buildings in the Inria Aerial Image Labeling Dataset, we found the following:

1. When the noise rate and noise level are low, the convolutional-neural-network is almost unaffected by the label noise, which may be due to the network's specific noise immunity. After the training set's label noise rate exceeds a certain threshold, the convolutional-neural-network's accuracy decreases significantly.
2. For the four-label noise, our proposed method of the WLN can maintain high accuracy and outperform the original method if the sample noise rate and noise level of the data set gradually increase.
3. The local optimum problem was found if we allowed the network to choose which samples were essential. This phenomenon might be universal and can be relieved by adjusting the class-label imbalance by adding the class-balance coefficient.

The antinoise framework proposed in this paper can help current segmentation models avoid noisy training labels. The local optimum problem that happens when we give a network the power to choose which training samples are necessary to learn should be carefully addressed in the future. This problem might be solved by giving a deep-learning network the ability to train more intelligently and efficiently with prior knowledge in the remote-sensing field, which is also worth investigating in the future.

Author Contributions: Conceptualization, S.G.; data curation, C.L.; investigation, L.S.; methodology, C.L. and S.G.; project administration, J.C.; resources, L.S.; software, Y.X.; supervision, S.G.; validation, X.Z.; visualization, L.S. and Y.Y.; writing—original draft, C.L.; writing—review and editing, S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Key Research and Development Program of China (Project No. 2017YFB0504203), the Natural Science Foundation of China project (41601212, 41801360, 41771403, and 41801358), and the Fundamental Research Foundation of Shenzhen Technology and Innovation Council (Project No. KCXFZ202002011006298).

Institutional Review Board Statement: Not applicable.

Acknowledgments: We thank all the GIS group members at the SIAT, Chinese Academy of Sciences, for their encouragement and discussion of the work presented here. We also thanks to the anonymous reviewers for their valuable suggestions on the earlier drafts of this study.

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

References

1. Jiang, J.; Ma, J.; Wang, Z.; Chen, C.; Liu, X. Hyperspectral image classification in the presence of noisy labels. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 851–865. [CrossRef]
2. Raykar, V.C.; Yu, S.; Zhao, L.H.; Valadez, G.H.; Florin, C.; Bogoni, L.; Moy, L. Learning from crowds. *J. Mach. Learn. Res.* **2010**, *11*, 1297–1322.
3. Huang, J.; Qu, L.; Jia, R.; Zhao, B. O2U-Net: A simple noisy label detection approach for deep neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3325–3333. [CrossRef]
4. Song, H.; Kim, M.; Park, D.; Lee, J.G. Two-phase learning for overcoming noisy labels. *arXiv* **2020**, arXiv:2012.04337.
5. Wu, P.; Zheng, S.; Goswami, M.; Metaxas, D.; Chen, C. A topological filter for learning with label noise. *arXiv* **2020**, arXiv:2012.04835.
6. Li, J.; Socher, R.; Hoi, S.C.H. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv* **2020**, arXiv:2002.07394.
7. Zhou, T.; Wang, S.; Bilmes, J. Robust Curriculum Learning: From clean label detection to noisy label self-correction. In Proceedings of the International Conference on Learning Representations, Lisbon, Portugal, 28–29 October 2021.
8. Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.W.; Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *arXiv* **2018**, arXiv:1804.06872.
9. Chen, X.; Gupta, A. Webly supervised learning of convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1431–1439. [CrossRef]
10. Sukhbaatar, S.; Bruna, J.; Paluri, M.; Bourdev, L.; Fergus, R. Training convolutional networks with noisy labels. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–11.
11. Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; Wang, X. Learning from massive noisy labeled data for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2691–2699. [CrossRef]
12. Yao, Y.; Liu, T.; Han, B.; Gong, M.; Deng, J.; Niu, G.; Sugiyama, M. Dual T: Reducing estimation error for transition matrix in label-noise learning. *arXiv* **2020**, arXiv:2006.07805.
13. Ghosh, A.; Kumar, H.; Sastry, P.S. Robust loss functions under label noise for deep neural networks. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 1919–1925.
14. Zhang, Z.; Sabuncu, M.R. Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv* **2018**, arXiv:1805.07836.
15. Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; Bailey, J. Symmetric cross entropy for robust learning with noisy labels. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 322–330. [CrossRef]
16. Liu, T.; Tao, D. Classification with Noisy Labels by Importance Reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 447–461. [CrossRef] [PubMed]
17. Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Xu, Z.; Zhou, S.; Meng, D. Meta-weight-net: Learning an explicit mapping for sample weighting. *arXiv* **2019**, arXiv:1902.07379.
18. Jenni, S.; Favaro, P. Deep bilevel learning. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 632–648. [CrossRef]
19. Ren, M.; Zeng, W.; Yang, B.; Urtasun, R. Learning to reweight examples for robust deep learning. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 10, pp. 6900–6909.
20. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Can semantic labeling methods generalize to any city? In The inria aerial image labeling benchmark. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 3226–3229. [CrossRef]
21. Frénay, B.; Verleysen, M. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2014**, *25*, 845–869. [CrossRef] [PubMed]
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 6000–6010. [CrossRef]
23. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 20 June 2019; pp. 3141–3149. [CrossRef]
24. 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.
25. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2020**, arXiv:1709.01507.