remote sensing

MDPI

Article

ERN: Edge Loss Reinforced Semantic Segmentation Network for Remote Sensing Images

Shuo Liu [1], Wenrui Ding [2], Chunhui Liu [2,*], Yu Liu [1], Yufeng Wang [1] and Hongguang Li [2]

[1] School of Electronic and Information Engineering, Beihang University, Beijing 100191, China; liush@buaa.edu.cn (S.L.); liu_yu@buaa.edu.cn (Y.L.); wyfeng@buaa.edu.cn (Y.W.)

[2] Unmanned Systems Research Institute, Beihang University, Beijing 100191, China; ding@buaa.edu.cn (W.D.); lihongguang@buaa.edu.cn (H.L.)

* Correspondence: liuchunhui2134@126.com

Received: 7 July 2018; Accepted: 14 August 2018; Published: 22 August 2018

Abstract: The semantic segmentation of remote sensing images faces two major challenges: high inter-class similarity and interference from ubiquitous shadows. In order to address these issues, we develop a novel edge loss reinforced semantic segmentation network (ERN) that leverages the spatial boundary context to reduce the semantic ambiguity. The main contributions of this paper are as follows: (1) we propose a novel end-to-end semantic segmentation network for remote sensing, which involves multiple weighted edge supervisions to retain spatial boundary information; (2) the main representations of the network are shared between the edge loss reinforced structures and semantic segmentation, which means that the ERN simultaneously achieves semantic segmentation and edge detection without significantly increasing the model complexity; and (3) we explore and discuss different ERN schemes to guide the design of future networks. Extensive experimental results on two remote sensing datasets demonstrate the effectiveness of our approach both in quantitative and qualitative evaluation. Specifically, the semantic segmentation performance in shadow-affected regions is significantly improved.

Keywords: CNN; deep learning; edge loss reinforced network; remote sensing; semantic segmentation

1. Introduction

With the rapid development of remote sensing, it has become much easier to obtain high-resolution images [1–3]. The ever-increasing amount of data places more emphasis on automated interpretation of remote sensing images [2,4]. As an important step towards scene understanding [5], segmentation plays a vital role in many important remote sensing applications [6], such as natural hazards detection [7], urban planning [8,9], land cover mapping [10] and so on. Unlike the classical paradigm in geographic object-based image analysis that unsupervised segmentation is followed by classification [11–14], semantic segmentation employs a pixel-level supervised style and assigns each pixel with a pre-designed label.

Recently, deep convolutional neural network (CNN)-based semantic segmentation has drawn a great deal of attention due to excellent performances [15,16]. In particular, the encoder–decoder architecture has been proven highly effective in generating pixel-wise predictions in an end-to-end style [17–19]. Despite this success, some detail is lost after down-sampling during the encoder forward stage, which means that predictions tend to be less accurate near boundaries [20,21]. A typical idea is to add skip connections that assemble high-resolution feature maps from the encoder to learn a more precise output [22]. For remote sensing images, Volpi et al. [23] proposed a full patch labeling (FPL) network to up-sample the rough spatial maps with successive deconvolution layers. Liu et al. [24]

proposed an hourglass-shaped network (HSNet) and replaced some convolution layers with inception and residual blocks to further enhance the ability of context information extraction.

Even though the above works have achieved remarkable progress, semantic segmentation for remote sensing images is far from being solved and remains a challenging task due to the complex surface environments. The inter-class variance of different surface types in remote sensing images is extremely low, which makes accurate labeling near boundaries difficult. For example, the buildings and surfaces in Figure 1a present a very similar visual appearance that confuses the network. In addition, the ubiquitous shadows in remote sensing images further decrease the inter-class variance, resulting in a large amount of semantic ambiguity and intensifying the challenge of semantic segmentation. Figure 1b shows the incorrect labeling under the interference of shadows.

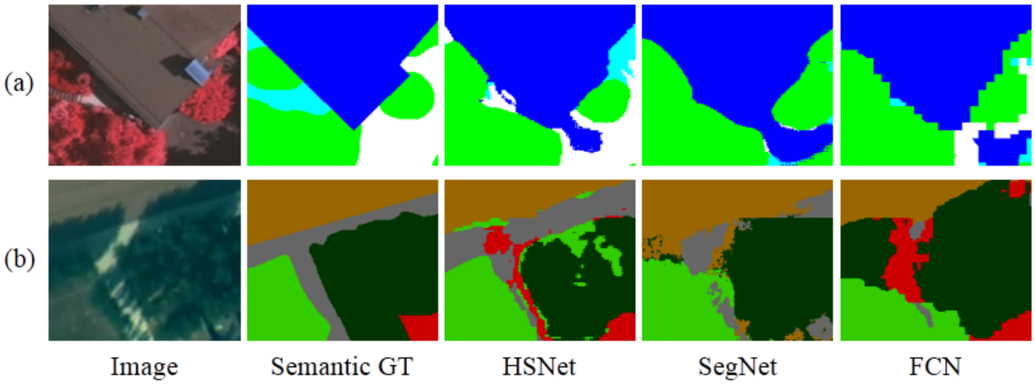


Figure 1. Examples of remote sensing images that are challenging for semantic segmentation. (**a**) similar appearance between a building and its surroundings, in which the impervious surface was incorrectly recognized as a building by HSNet [24], SegNet [19], and FCN [17]. white: impervious surface; blue: buildings; cyan: low vegetation; green: trees. (**b**) interference of shadows, which results in poor performance; red: buildings; gray: roads; bright green: grass; dark green: trees; brown: land.

Typical semantic segmentation approaches mainly focus on mitigating semantic ambiguity via providing rich information [19,24]. However, redundant and noisy semantic information from high-resolution feature maps may clutter the final pixel-wise predictions [25]. Consideration of which kind of information can directly help the network to better distinguish different semantics is needed. The boundary information is simple but effective to indicate the semantic separation between different regions. In fact, the traditional high-order conditional random field (CRF) based semantic segmentation methods utilize superpixels to retain boundary information [26,27]. There are also some superpixel-based CNN models for semantic segmentation [15,28]. Their main shortcoming is that the superpixel is unlearnable and not robust. Thanks to the holistically-nested edge detection (HED) network [29], deep net-style edges have shown the capacity to improve the performance of high-level semantic tasks [30,31]. Chen et al. [32] proposed an edge-preserving filtering method using domain transform to enhance object localization accuracy in semantic segmentation. Cheng et al. [33] fused semantic segmentation net and edge net with a regularization method to refine entire network. Marmanis et al. [34] proposed a model that cascades the edge net (HED [29]) and semantic segmentation net (FCN [17]/SegNet [19]), where the model is complex and the training phase must be carefully fine-tuned. In contrast to these works, we sought to establish a simple and scalable model that integrates multiple weighted edge structures into semantic segmentation.

To this end, this paper proposes a novel edge loss reinforced semantic segmentation network (ERN). The framework follows the encoder–decoder architecture shown in Figure 2. The edge loss reinforced structures are constructed at the encoder and decoder parts, which consist of convolution layers and an edge ground truth supervision (only in the training phase). ERN leverages the edge loss reinforced structures to focus on the low-level boundary features and further reduce semantic ambiguity. It should be noted that the edge ground truth is obtained by a simple calculation of the semantic ground truth gradient, which does not require extra manual labeling effort.

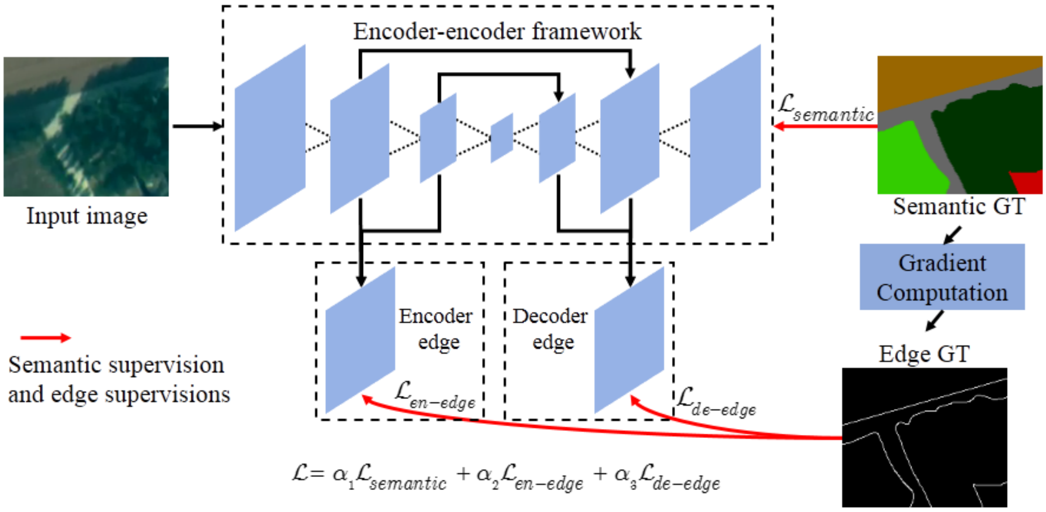

Encoder-encoder framework
Input image
$\mathcal{L}_{semantic}$
Semantic GT
Gradient Computation
Edge GT
Encoder edge
Decoder edge
$\mathcal{L}_{en-edge}$
$\mathcal{L}_{de-edge}$
Semantic supervision and edge supervisions
$\mathcal{L} = \alpha_1 \mathcal{L}_{semantic} + \alpha_2 \mathcal{L}_{en-edge} + \alpha_3 \mathcal{L}_{de-edge}$

Figure 2. Framework of the proposed edge loss reinforced semantic segmentation network (ERN). The encoder edge and decoder edge are constructed to promote the semantic segmentation. In the training phase, the semantic ground truth (GT) and edge ground truth simultaneously provide the supervision information in different layers. Thus, the loss ($\mathcal{L}$) of the network includes semantic loss ($\mathcal{L}_{semantic}$) and edge loss ($\mathcal{L}_{en-edge}$, $\mathcal{L}_{de-edge}$).

The main contributions of our paper are as follows: (1) we propose a novel edge loss reinforced semantic segmentation network for remote sensing. By introducing multiple weighted edge supervisions, the network can better preserve the spatial boundary information and significantly improve semantic segmentation performance; (2) the main representations of the network are shared between the edge loss reinforced structures and semantic segmentation, which means that the ERN simultaneously achieves semantic segmentation and edge detection without significantly increasing the model complexity; and (3) different ERN schemes are explored and discussed to provide guidelines for applying edge loss reinforced structures to future networks.

We evaluate the performance of ERN on two remote sensing datasets: (1) UAV (unmanned aerial vehicle) Image Dataset, which is collected by a medium-altitude UAV; and the (2) ISPRS Vaihingen Dataset [35], which is a publicly available dataset for 2D semantic labeling. Experimental results show that our ERN achieved a performance that excelled the referenced methods and was significantly improved for regions near the boundary or in shadow. The remainder of the paper is organized as follows. In Section 2, we review recent related literature. We describe the ERN architecture in Section 3. In Section 4, we design the experiments and compare the performance of ERN with several CNN-based baselines. Section 5 discusses experimental results and the limitation of ERN, as well as future research directions. In addition, Section 6 gives a brief statement of our work.

2. Background

Semantic segmentation is one of the key problems in the field of computer vision [36]. In contrast to classification and recognition tasks [37–39], which outputs one label for the whole image, semantic segmentation assigns a pre-designed label to each pixel in an image and is thus also called pixel-wise labeling. We found that different pixel-wise labeling tasks share similar principles and frameworks. We first introduce the semantic segmentation models and then briefly present some related pixel-wise labeling methods.

Semantic segmentation models. Before the widespread application of CNN-based architectures, one kind of successful traditional method formulated semantic segmentation as a CRF-based energy minimization problem [26,27]. The expression for energy consists of a unary energy term, a pairwise energy term, and a superpixel-based high-order energy term. The superpixels are usually obtained by bottom-up image segmentation methods [40–42], which include a large amount of image boundary information. Our work is partly inspired by the idea that retaining the boundary information can help to establish spatial context constraints. Since they were limited by the ability of hand-crafted features [43], traditional CRF based methods were gradually surpassed by CNN-based architectures.

A fully convolutional network [17] was proposed where the fully connected layers of classification models [37,44] were replaced with deconvolution layers to produce dense pixel-wise predictions, demonstrating how CNNs can be trained end-to-end for semantic segmentation. However, deconvolutional layers produce coarse segmentation maps because of a loss of information during pooling. Two different classes of architectures have evolved in the literature to tackle this issue. The first strengthens the power of the decoder part, in which skip connections from the encoder to the decoder, along with gradual deconvolution, help to more effectively recover details [19,22,24]. Apart from the above architecture, another insightful work came from dilated/à-trous convolutions [45–48], which support exponentially expanding receptive fields without losing resolution. In addition, CRF has been applied to refine the semantic segmentation results [46,47,49,50].

Related models. From a broader perspective, edge detection, salient object detection, and object symmetry detection can all be regarded as pixel labeling problems. The difference between them is that their label value spaces are different, including “edge” or “non-edge”, “object” or “non-object”.

The HED [29] comprises a single-stream deep network with multiple side outputs. Each side-output layer is also associated with a classifier to perform deep layer supervision to "guide" early classification results, as in [51]. Thus, the loss of HED consists of side output loss and fuse loss. Hou et al. [52] further propose a deeply supervised salient object detection method (DSS) by introducing short connections to the skip-layer structures within the HED architecture. Ke et al. [53] propose a side-output residual network (SRN) for symmetry detection based on HED architecture. SRN leverages output residual units (RUs) to fit the errors between side outputs and ground truth. The short connection in DSS and residual units in SRN is extremely similar.

The above literature shows a strong correlation between different pixel labeling tasks. The network for one task can be applied to other tasks with slight or even no modification. Therefore, we were inspired to combine semantic segmentation with edge detection in a single network in which the edge outputs are deeply supervised within additional edge loss reinforced structures.

3. Proposed Edge Loss Reinforced Semantic Segmentation Network

3.1. Model Overview

Our proposed ERN model is illustrated in Figures 2 and 3. ERN consists of an encoder–decoder semantic segmentation net with two additional edge loss reinforced structures constructed from encoder and decoder parts, respectively. The corresponding semantic loss and edge loss are jointly trained end-to-end in order to optimize the process.

The first component is an encoder–decoder semantic segmentation net based on the HSNet [24]. The encoder part includes convolution layers, inception blocks [54], and max pooling layers. The spatial resolution of feature maps gradually decreases after pooling. The decoder part mainly includes deconvolution layers and inception blocks. The deconvolution is used to progressively up-sample the feature maps to the original spatial resolution of the input images. The skip connections from the encoder to the decoder use residual blocks [44].

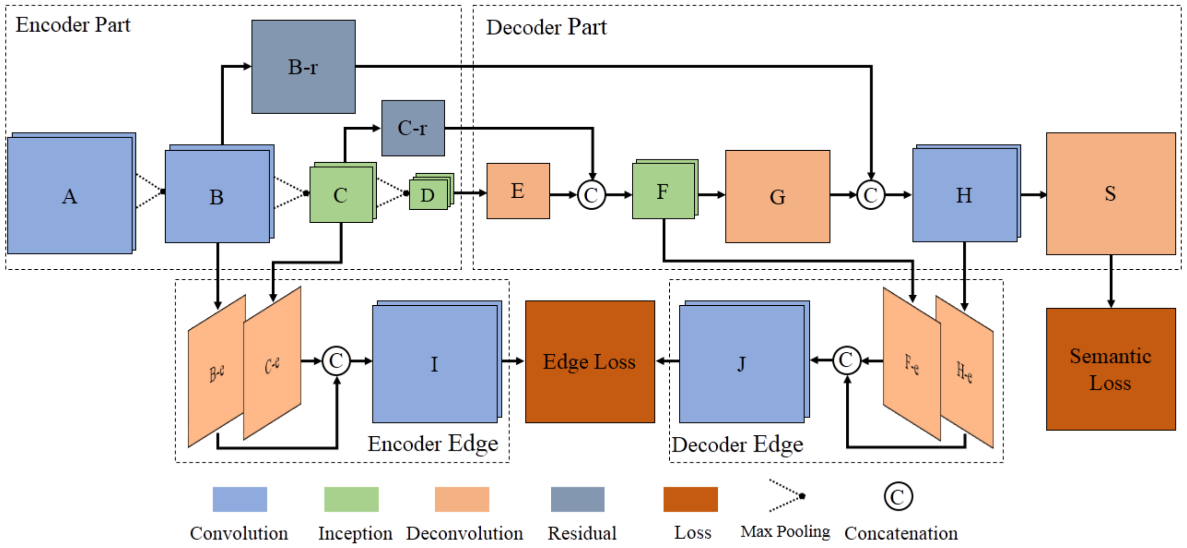


Figure 3. Architecture of the proposed ERN. A, B, H, I, and J are convolutional layers; C, D, and F are inception blocks; B-r and C-r are residual blocks; E, G, S, B-e, C-e, F-e, and H-e are deconvolutional layers.

The second component is the edge loss reinforced structures. Feature maps with different spatial resolution (e.g., after convolution layer H and inception block F) are directly deconvoluted to the original input size and then concatenated to further produce edge predictions. The edge loss reinforced structure is supervised by the edge ground truth. The configurations of ERN are listed in Table 1, which lists the kernel size, output number, and spatial resolution of each layer. The configurations of inception blocks (C, D, and F) and residual blocks (B-r and C-r) are discussed in Section 3.3.

Table 1. Configurations of the ERN.

	Layer ID	Type	Filter Size	Spatial Resolution
Encoder	A	convolution × 2	3 × 3, 64	256 × 256
	B	convolution × 2	3 × 3, 128	128 × 128
	B-r	residual block	-, 128	128 × 128
	C	inception block× 2	-, 256	64× 64
	C-r	residual block	-, 256	64 ×64
	D	inception block× 3	-, 512	32× 32
Decoder	E	deconvolution	-,256	64 × 64
	F	inception block × 2	-,256	64 × 64
	G	deconvolution	-, 128	128 × 128
	H	convolution × 2	3 × 3, 128	128 × 128
	SL	deconvolution	-, 6	256 × 256
Edge Loss	B-e	deconvolution	-,2	256 × 256
	C-e	deconvolution	-,2	
	I	convolution × 2	3 × 3, 64	
	EEL	convolution	3 × 3, 2	
	F-e	deconvolution	-,2	
	H-e	deconvolution	-,2	
	J	convolution × 2	3 × 3, 64	
	DEL	convolution	3 × 3, 2	

3.2. The Edge Loss Reinforced Structure Based on Short Connection

The edge loss reinforced structure in ERN was inspired by HED [29] and DSS [52]. Different architectures are illustrated in Figure 4. Figure 4a shows a simplified version of HED, showing the proposed scheme with deep supervision for each side output and fuse output. Thus, a series of side losses are added after each side output to preserve more detailed edge information. DSS further connects feature maps at different scales before output, as shown in Figure 4b.

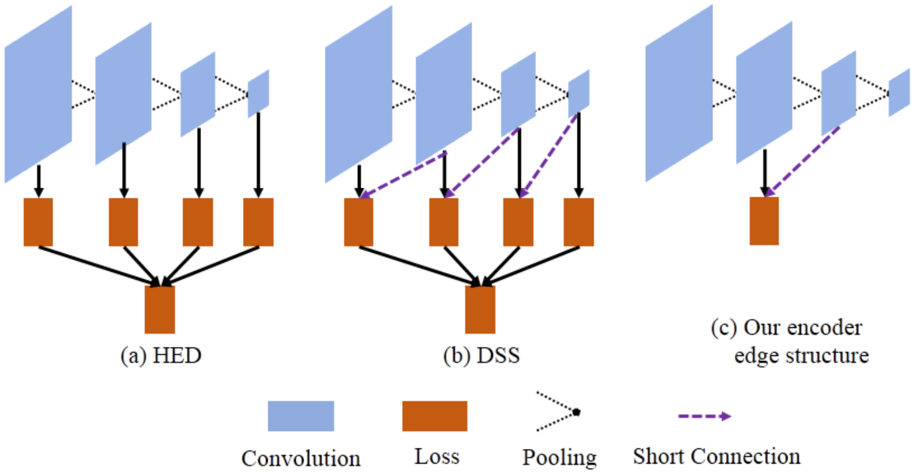


Figure 4. Illustration of different architectures. HED [29] and DSS [52] have several side output losses and one fuse loss. Our edge structure is much simpler.

HED [29] and DSS [52] were designed to accomplish one specific task. However, the edge loss reinforced structure within ERN is auxiliary to the semantic segmentation. It was designed to be as simple as possible, to balance the trade-off between edge detection performance and model complexity. Therefore, we simplify the edge loss reinforced structure so that the entire model is not too complicated. Moreover, the lower-resolution feature map has poorer spatial accuracy because of pooling operations, as discussed in Section 1. For these reasons, we did not design side output loss in the same way as HED [29] and DSS [52].

Finally, our edge loss reinforced structure simply concatenates two middle-scale feature maps without side output, as shown in Figure 4c. The decoder edge loss reinforced structure is symmetrical to the encoder one. See Figure 3 and Table 1 for details. The edge dependent loss in ERN comprises two terms: encoder edge loss and decoder edge loss. The encoder one plays the role of deep supervision like the shallow side output loss in HED [29] and DSS [52].

3.3. Inception and Residual Learning

The inception block is introduced to replace the convolutional layers (e.g., layers C, D, and F in Figure 3). The structure of the inception block is shown in Figure 5a, and the corresponding configurations are listed in Table 2. The inception block is composed of four branches. Three branches comprise two banks of convolution filters: the convolution size of the first is 1 × 1, and the second is 3 × 3, 5 × 5, and 7 × 7. The last branch comprises one bank of convolution filters with size 1 × 1. Each convolution is followed by batch normalization and a rectified linear unit (ReLu). Convolution filters of different sizes are assembled in one inception block to enable multi-scale inference through the network.

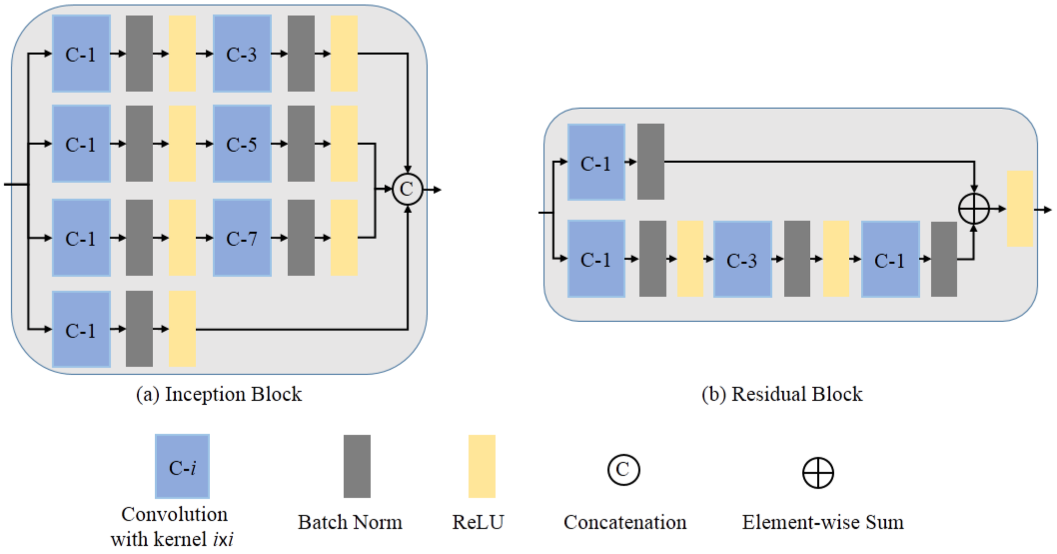


Figure 5. Structure of the inception block and the residual block.

Table 2. Configurations of the inception blocks.

Layer ID	Convolution Configurations	Operation	Output Number
C	$1 \times 1, 128$ $3 \times 3, 128$ $1 \times 1, 32$ $5 \times 5, 32$ $1 \times 1, 32$ $7 \times 7, 32$ $1 \times 1, 64$	concatenation	256
D	$1 \times 1, 192$ $3 \times 3, 256$ $1 \times 1, 64$ $5 \times 5, 128$ $1 \times 1, 32$ $7 \times 7, 64$ $1 \times 1, 64$	concatenation	512
F	$1 \times 1, 256$ $3 \times 3, 128$ $1 \times 1, 64$ $5 \times 5, 32$ $1 \times 1, 32$ $7 \times 7, 32$ $1 \times 1, 64$	concatenation	256

The residual block in ERN is shown in Figure 5b, and the corresponding configurations are listed in Table 3. The residual block is composed of two branches. Branch one is a bank of convolution filters with size 1×1, followed by batch normalization. Another branch consists of three banks of convolution filters with size 1×1, 3×3, and 1×1. We use batch normalization after every convolution. An element-wise summation of two branches is carried out before the ReLu of the final convolution.

Table 3. Configurations of the residual blocks.

Layer ID	Convolution Configurations	Operation	Output Number
B-r	$1 \times 1, 128$ $1 \times 1, 64$ $3 \times 3, 128$ $1 \times 1, 128$	element-wise sum	128
C-r	$1 \times 1, 256$ $1 \times 1, 64$ $3 \times 3, 128$ $1 \times 1, 256$	element-wise sum	256

3.4. Joint Semantic Loss and Edge Loss

We denote the input training data set with N samples $S = \{(X_n, Y_n, YE_n)\}_{n=1}^{N}$, where $X_n = \{x_j^{(n)}, j = 1, ..., T\}$ denotes the raw input image with T pixels. $Y_n = \{y_j^{(n)}, j = 1, ..., T\}$ and $YE_n = \{ye_j^{(n)}, j = 1, ..., T\}$ denote the corresponding semantic ground truth and edge ground truth, respectively, for image X_n. $y_j^{(n)} = c$ denotes that the pixel belongs to the c^{th} semantic label where $c \in \{1, ..., C\}$. $ye_j^{(n)} \in \{1, 0\}$ denotes whether or not the pixel lies on an edge. It is worth mentioning that the edge ground truth is obtained by calculating the gradient of the semantic ground truth, $YE_n = \nabla Y_n$.

For simplicity, we represent the collection of all standard network layer parameters by $\mathbf{W}$. The semantic segmentation output and each edge output are associated with a classifier, in which the corresponding weights are denoted w_s, w_{encode}, and w_{decode}.

The cross-entropy loss function summed over all pixels is used. However, when applied to semantic segmentation of remote sensing images [55] and edge detection tasks [29,56], the ordinary cross-entropy loss can be heavily affected by the imbalance of the class distribution . We adopted a weighted loss function where the calculation of trade-off weight for biased sampling is based on median frequency balancing [16].

The semantic loss is defined as:

$$\mathcal{L}_{semantic}(\mathbf{W}, w_s) = -\sum_{j,n,c} \beta_c^{(semantic)} \cdot \log Pr(y_j^n = c|X; \mathbf{W}, w_s), \tag{1}$$

where $\beta_c^{(semantic)} = \frac{frequency(c)}{\sum_c frequency(c)}$ denotes the weight of class c. Pr denotes probability.

The encoder edge loss is defined as:

$$\begin{aligned}\mathcal{L}_{en-edge}(\mathbf{W}, w_{encode}) = &-\beta^{(edge)} \cdot \sum_{j,n} \log Pr(ye_j^n = 1|X; \mathbf{W}, w_{encode}) \\ &- (1 - \beta^{(edge)}) \cdot \sum_{j,n} \log Pr(ye_j^n = 0|X; \mathbf{W}, w_{encode}),\end{aligned} \tag{2}$$

where $\beta^{(edge)} = \frac{frequency(edge)}{frequency(edge)+frequency(non-edge)}$ indicates the weight of edge pixels and $(1 - \beta^{(edge)})$ denotes the weight of non-edge pixels. The decoder edge loss $\mathcal{L}_{de-edge}$ is identical to the encoder edge loss.

Putting the semantic loss and edge loss together, we minimize the following objective function via back-propagation:

$$(\mathbf{W}, w)^* = argmin(\alpha_1 \cdot \mathcal{L}_{semantic}(\mathbf{W}, w_s) + \alpha_2 \cdot \mathcal{L}_{en-edge}(\mathbf{W}, w_{encode}) + \alpha_3 \cdot \mathcal{L}_{de-edge}(\mathbf{W}, w_{decode})), \tag{3}$$

where α_1, α_2, and α_3 are continuous hyper-parameters and denote the weights of semantic loss, encoder edge loss, and decoder edge loss, respectively. In our experiments, the α_1 was fixed to 1, $\alpha_2 = \alpha_3 = 20$ for the UAV Image Dataset, and $\alpha_2 = \alpha_3 = 4$ for the ISPRS Vaihingen Dataset. More discussions about the value of α_1, α_2, and α_3 can be found in Section 5.1.

4. Experimental Design and Results

We performed extensive experiments to evaluate the effectiveness of the proposed ERN architecture. In this section, we describe the datasets used and our experimental settings, and report quantitative and qualitative results. The full implementation and trained networks are publicly available at: https://github.com/liushuo2018/ERN.

4.1. Datasets

We evaluated the proposed ERN on two datasets for semantic segmentation.

UAV Image Dataset. This dataset consists of 200 images which were obtained by a medium-altitude UAV in a plain region located in east China, where the main landforms include cities, villages, and open fields. The images were acquired by a visible light camera and composed of three channels: red (R), green (G) and blue (B). Each of the images has 1280 × 1024 pixels at a GSD (Ground Sample Distance) ranges from 35 cm to 60 cm. All of the image pixels were labeled as one of the following six classes: building, road, grassland, tree, land, and clutter. Half of the UAV images were randomly selected as the training sets, and the others were reserved for testing. Further information of this dataset will be updated in the project page.

ISPRS Vaihingen Dataset [35]. This dataset is publicly avaliable and consists of 33 very high resolution true orthophoto (TOP) tiles, as well as corresponding DSM (digital surface model) and nDSM (normalized digital surface model) data. TOP images were acquired by an airbone color-infrared camera and composed of three channels: near-infrared (NIR), red (R) and green (G). In addition, the corresponding DSM data was acquired by LiDAR and composed of one channel. Each of the tiles have $\approx$2500 × 2000 pixels at a GSD $\approx$ 9 cm. Pixels were labeled as one of the following six classes: impervious surfaces, building, low vegetation, tree, car, and clutter/background. Following the HSNet [24], eleven tiles (areas 1, 3, 5, 7, 13, 17, 21, 23, 26, 32, 37) were selected for training, while the other five tiles (areas: 11, 15, 28, 30, 34) were reserved for testing.

4.2. Training and Testing

Training. In the training phase, data augmentation was employed to mitigate overfitting. The images were split into fixed size patches (256 × 256) with 50% overlap. Each image patch was rotated at a 90 degree interval and flipped vertically and horizontally to produce eight augmented patches.

The adaptive moment estimation (ADAM) [57] optimization algorithm was used to train the networks. ADAM is a variant of stochastic gradient descent (SGD) with two moments $m_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2$ and $v_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$ (g_t is the pre-set first momentum). The update rule in ADAM is:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t, \tag{4}$$

where $\hat{m}_t = \frac{m_t}{1-\beta_1}$ and $\hat{v}_t = \frac{v_t}{1-\beta_2}$ are the bias-corrected moments, and η is the learning rate. $(\beta_1, \beta_2, \epsilon)$ are parameters. In this paper, we chose $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. The learning rate was set to be divided by a factor of 10 every 10 epochs from an initial value of 10^{-5}. All the trainable parameters in the kernel of convolution and deconvolution layers were initialised following [58].

The training processes were performed on a Linux PC machine equipped with an single Nvidia GeForce 1080Ti graphics card. We implemented our deep network under Caffe [59] framework, and pre-processed original images with Python.

Testing. Limited by the GPU memory, the test images were also first split into small-size patches (256 × 256) to perform network inference. Then, the semantic segmentation result of the whole image was obtained by stitching the corresponding patch results. Overlap inference (OI) is widely used to mitigate erroneous artifacts caused by split and stitching pattern. We also performed multi-hypothesis prediction where the class for each pixel was identified in several overlapping patches to further improve the segmentation performance of whole image. In our overlap inference experiments, we classified overlapping patches with a stride of 128 pixels and then summed the results.

The testing processes were performed under the same environment with the training processes. In addition, the post-processing (stitching, multi-hypothesis prediction, etc) was also implemented with Python.

4.3. Evaluation Metrics

We used the same evaluation metrics as in HSNet [24] and evaluated the performance of different methods based on three criteria: per-class F-score, overall accuracy, and average F-score. The F-score is defined as:

$$\text{F-score} = 2 \times \frac{precision \times recall}{(precision + recall)}. \tag{5}$$

The overall accuracy is the total number of correctly-labeled pixels divided by the total number of pixels.

In the UAV Image Dataset and ISPRS Vaihingen Dataset [35], the clutter class accounts for an extremely small number of pixels. As a result, we neglected the clutter class when reporting the results, following the common practice [23,24].

4.4. Results

We compare our results with those of FCN [17], SegNet [19], and HSNet [24]. To produce the results of the above baselines, we employed the publicly available networks provided by the original authors and trained them under the same settings in Section 4.2.

4.4.1. Results of UAV Image Dataset

Numerical results. Table 4 reports the experimental results obtained from the UAV images. The results are organized into two groups, corresponding to the normal inference results and overlap inference results. From the table, it can be observed that the proposed ERN outperformed the other networks. The average F-score and the overall accuracy of ERN reached 87.74% and 91.90%, respectively, and ERN showed a better performance for all classes. Overlap inference (OI) systematically improved the prediction accuracy for all methods and all classes. The average F-score and overall accuracy of ERN further increased to 88.81% and 92.66%, respectively. This proves the effectiveness of overlap inference. The confusion matrices are further provided in Appendix A—Table A1.

Table 4. Experimental results on the UAV image dataset. OI: overlap inference.

Methods	Buildings	Road	Grass	Tree	Land	Average F-Score	Overall Accuracy
FCN [17]	91.88	84.06	81.12	60.96	94.59	82.52	87.09
SegNet [19]	91.45	73.85	85.52	67.85	93.48	82.43	87.98
HSNet [24]	92.78	81.92	85.35	63.45	95.45	83.79	89.42
ERN	**94.43**	**85.27**	**90.17**	**72.43**	**96.38**	**87.74**	**91.90**
FCN [17] + OI	92.27	84.96	81.58	61.39	94.87	83.01	87.50
SegNet [19] + OI	92.33	76.97	86.32	68.77	94.12	83.70	88.91
HSNet [24] + OI	93.47	84.04	86.21	65.26	95.88	84.97	90.25
ERN + OI	**95.02**	**87.20**	**91.17**	**73.88**	**96.76**	**88.81**	**92.66**

Qualitative Results. For a visual demonstration, Figure 6 shows the semantic segmentation results for four complete UAV images, while Figure 7 magnifies certain areas, showing more detail.

From Figure 6, it can be observed that most of the methods performed well for buildings and land. The performance on the road, tree, and grass was relatively poor, especially at the boundaries between different regions. ERN had much cleaner boundary results than other methods. These results were consistent with the numerical results in Table 4.

Figure 7 presents some local results in detail, which better shows the strength of ERN. From Figure 7a, it can be seen that the shadows from trees posed great difficulties for semantic segmentation. Both FCN and SegNet labeled part of the road as a tree. HSNet managed to detect the road, but the segmentation accuracy was quite low. ERN successfully recognized the road, tree, grass, and land, even under shadows. In Figure 7b–e, the results of ERN outperformed the other models, giving much more accurate boundaries. We argue that the edge loss reinforced structures contributed to the good performance by improving the boundary accuracy and reducing semantic ambiguity between different regions.

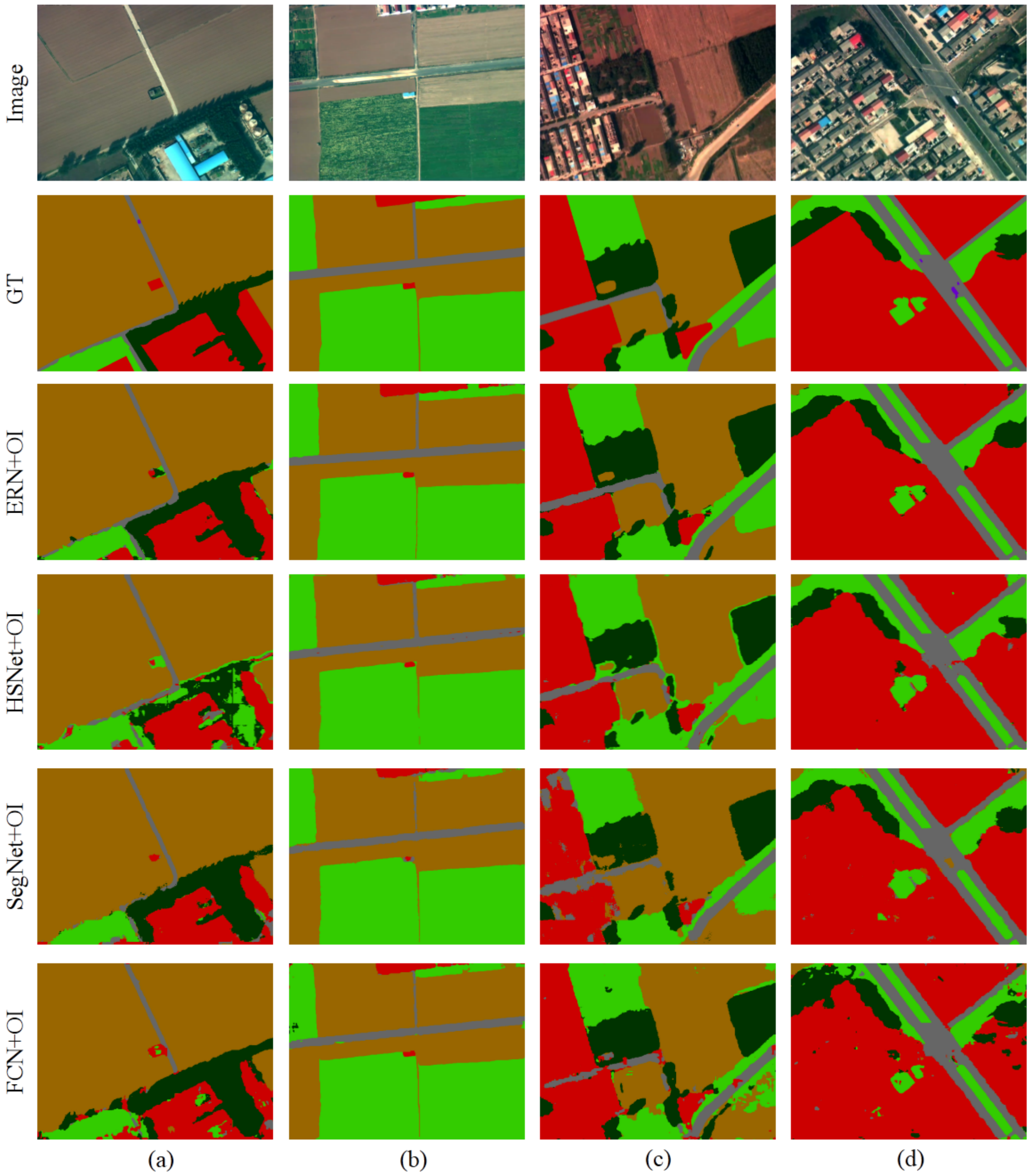


Figure 6. Semantic segmentation results of the UAV images from ERN, HSNet [24], SegNet [19] and FCN [17]. GT: ground truth. OI: overlap inference. (**a**–**d**) are four different scenes of UAV images, ranging from villages to cities; red: buildings; gray: roads; bright green: grass; dark green: trees; brown: land; purple: clutter.

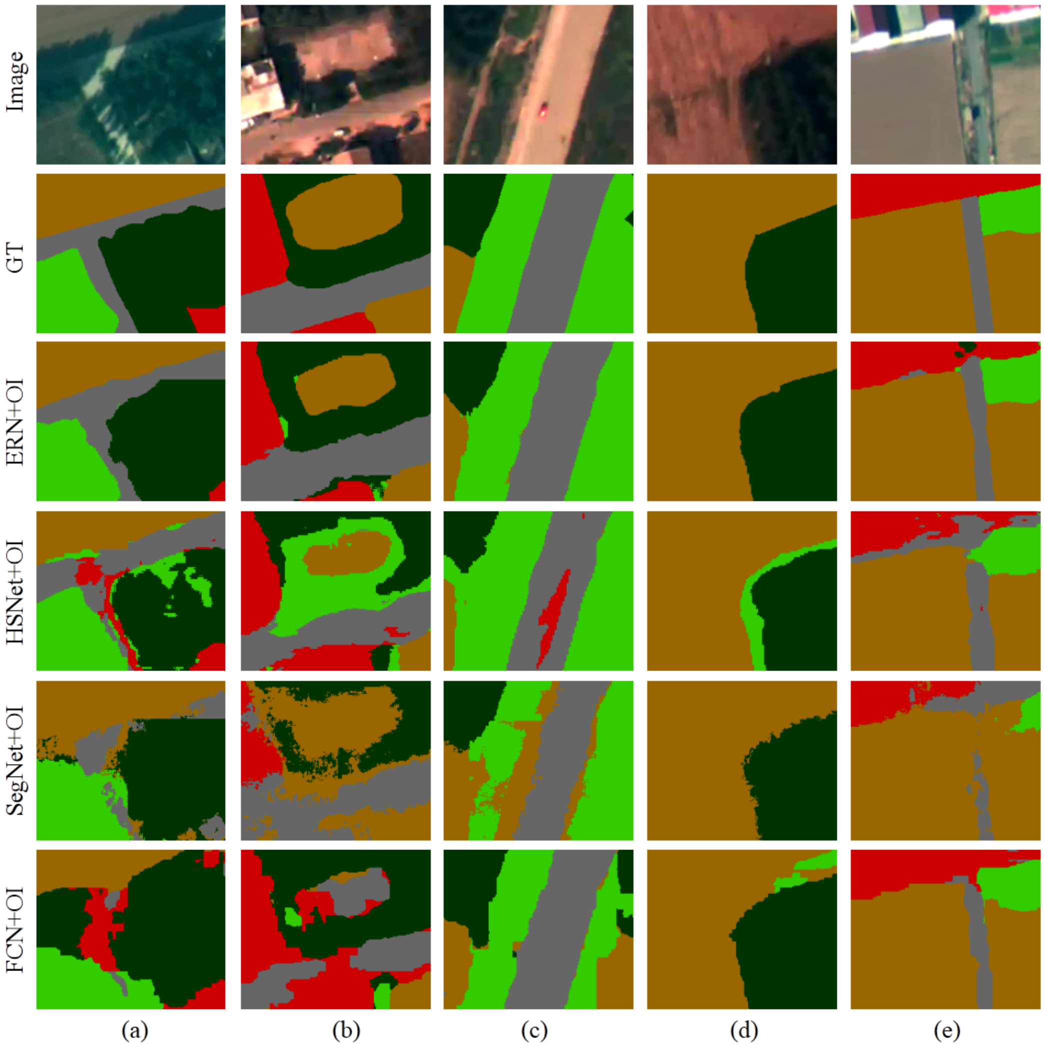


Figure 7. Semantic segmentation results for some local details of the UAV images. (**a**) shadow-affected region; (**b**) similar appearance between road and land; (**c**–**e**) challenge cases near the boundary; red: buildings; gray: roads; bright green: grass; dark green: trees; brown: land; purple: clutter.

4.4.2. Results of ISPRS Vaihingen Dataset

Numerical results. Aside from the conventional pixel-wise ground truth, border-eroded ground-truth label images are also available for the ISPRS Vaihingen Dataset. In these images, borders between classes are eroded with a disk radius of three pixels. We report results for both ground-truth versions. All pixels were considered for the conventional pixel-wise ground-truth version, while for the eroded version, border pixels were not accounted for.

Table 5 reports the results of different methods for the ISPRS Vaihingen Dataset. The F-scores for each class and overall performance are shown respectively for GT and er-GT (eroded ground-truth version). From the table, it can be observed that the proposed ERN outperformed the other networks. The average F-score and the overall accuracy of ERN reached 88.64% and 88.88%, respectively, and ERN reached a better performance for all classes. ERN improved the segmentation accuracy particularly well for the car class. It is worth mentioning that the experiments in Table 5 did not utilize overlap inference. The confusion matrices are further provided in Appendix A—Table A2.

Table 5. Experimental results on the ISPRS Vaihingen Dataset. er-GT: eroded ground-truth; Imp.Surf: impervious surface; LowVeg: low vegetation.

	Methods	Imp.Surf	Buildings	LowVeg	Tree	Car	Average F-Score	Overall Accuracy
er-GT	FCN [17]	89.41	93.80	76.46	86.63	71.32	83.52	86.75
	SegNet [19]	90.15	94.11	77.35	87.40	77.31	85.27	87.59
	HSNet [24]	90.89	94.51	78.83	87.84	81.87	86.79	88.32
	ERN	**91.48**	**95.11**	**79.42**	**88.18**	**89.00**	**88.64**	**88.88**
GT	FCN [17]	85.82	91.27	72.39	83.30	63.10	79.18	83.18
	SegNet [19]	86.68	91.74	73.22	83.99	71.36	81.40	84.07
	HSNet [24]	87.57	92.20	75.03	84.44	75.16	82.88	84.92
	ERN	**88.34**	**93.03**	**75.66**	**84.78**	**82.15**	**84.79**	**85.61**

Qualitative Results. As a visual demonstration, Figure 8 shows the semantic segmentation results for the ISPRS Vaihingen Dataset, while Figure 9 shows certain areas in more detail. The ISPRS Vaihingen dataset provides a nDSM, which can help the network to distinguish buildings and surfaces, trees and vegetation. We connected the nDSM with near-infrared (NIR), red (R), and green (G) as an additional channel for all methods.

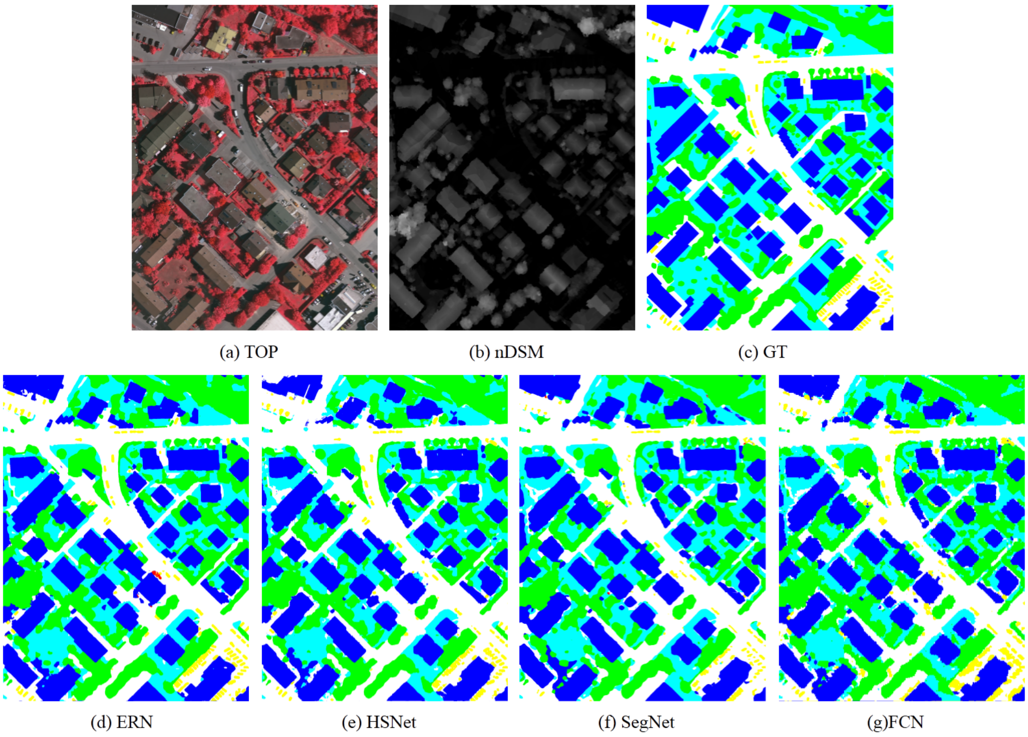

(a) TOP (b) nDSM (c) GT
(d) ERN (e) HSNet (f) SegNet (g)FCN

Figure 8. Full tile prediction. (**a**) TOP (true orthophoto); (**b**) nDSM (normalized digital surface model); (**c**) GT (ground truth). (**d**–**g**) the inference results from ERN, HSNet [24], SegNet [19], and FCN [17], respectively; white: impervious surfaces; blue: buildings; cyan: low vegetation; green: trees; yellow: cars; red: other.

Figure 9 gives more results in detail. From Figure 9a, it can be seen that cars are very close to each other, and it is difficult to separate them with HSNet [24], SegNet [19], and FCN [17]. By introducing the edge loss reinforced structures, the proposed ERN can better separate densely located cars. Figure 9b shows that similar appearance between buildings and impervious surfaces confuses the network, but ERN correctly segmented neighboring regions. Moreover, Figure 9c–e show that the shadows from trees or buildings pose difficulties for semantic segmentation. In this case, ERN segmented the impervious surfaces and low vegetation with a higher accuracy. We argue that the edge loss reinforced structures helped improve the boundary accuracy in regions without effective infromation from the nDSM.

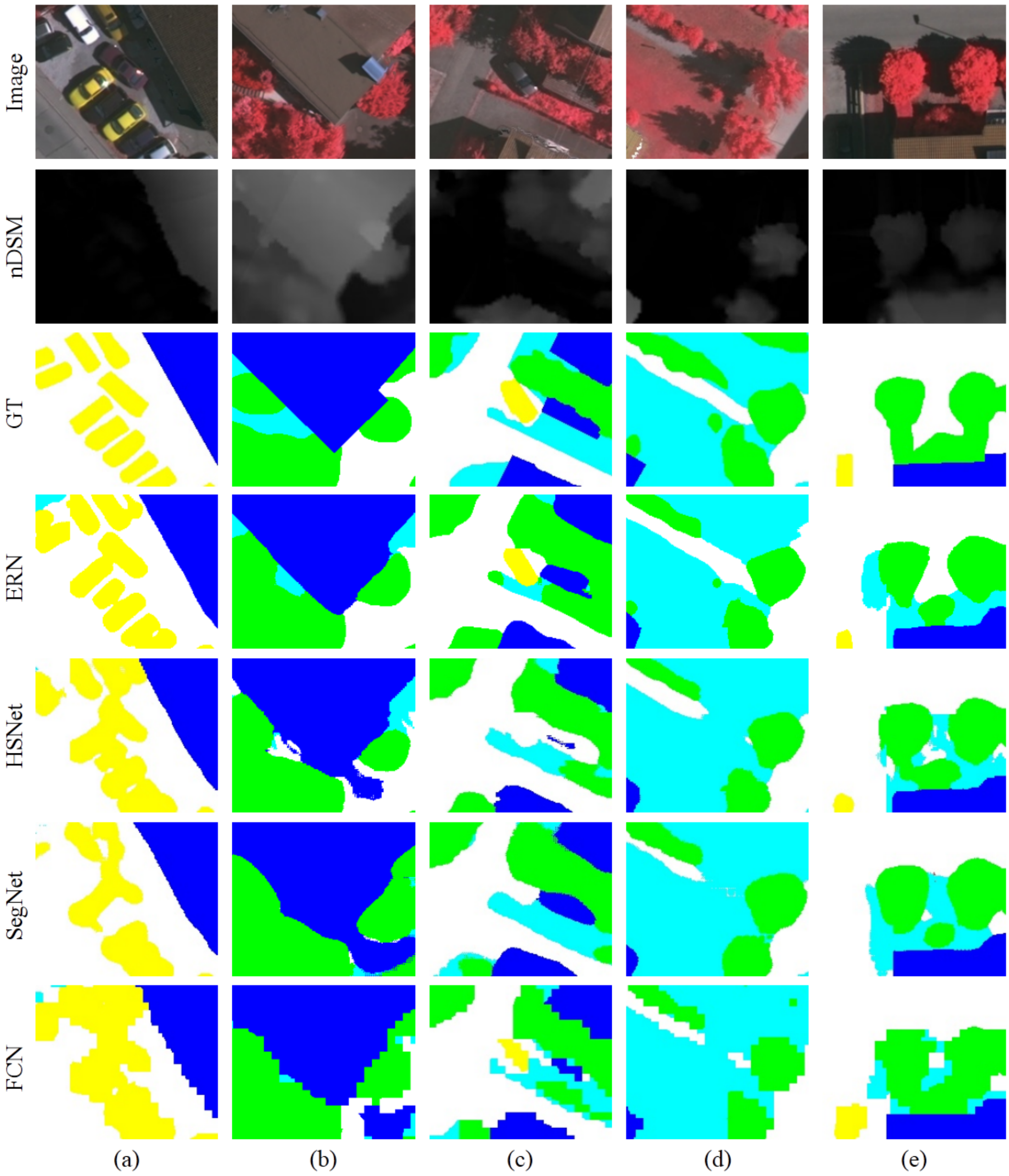


Figure 9. Semantic segmentation results for some local details of the ISPRS Vaihingen Dataset. (**a**) area with dense cars; (**b**) similar appearance between a building and its surroundings; (**c**–**e**) shadow-affected regions; white: impervious surfaces; blue: buildings; cyan: low vegetation; green: trees; yellow: cars; red: other.

4.5. Performance in Shadow-Affected Regions

In this section, we present the performance of semantic segmentation in shadow-affected regions. The major challenge for comparison comes from the fact that there is no ground truth to indicate which pixel is covered by shadow. Traditional unsupervised shadow detection methods [60] often failed in the dark regions, such as black cars and roofs. Therefore, we manually labeled the shadow masks for the test tiles (areas: 11, 15, 28, 30, 34) in ISPRS Vaihingen Dataset [35]. We first pre-processed the original images with the contrast preserving decolorization technology [61], which can help human better separate shadow regions from surroundings. And the semantic ground truth was also utilized to help distinguish whether the pixel belongs to the shadow or dark-color cars during annotation. Figure 10 shows an example of the labeled shadow mask. According to statistics, 15%∼25% of the pixels are in shadow. And the shadow-affected pixels are mainly located in the regions of impervious surfaces and low vegetation.

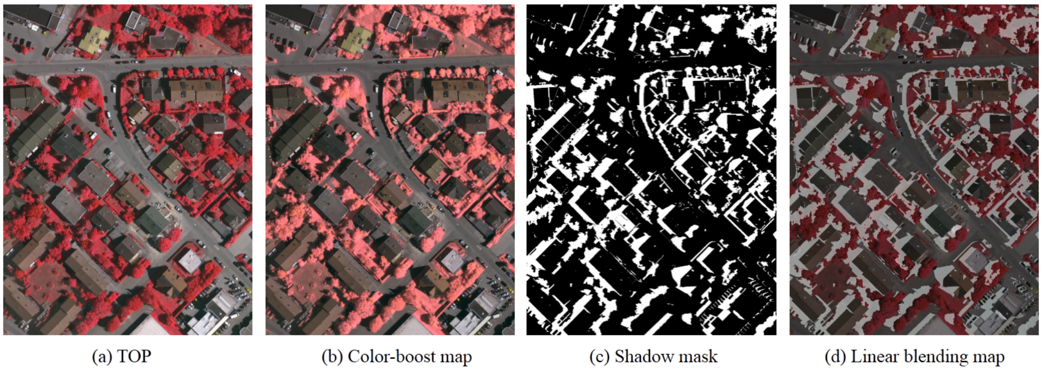

(a) TOP (b) Color-boost map (c) Shadow mask (d) Linear blending map

Figure 10. The example of labeled shadow mask (area 30). (**a**) TOP; (**b**) color-boost map by contrast preserving decolorization [61]; (**c**) manually labeled shadow mask, white pixels denote the shadow; (**d**) linear blending map of (**a**) and (**c**).

The semantic segmentation performance in the shadow-affected regions has been re-evaluated and listed in Table 6. Compared with the results in whole image (see Table 5), the performance in shadow-affected regions is much poorer due to the interference of shadow. From Table 6, it can be observed that the proposed ERN outperforms the other networks. ERN reaches the best performance in F-scores of all classes, average F-score and the overall accuracy, which demonstrates its effectiveness and robustness in shadow-affected regions.

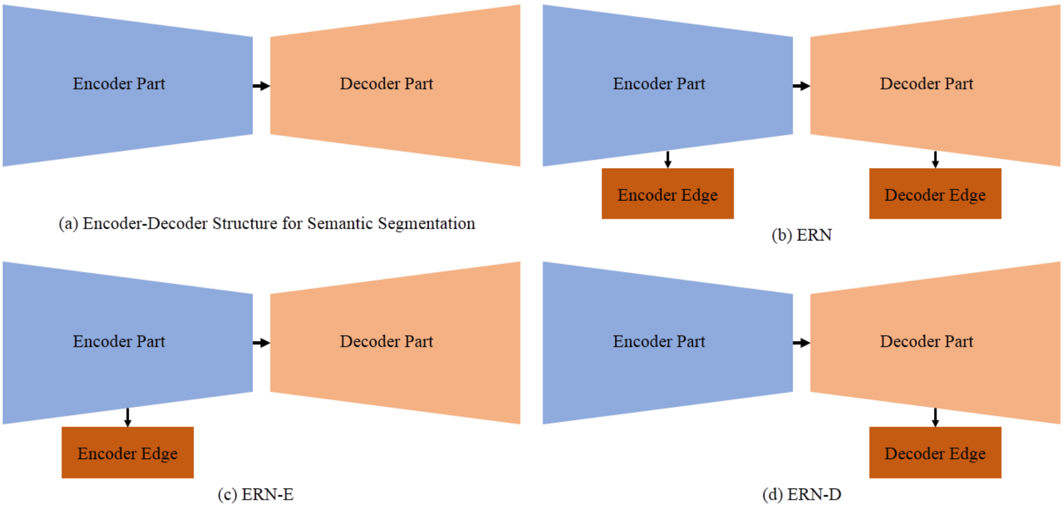
Table 6. Experimental results on the shadow-affected regions in the ISPRS Vaihingen Dataset. Imp.Surf: impervious surface; LowVeg: low vegetation.

Methods	Imp.Surf	Buildings	LowVeg	Tree	Car	Average F-Score	Overall Accuracy
FCN [17]	75.70	66.10	67.20	70.22	29.30	61.70	69.70
SegNet [19]	76.49	69.26	68.50	69.30	26.40	61.99	70.77
HSNet [24]	79.50	69.18	69.51	72.22	51.30	68.34	73.17
ERN	**80.39**	**71.02**	**70.33**	**74.21**	**62.74**	**71.74**	**74.37**

5. Discussion

5.1. Edge Loss Analysis

To evaluate the performance brought by edge loss reinforced structures in the proposed ERN, extensive experiments of different edge loss constraints were further conducted. ERN includes two edge loss reinforced structures: encoder edge and decoder edge. In this section, we explore two variations of ERN—ERN-E and ERN-D, which are shown in Figure 11. ERN-E is the encoder–decoder semantic segmentation net with only the encoder edge, while ERN-D uses only the decoder edge. The training process for ERN-E and ERN-D is identical to ERN.



(a) Encoder-Decoder Structure for Semantic Segmentation

(b) ERN

(c) ERN-E

(d) ERN-D

Figure 11. Variations of ERN–ERN-E and ERN-D–with different edge loss reinforced structures.

Table 7 reports the experimental results on the UAV images for ERN, ERN-E, and ERN-D with different edge loss weight. The edge loss weight in ERN-E and ERN-D are denoted as α_{en} and α_{de}, respectively. Figure 12 shows the comparison of semantic output predictions and edge output predictions of ERN, ERN-E, and ERN-D.

Table 7. Experimental results on the UAV Image Dataset. ERN-E is the encoder–decoder semantic segmentation net with only the encoder edge, while ERN-D is with only the decoder edge. OI: overlap inference.

Methods	Edge Loss Weights	Buildings	Road	Grass	Tree	Land	Average F-Score	Overall Accuracy
ERN-E	1	92.11	80.55	82.84	56.79	95.37	81.53	87.60
	10	94.06	83.93	86.18	67.26	95.62	85.41	90.00
ERN-D	1	92.48	81.41	82.26	60.26	95.23	82.33	88.07
	10	93.34	84.64	89.32	67.95	96.27	86.30	90.78
ERN	10, 10	94.38	84.69	88.82	71.81	96.19	87.18	91.45
	20, 20	**94.43**	**85.27**	**90.17**	**72.43**	**96.38**	**87.74**	**91.90**
ERN-E+OI	10	94.65	85.71	86.88	68.09	95.97	86.26	90.62
ERN-E+OI	10	93.88	86.46	90.33	69.26	96.62	87.31	91.51
ERN+OI	10, 10	94.94	86.67	89.94	73.31	96.54	88.28	92.24
	20, 20	**95.02**	**87.20**	**91.17**	**73.88**	**96.76**	**88.81**	**92.66**

When setting $\alpha_{en} = \alpha_{de} \leq 1$, the performance of ERN-E and ERN-D was slightly decreased compared with the original encoder–decoder framework. We checked the edge structure and found that it output hardly any edge information. When setting $\alpha_{en} = \alpha_{de} = 10$, the performances of ERN-E and ERN-D were improved. We found that the edge structure could correctly output the edge information; see Figure 12. Under the same edge loss weight, ERN ($\alpha_1 = 1$, $\alpha_2 = \alpha_3 = 10$) clearly outperformed the ERN-E ($\alpha_{en} = 10$) and ERN-D ($\alpha_{de} = 10$); see Table 7 and Figure 12. Comparing the edge results of ERN and ERN-D from Figure 12, we found that the edge predictions from the ERN (decoder edge) were superior to those from ERN-D.

We argue that the edge loss reinforced structure can be incorporated into any encoder–decoder architecture with a simple modification. The instructions are as follows: (1) the edge loss weight (α_{edge}) should be larger than the semantic loss weight ($\alpha_{semantic}$), because the edge loss ($\mathcal{L}_{edge}$) is always smaller than the semantic loss ($\mathcal{L}_{semantic}$). A too-small edge loss weight may lead to a failure of edge supervision; (2) the edge loss weight may differ between datasets, which helps the different kinds of losses ($\alpha_{semantic} \cdot \mathcal{L}_{semantic}$, $\alpha_{edge} \cdot \mathcal{L}_{edge}$) adapt to the same order of magnitude; and (3) the performance gain of semantic prediction is proportional to the accuracy of edge prediction, indicating that a better edge detection improves semantic segmentation.

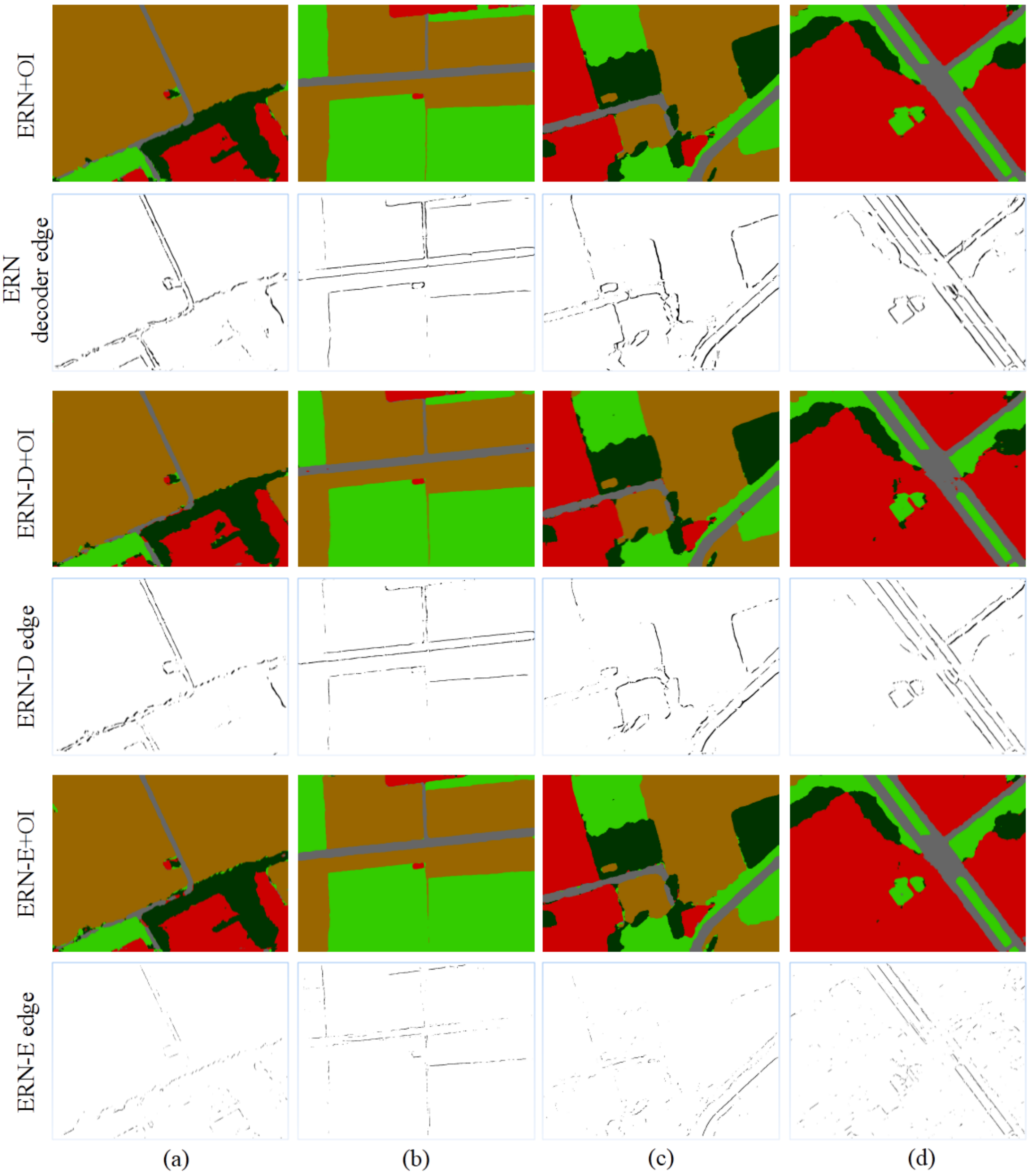


Figure 12. The results of semantic predictions and edge predictions. ERN ($\alpha_1 = 1$, $\alpha_2 = \alpha_3 = 10$), ERN-D ($\alpha_{de} = 10$), ERN-E ($\alpha_{en} = 10$). (**a**–**d**) are four different scenes from the UAV image dataset.

5.2. General Analysis

To the best of our knowledge, the most similar works to our ERN are from Chen et al. [32] and Cheng et al. [33], where they build edge-aware nets to further filter the semantic segmentation results using domain transfer technology and regularization method, respectively. ERN constructs multiple edge loss reinforced structures from the encoder and decoder separately (namely, encoder edge and decoder edge), while only one edge-aware net has been constructed in [32] (similar to our encoder edge) and [33] (constructed by concatenating hierarchical features cross encoder and decoder). Multiple structures and corresponding weighted edge losses are introduced to strengthen the ability of preserving the boundary information rather than post-fine-tuning the semantic segmentation results. The encoder edge loss leverages the benefits of deep supervision in shallow layers like HED [29], and the decoder one aims to further assist the high-level semantic parsing. Moreover, the weighted style of edge loss helps better shape and reinforce the semantic segmentation net.

Compared with the models for general images, the network designed for remote sensing images should not only face the inherent difficulties of semantic segmentation, but also deal with the special issues derived from characters of remote sensing images. Thus, this paper mainly focuses on improving the poor segmentation performance caused by appearance similarity and shadow interference, which are ubiquitous in remote sensing images.

The experimental results in Section 4 demonstrate that our approach achieves state-of-the-art performance on two remote sensing datasets. The proposed approach outperforms reference methods by substantial margins in terms of both average F-score and overall accuracy. In particular, the easily confusing pixels with similar visual appearance have been correctly labeled (See Figure 9b,c). In addition, the semantic segmentation performance is also significantly improved in the challenging situation of shadow interference (See Figures 7a and 9d,e). Specifically, the shadow masks have been manually labeled and the numerical comparison within the shadow-affected regions has been further reported in Section 4.5, which shows the advantage of the proposed ERN.

We attribute the effectiveness of the proposed approach mainly to the design of multiple weighted edge loss reinforced structures in the network. The above two problems are essentially due to the low inter-class variance and the large semantic ambiguity, which make it difficult for the network to correctly distinguish different semantics. By introducing the multiple weighted edge loss reinforced structures, more boundary information can be preserved in the network and further helps to reduce the semantic ambiguity.

It is interesting to find that our approach significantly improves the segmentation accuracy of car in the ISPRS Vaihingen dataset. The low accuracy of car segmentation is usually thought to be caused by the small sample number. However, we find that the surface between cars is often incorrectly classified when the cars are extremely close to each other (see Figure 9a). We argue that the boundary information help the network better segment cars.

5.3. Efficiency Limitation

Even though ERN provides the best overall accuracy and average F-score, it requires the highest segmentation time when compared with other networks. The efficiency of ERN is one of the biggest limitations.

Table 8 shows the average semantic segmentation time per image on the test dataset (100 images for UAV Image Dataset, five images for ISPRS Vaihingen Dataset). The running time list in the Table 8 is the sum of inference time and stitching time. The proposed ERN takes 118.83 s and 19.55 s to finish inference and stitching on the test images of UAV Image Dateset and ISPRS Vaihingen Dataset respectively. The environment is same as Section 4.2.

Table 8. Average semantic segmentation time per image in the experiments.

Average Time (s)	FCN [17]	SegNet [19]	HSNet [24]	ERN
UAV Image Dateset	0.33	0.71	0.96	1.19
ISPRS Vaihingen Dataset [35]	1.01	2.23	3.41	3.91

5.4. Future Work

Possible directions for future research include designing a more powerful edge loss reinforced structure while keeping efficiency for high quality semantic segmentation and automatically learning the edge loss weight rather than based on empirical settings.

6. Conclusions

Semantic segmentation for remote sensing images is a challenging task due to low inter-class variance and interference from areas containing shadows. In this paper, we present a new end-to-end semantic segmentation network. By introducing multiple weighted edge loss reinforced structures, the spatial boundary information is preserved and used to reduce semantic ambiguity. The performance of semantic segmentation is significantly improved in the whole image as well as the shadow-affected regions. On the UAV Image Dataset and ISPRS Vaihingen Dataset [35], the average F-score of ERN has reached 88.81% and 88.64% , and the overall accuracy has reached 92.66% and 88.88%, respectively. In addition, the F-score of Car has been impressively improved nearly 7%. In addition, the edge loss reinforced structures share most network parameters with the original network, indicating that additional structure does not greatly increase model complexity. Finally, we have compared and analyzed different edge loss constraints and clarified the working conditions where edge detection promotes semantic segmentation. The edge loss reinforced structure can be easily integrated into any encoder–decoder semantic segmentation networks. The full implementation in this project is available at: https://github.com/liushuo2018/ERN.

Author Contributions: Conceptualization, S.L. and Y.L.; Data curation, Y.W.; Funding acquisition, W.D. and H.L.; Investigation, S.L. and C.L.; Methodology, S.L., W.D. and Y.L.; Project administration, C.L.; Resources, W.D.; Software, S.L. and Y.L.; Supervision, W.D.; Validation, S.L. and C.L.; Visualization, C.L.; Writing—Original draft, S.L., Y.L. and Y.W.; Writing—Review and editing, W.D. and H.L.

Acknowledgments: This work was supported by the National Key Research and Development Program of China (Grant No. 2016YFB0502600) and the National Natural Science Foundation of China (Grant No. 61601014).

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

Appendix A. Confusion Matrices

In this section, we report the confusion matrices for both the proposed ERN and the reference techniques tested on the UAV Image Dateset and ISPRS Vaihingen Dataset. The values are given in percentages, and the diagonal elements are highlighted in bold.

Table A1. Confusion matrix on the UAV Image Dateset.

		Buildings	Road	Grass	Tree	Land
FCN [17]	**Buildings**	**89.90**	1.47	1.91	6.10	0.62
	Road	7.5	**83.16**	3.5	1.9	3.94
	Grass	2.44	0.63	**74.10**	19.53	3.3
	Tree	8.28	0.26	1.52	**87.90**	2.04
	Land	1.09	0.38	2.66	2.81	**93.06**
SegNet [19]	**Buildings**	**86.49**	4.64	2.34	2.86	3.67
	Road	2.04	**78.67**	2.20	0.66	16.42
	LowVeg	0.82	0.87	**81.04**	13.41	3.87
	Tree	7.52	0.45	2.65	**79.88**	9.5
	Land	0.15	0.18	2.16	0.49	**97.03**
HSNet [24]	**Buildings**	**91.28**	1.97	3.33	3.16	0.26
	Road	8.03	**84.21**	5.32	0.25	2.18
	Grass	2.12	0.77	**92.46**	3.29	1.37
	Tree	10.90	1.02	6.79	**78.44**	2.58
	Land	0.39	0.68	5.46	0.29	**93.18**
ERN	**Buildings**	**92.94**	1.12	1.11	4.49	0.33
	Road	9.57	**84.25**	2.60	1.35	2.23
	Grass	0.94	0.72	**90.28**	6.44	1.62
	Tree	6.77	0.59	7.09	**82.34**	3.21
	Land	0.22	0.37	3.26	0.92	**95.22**

Table A2. Confusion matrix on the ISPRS Vaihingen Dataset. Imp.Surf: impervious surface; LowVeg: low vegetation.

		Imp.Surf	Buildings	LowVeg	Tree	Car
FCN [17]	**Imp.Surf**	**88.99**	3.14	5.39	1.09	1.38
	Buildings	3.89	**93.21**	2.22	0.57	0.11
	LowVeg	5.88	2.47	**74.11**	17.32	0.22
	Tree	0.92	0.37	9.36	**89.35**	0.01
	Car	15.60	1.71	1.00	0.57	**81.11**
SegNet [19]	**Imp.Surf**	**91.68**	2.46	3.87	1.18	0.81
	Buildings	4.16	**93.22**	2.02	0.55	0.05
	LowVeg	6.62	2.44	**73.63**	17.22	0.09
	Tree	0.93	0.46	14.28	**84.32**	0.01
	Car	17.31	0.80	0.90	0.72	**80.27**
HSNet [24]	**Imp.Surf**	**92.64**	2.54	3.71	0.65	0.46
	Buildings	3.50	**94.11**	2.18	0.18	0.03
	LowVeg	6.73	2.44	**78.09**	12.67	0.08
	Tree	1.24	0.35	10.96	**87.44**	0.01
	Car	15.91	1.96	1.22	0.32	**80.59**
ERN	**Imp.Surf**	**91.18**	2.63	4.62	1.13	0.33
	Buildings	2.67	**94.80**	2.15	0.35	0.02
	LowVeg	5.07	1.88	**77.96**	15.06	0.03
	Tree	0.80	0.20	8.82	**90.17**	0.01
	Car	6.39	3.07	0.46	0.71	**89.23**

References

1. Plaza, A.; Plaza, J.; Paz, A.; Sanchez, S. Parallel Hyperspectral Image and Signal Processing [Applications Corner]. *IEEE Signal Process. Mag.* **2011**, *28*, 119–126. [CrossRef]
2. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
3. Zhang, B.; Gu, J.; Chen, C.; Han, J.; Su, X.; Cao, X.; Liu, J. One-two-one networks for compression artifacts reduction in remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2018**. [CrossRef]
4. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
5. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915–1929. [CrossRef] [PubMed]
6. Gaetano, R.; Scarpa, G.; Poggi, G. Hierarchical Texture-Based Segmentation of Multiresolution Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 2129–2141. [CrossRef]
7. Martha, T.R.; Kerle, N.; Van Westen, C.J.; Jetten, V.G.; Kumar, K.V. Segment Optimization and Data-Driven Thresholding for Knowledge-Based Landslide Detection by Object-Based Image Analysis. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 4928–4943. [CrossRef]
8. Yao, X.; Han, J.; Cheng, G.; Qian, X.; Guo, L. Semantic Annotation of High-Resolution Satellite Images via Weakly Supervised Learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3660–3671. [CrossRef]
9. Liu, Y.; Fan, B.; Wang, L.; Bai, J.; Xiang, S.; Pan, C. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. *ISPRS J. Photogramm. Remote Sens.* **2017**. [CrossRef]
10. Marcos, D.; Volpi, M.; Kellenberger, B.; Tuia, D. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models. *ISPRS J. Photogramm. Remote Sens.* **2018**. [CrossRef]
11. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [CrossRef]
12. Wang, P.; Huang, C.; Tilton, J.C.; Tan, B.; Colstoun, E.C.B.D. HOTEX: An approach for global mapping of human built-up and settlement extent. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, TX, USA, 23–28 July 2017.

13. Liu, T.; Abd-Elrahman, A.; Zare, A.; Dewitt, B.A.; Flory, L.; Smith, S.E. A fully learnable context-driven object-based model for mapping land cover using multi-view data from unmanned aircraft systems. *Remote Sens. Environ.* **2018**, *216*, 328–344. [CrossRef]
14. Chen, G.; Weng, Q.; Hay, G.J.; He, Y. Geographic object-based image analysis (GEOBIA): Emerging trends and future opportunities. *GISci. Remote Sens.* **2018**, *55*, 159–182. [CrossRef]
15. Mostajabi, M.; Yadollahpour, P.; Shakhnarovich, G. Feedforward semantic segmentation with zoom-out features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3376–3385.
16. Eigen, D.; Fergus, R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2016.
17. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
18. Noh, H.; Hong, S.; Han, B. Learning Deconvolution Network for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
19. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. *IEEE Transact. Pattern Anal. Machine Intell.* **2017**, *39*, 2481–2495. [CrossRef] [PubMed]
20. Lin, G.; Milan, A.; Shen, C.; Reid, I.D. RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation. *IEEE Conf. Comput. Vis. Pattern Recognit.* **2017**, *1*, 5168–5177.
21. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
22. 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, Quebec, QC, Canada, 10–14 September 2015.
23. Volpi, M.; Tuia, D. Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 881–893. [CrossRef]
24. Liu, Y.; Minh Nguyen, D.; Deligiannis, N.; Ding, W.; Munteanu, A. Hourglass-ShapeNetwork Based Semantic Segmentation for High Resolution Aerial Imagery. *Remote Sens.* **2017**, *9*, 522. [CrossRef]
25. Ghiasi, G.; Fowlkes, C.C. Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
26. Kohli, P.; Torr, P.H.S. Robust higher order potentials for enforcing label consistency. *Int. J. Comput. Vis.* **2009**, *82*, 302–324. [CrossRef]
27. Russell, C. Associative hierarchical CRFs for object class image segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Angers, France, 17–21 May 2010.
28. Arnab, A.; Jayasumana, S.; Zheng, S.; Torr, P.H.S. Higher Order Conditional Random Fields in Deep Neural Networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016.
29. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1395–1403.
30. Bertasius, G.; Shi, J.; Torresani, L. High-for-Low and Low-for-High: Efficient Boundary Detection from Deep Object Features and Its Applications to High-Level Vision. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2015; pp. 504–512.
31. Kokkinos, I. Pushing the Boundaries of Boundary Detection using Deep Learning. In Proceedings of the International Conference on Learning Representations, San Juan, PR, USA, 2–4 May 2016.
32. Chen, L.C.; Barron, J.T.; Papandreou, G.; Murphy, K.; Yuille, A.L. Semantic Image Segmentation with Task-Specific Edge Detection Using CNNs and a Discriminatively Trained Domain Transform. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4545–4554.

33. Cheng, D.; Meng, G.; Xiang, S.; Pan, C. FusionNet: Edge Aware Deep Convolutional Networks for Semantic Segmentation of Remote Sensing Harbor Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2017**, *10*, 5769–5783. [CrossRef]
34. Marmanis, D.; Schindler, K.; Wegner, J.D.; Galliani, S.; Datcu, M.; Stilla, U. Classification with an edge: Improving semantic image segmentation with boundary detection. *ISPRS J. Photogramm. Remote Sens.* **2018**, *135*, 158–172. [CrossRef]
35. ISPRS 2D Semantic Labeling Contest. Available online: http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html (accessed on 7 July 2018).
36. Garciagarcia, A.; Ortsescolano, S.; Oprea, S.; Villenamartinez, V.; Rodriguez, J.G. A Review on Deep Learning Techniques Applied to Semantic Segmentation. *arXiv* **2017**, arXiv:1704.06857.
37. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Available online: http://arxiv.org/pdf/1409.1556v6.pdf (accessed on 18 December 2015).
38. Zhang, B.; Yang, Y.; Chen, C.; Yang, L.; Han, J.; Shao, L. Action Recognition Using 3D Histograms of Texture and A Multi-class Boosting Classifier. *IEEE Trans. Image Process.* **2017**, *26*, 4648–4660. [CrossRef] [PubMed]
39. Luan, S.; Chen, c.; Zhang, B.; han, j.; Liu, J. Gabor Convolutional Networks. *IEEE Trans. Image Process.* **2018**, *27*, 4357–4366. [CrossRef] [PubMed]
40. Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 603–619. [CrossRef]
41. Felzenszwalb, P.F.; Huttenlocher, D.P. Efficient Graph-Based Image Segmentation. *Int. J. Comput. Vis.* **2004**, *59*, 167–181. [CrossRef]
42. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Susstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef] [PubMed]
43. Shotton, J.; Johnson, M.; Cipolla, R. Semantic texton forests for image categorization and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
44. 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, 26 June–1 July 2016; pp. 770–778.
45. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. In Proceedings of the International Conference on Learning Representations, San Juan, PR, 2–4 May 2016.
46. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.P.; Yuille, A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
47. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.P.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
48. Persello, C.; Stein, A. Deep Fully Convolutional Networks for the Detection of Informal Settlements in VHR Images. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 2325–2329. [CrossRef]
49. Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P.H.S. Conditional Random Fields as Recurrent Neural Networks. In Proceedings of the IEEE Conference on International Conference on Computer Vision, Los Alamitos, CA, USA, 7–13 December 2015.
50. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. High-Resolution Aerial Image Labeling with Convolutional Neural Networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7092–7103. [CrossRef]
51. Lee, C.; Xie, S.; Gallagher, P.W.; Zhang, Z.; Tu, Z. Deeply-Supervised Nets. In Proceedings of the International Conference on Artificial Intelligence and Statistics, San Diego, CA, USA, 9–12 May 2015; pp. 562–570.
52. Hou, Q.; Cheng, M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P.H.S. Deeply Supervised Salient Object Detection with Short Connections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honplulu, HI, USA, 21–26 July 2017; pp. 5300–5309.
53. Ke, W.; Chen, J.; Jiao, J.; Zhao, G.; Ye, Q. SRN: Side-Output Residual Network for Object Symmetry Detection in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honplulu, HI, USA, 21–26 July 2017; pp. 302–310.

54. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
55. Kampffmeyer, M.; Salberg, A.B.; Jenssen, R. Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
56. Hwang, J.J.; Liu, T.L. Pixel-wise Deep Learning for Contour Detection. *arXiv* **2015**, arXiv:1504.01989.
57. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
58. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE International Conference on Computer Vision, Washington, DC, USA, 7–13 December 2016.
59. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the Acm International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
60. Tsai, V.J.D. A comparative study on shadow compensation of color aerial images in invariant color models. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 1661–1671. [CrossRef]
61. Lu, C.; Xu, L.; Jia, J. Contrast preserving decolorization. In Proceedings of the IEEE International Conference on Computational Photography, Seattle, WA, USA, 28–29 April 2012; pp. 1–7.