E-D-Net: Automatic Building Extraction From High-Resolution Aerial Images With Boundary Information

Yuting Zhu, Zili Liang, Jingwen Yan, Gao Chen, and Xiaoqing Wang

***Abstract*—The automatic extraction of buildings from high-resolution aerial imagery plays a significant role in many urban applications. Recently, the convolution neural network (CNN) has gained much attention in remote sensing field and achieved a remarkable performance in building segmentation from visible aerial images. However, most of the existing CNN-based methods still have the problem of tending to produce predictions with poor boundaries. To address this problem, in this article, a novel semantic segmentation neural network named edge-detail-network (E-D-Net) is proposed for building segmentation from visible aerial images. The proposed E-D-Net consists of two subnetworks E-Net and D-Net. On the one hand, E-Net is designed to capture and preserve the edge information of the images. On the other hand, D-Net is designed to refine the results of E-Net and get a prediction with higher detail quality. Furthermore, a novel fusion strategy, which combines the outputs of the two subnetworks is proposed to integrate edge information with fine details. Experimental results on the INRIA aerial image labeling dataset and the ISPRS Vaihingen 2-D semantic labeling dataset demonstrate that, compared with the existing CNN-based model, the proposed E-D-Net provides noticeably more robust and higher building extraction performance, thus making it a useful tool for practical application scenarios.**

***Index Terms*—Convolutional neural networks (CNNs), edge information, fully convolutional networks, high resolution, remote-sensing, semantic segmentation.**

I. Introduction

ESTABLISHING and updating large scale building maps from remote sensing imagery is a tedious, expensive, and often manual process. It is widely used in urban dynamics, such as estimating population and facilitating urban planning, and many other applications in socio-economics studies [1], [2]. With the rapid development of sensor technology, a large number of aerial image data are becoming more accessible and affordable. High-resolution aerial images provide sufficient structural and texture information for image segmentation while also raise new challenges for automatically extracting buildings from aerial images. The challenges arising from the variations in appearance of buildings, different scales of buildings, and occlusions increase the difficulties [3]. Exploring effective and efficient algorithms to realize building extraction automatically is highly demanded.

Recently, due to its powerful ability in effectively extracting high-level features without the involvement of human ingenuity in feature engineering, the state-of-the-art convolutional neural network (CNN) has gained notable success in a wide range of applications in remote sensing field, e.g., heterogeneous image change detection [5], hyperspectral image classification [6], and object detection in remote sensing images [7]. More recently, CNN has shown promising results in the field of automatic building extraction and various CNN architectures have been adopted for automatic building extraction [2], [8], [9]. While early works mainly use restricted Boltzmann machine [8] or patched-based CNN [10], recent progress has taken the fully convolutional network (FCN) [9], [11].

FCN is the first network that effectively converts classification deep CNNs for dense labeling. This key feature permits FCN to take advantage of the pretrained classification CNN model and to generate prediction maps that are the same size as the input images [2]. With many public building datasets, FCN has achieved a remarkable performance in accuracy as well as computational time. However, repeated convolutional and pooling operations employed in FCN result in a local reception field and ignore some detailed information, leading to the poor prediction [2], [12]. In more detail, FCN-based models such as Deeplabv3 finally use 4 times bilinear interpolation for upsampling, which is very unfavorable for the prediction of building edges. Therefore, most of the FCN-based approaches are lacking boundary details for small buildings, as shown in Fig. 1(c). Identifying accurate boundary for buildings is particularly important in the building extraction task, as it aims to establish a building footprint map that provides the outline of a building drawn along the exterior walls, with a description of the exact size, shape, and its location. However, "Not all pixels are equal: Difficulty-aware semantic

Manuscript received December 29, 2020; revised March 4, 2021 and April 2, 2021; accepted April 9, 2021. Date of publication April 19, 2021; date of current version May 17, 2021. This work was supported in part by the NSF of China under Grant 61672335 and Grant 61601276 and in part by the Department of Education of Guangdong Province under Grant 2016KZDXM012, Grant 2017KCXTD015, and Grant 2016A030310077. *(Corresponding authors: Gao Chen and Xiaoqing Wang.)*

Yuting Zhu and Xiaoqing Wang are with the School of Electronic, and Communication Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: 15951001995@163.com; wangxq58@mail.sysu.edu.cn).

Zili Liang and Jingwen Yan are with the Department of Electronic Engineering, Shantou University, Shantou 515063, China (e-mail: 20zlliang1@stu.edu.cn; jwyan@stu.edu.cn).

Gao Chen is with the School of Electrical Engineering, and Intelligentization, Dongguan University of Technology, Dongguan 523808, China (e-mail: chengao@dgut.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3073994

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

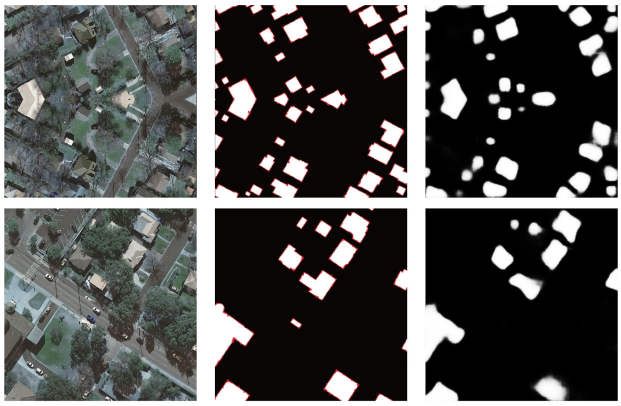


Fig. 1. (a) Two examples of high-resolution aerial images. (b) Difficulty level of pixels of building footprints. (c) Corresponding predictions by a fully convolutional network [4]. The difficulty level of pixels is visualized in the middle image, where the pixels are divided into three sets, including the "easy" (Black), "moderate" (White), and "extremely hard" (Red) sets. It can be seen that the prediction overlaps well with the label but fails to reflect the boundaries that exist in label.

segmentation via deep layer cascade" [13] has made detailed statistics on semantic segmentation, in which the pixels are easy to lead to misjudged in classifying, the edge of the object, as shown in red in Fig. 1(b). Jiang *et al.* [14] also pointed out that in the task of building extraction, different categories of boundary recognition capabilities are different, which will have a certain impact on the final segmentation accuracy.

To address this problem, several works have considered the boundaries of objects when designing the network, leading to a superior performance of object extraction. Marmains *et al.* [15] proposed to combine feature maps from multiple networks at different feature scales and make the final predictions on top of these concatenated feature maps. Their proposed network outputs a set of scale-dependent class boundaries before each pooling, which are integrated into the final multiscale edge prediction. Bischke *et al.* [16] addressed the problem of preserving semantic segmentation boundaries in high-resolution satellite imagery by introducing a new cascaded multitask loss. Their goal is to rely besides the semantic term also a geometric term, which incorporates the boundary information of the segmentation mask into a single loss function. In [17], Tu *et al.* designed an edge-oriented region growing algorithm, where growing seeds are selected from the airport support regions with the help of edge information in SAR images. Chen *et al.* [18] introduced a novel automatic building extraction method that integrates LiDAR data and high spatial resolution imagery using adaptive iterative segmentation and hierarchical overlay analysis based on data fusion, which adapted to the variability of building shape and the environmental complexity surrounding buildings.

Recently, some novel two-step methods have been proposed to improve extracting buildings from aerial images. Li [19] *et al.* proposed an improved model based on morphological methods to reduce edge misclassification, and then further made full use of the saliency features of buildings to improve the expression of edge information. Similarly, the method of Xie [20] *et al.* includes two steps. First, the improved network is applied to achieve pixel-level segmentation of buildings. Second, used morphological filtering to optimize building boundaries, and improve boundary regularity. Similar to our method, Jiang *et al.* [14] proposed a predictive optimization architecture, which consists of an encoder–decoder network and residual refinement modules responsible for prediction and refinement. To enhance the building's expression ability, the authors also introduce a composite loss function. Although these works could improve the segmentation accuracy, one drawback of these methods is that the boundary information is not applied directly.

As shown in Fig. 1(b), the difficulty level (e.g., recognizability) of pixels is visualized in the middle image, where pixels are partitioned into three sets, including easy (Black), moderate (White), and extremely hard (Red) sets. In this article, we aim to improve the recognition ability of arbitrary shaped buildings by improving the segmentation accuracy of the "extremely hard" pixels. Li *et al.* [13] show that 70% pixels in HS are located at object boundaries, which have large ambiguity. We extract boundary deviates from the valid $5-10$ pixels as edge information and take full advantage of them. In this way, a novel semantic segmentation neural network named edge-detail-network (E-D-Net) is proposed for building segmentation from visible aerial images. The network framework consists of three components. First, the edge information generation network (E-Net) is designed to extract edge information. Then, the detail recovery network (D-Net) is used to refine the results of E-Net and get a prediction with higher detail quality. Different from the work in [14] using different loss functions to make the network pay attention to the significance map of the different-level pixels, we propose a new fusion strategy to ensembling the outputs of the two networks in a weighted manner. Experimental results on INRIA aerial image labeling dataset and the ISPRS Vaihingen 2-D semantic labeling dataset demonstrate that the performance of E-D-Net is competitive with that of several state-of-the-art methods.

The main contributions of this article can be summarized as follows.

1) By considering the boundary and details of buildings simultaneously, a novel semantic segmentation neural network E-D-Net for automatic building extraction is proposed in this article. To the best of authors' knowledge, this article makes the first attempt to directly apply the edge information of the target into the network for building extraction.
2) A novel fusion strategy is introduced to combine the edge information and refine results, and to make the whole network cooperate normally, which is crucial for end-to-end training.
3) Experimental results on INRIA aerial image labeling dataset and the ISPRS Vaihingen 2-D semantic labeling dataset demonstrate that the proposed E-D-Net can address the problem of produce predictions with poor boundaries, and has achieved certain improvements in terms of accuracy and intersection over union (IoU).

The rest of this article is organized as follows. Section II describes our proposed E-D-Net in detail. Experimental results

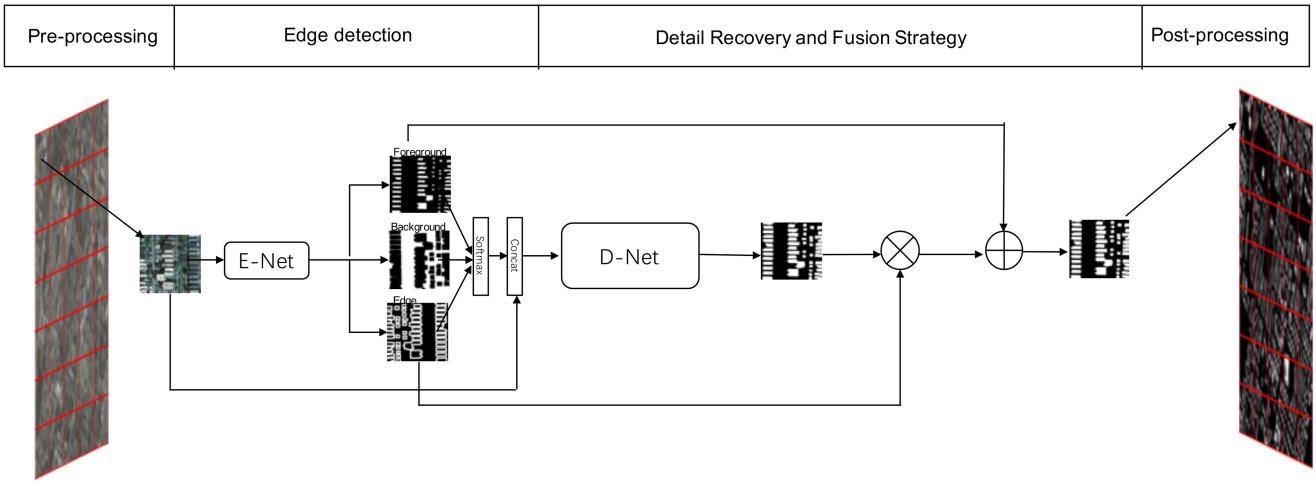


Fig. 2. E-D-Net framework for high-resolution satellite image segmentation. Our model consists of E-Net for edge detection and D-Net for detail recovery, with a new fusion strategy to generate the final result.

on the INRIA aerial image labeling dataset and the ISPRS Vaihingen 2-D semantic labeling dataset are presented in Section III. Section IV concludes this article.

II. Our Method

In this section, we elaborate the design of the proposed E-D-Net. The whole architecture of E-D-Net is shown in Fig. 2. The entire network takes a remote sensing image (usually a three-channel RGB image) as an input and directly outputs a one-channel mask of the same size. Note that there is no other auxiliary information to get the edge information of the image.

Our goal is to capture delicate building edge information and make full use of it to improve the segmentation predictions of building footprints. To this end, in E-D-Net two different subnetworks are designed to handle the two tasks separately and a new fusion strategy is proposed to achieve the shared representation of semantics and geometric features. In what follows, we elaborate the key elements of E-D-Net, including two subnetworks E-Net and D-Net, new fusion strategy, and implementation details.

A. Edge Information Generation Network

E-Net is designed to distinguish buildings from background and classify the pixels with extremely hard level between the building and the background as edges. This can be defined as a three-class segmentation task, i.e., foreground, background, and edge regions. Therefore, the output of E-Net is a three-channel map indicating the possibility that each pixel belongs to each of the three classes. In general, E-Net can be implemented by any of the state-of-the-art segmentation networks. In this article, we choose the FCN-based neural network U-Net [21] to construct E-Net for its efficacy. However, as mentioned in Section I, repeated convolutional and pooling operations employed in FCN-based methods results in a local reception field and ignore some detailed information [see Fig. 1(c)]. In order to overcome this problem, we added additional dilated convolutional layers to the E-Net to increase the receptive field of the feature points and preserve more detailed information. Dilated convolution is an effective kernel that can be used to adjust the receptive fields of feature points without reducing the resolution of the feature maps. As shown in Fig. 3(a), we put three dilated convolution layers with dilation rate of 1, 2, 5 in the center part of the E-Net, then the acceptance field of each layer will be 3, 7, 17. Due to the introduction of the additional dilated convolution layers, the feature points on the last center layer will cover main part of the first center feature map, which improves the receptive field of the feature points. The dilated convolutional layers have shown strong ability to increase the segmentation accuracy and we will verify this in the ablation experiments.

In summary, the proposed E-Net has the following characteristics.

1) E-Net is a supervised learning operator. Its role is learning to extract the building edge information, we need rather than all the edge information in the image. Fig. 4 is an example of E-Net classification, in which Fig. 4(b) is the E-Net label we made using the dilated and eroded operations and Fig. 4(f) is the building edge information.
2) E-Net helps to steer the loss function. We have designed a composite loss function (2) in the fusion strategy, and the training errors in E-Net can be easily fed to the corresponding components through the backpropagation algorithm of the neural network, thereby achieving end-to-end training of the entire network.

B. Detail Recovery Network

E-Net aims to get edge information. Compared with E-Net, D-Net aims to refine the results of E-Net and get a prediction with higher detail quality. Many methods have proved that fusing different levels of semantic information in a neural network can obtain more accurate segmentation results. Deng *et al.* [22] optimized feature maps on multiple scales due to the relatively small area of the classical receptive field. Lee *et al.* [23] used

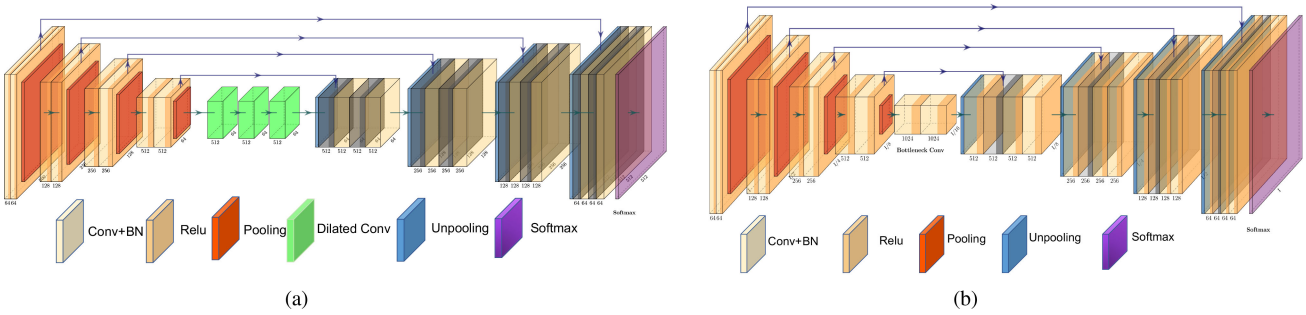


Fig. 3. (a) Structure of E-Net. It is based on U-Net, using VGG11 as the encoder, and the central part contains the dilated convolution layers with dilation rate of 1, 3, 5. (b) Structure of D-Net. It is also based on U-Net, using VGG11 as the encoder.

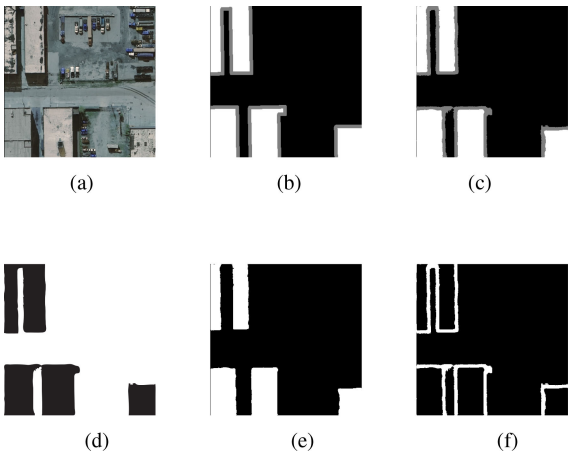


Fig. 4. Example of E-Net classification. (a) Color Image. (b) E-Net Label. (c) Classification result predicted by E-Net. (d) Background information of E-Net classification. (e) Foreground information of E-Net classification. (f) Building edge information of E-Net classification.

several methods to group different levels of semantic information to extract complex building boundaries from lidar and photogrammetric imagery. In this article, to establish a relationship between the low-level image primitives (original images) and the higher level semantic information (the outputs of E-Net: the foreground, background, and edge information), we cascaded the original image with the three-channel segmentation result from E-Net as a six-channel input.

Similar to E-Net, D-Net can be implemented by any of the FCN-based networks. We still choose U-Net [21] but we did not introduce additional dilated convolutional layers in order to increase training speed. As shown in Fig. 3(b), the encoder of the D-Net has eight convolutional layers and four pooling layers. The decoder has eight convolutional layers and four unpooling layers.

In summary, the proposed D-Net has the following characteristics.

(1) D-Net is a refinement model. Its role is helping refine the results of E-Net. In the training phase, the training label of D-Net is the ground truth label provided by the dataset.
(2) The loss function of D-Net is mean squared loss function, which helps to get a prediction with higher detail quality.

C. Fusion Strategy

The E-Net takes the aerial image as input and roughly extract edge region, and the D-Net takes the predicted foreground, background, edge, and original image as input and directly outputs the predicted result. To combine the edge semantic information preserved by E-Net and the refine results predicted by D-Net, a fusion strategy is proposed in this section.

Let $\boldsymbol{F}$, $\boldsymbol{B}$, $\boldsymbol{E}$ denote the foreground, background, and edge region predicted by E-Net, respectively. They are fed into a softmax layer to get the probability estimate of the pixel position, which can be written as $\boldsymbol{F_s}$, $\boldsymbol{B_s}$, and $\boldsymbol{E_s}$. Note that if a pixel locates in the edge region, it means that it has a higher recognition difficulty level. In this case, we perform element-wise product between $\boldsymbol{E_s}$ and the output of the D-Net $\boldsymbol{P_D}$ to get a finer probabilistic estimation of edge. The fusion strategy can be written as

$$\boldsymbol{P} = \boldsymbol{F_s} + \boldsymbol{E_s} \odot \boldsymbol{P_D} \tag{1}$$

where $\boldsymbol{P}$ denotes the final output after the fusion process, $\odot$ is the element-wise product operation between two matrices. Formally, during training process, the formula (1) represents the combination of the coarse semantics and fine detailed information. The element-wise product between $\boldsymbol{E_s}$ and $\boldsymbol{P_D}$ indicates fine edge details. From this formula we can see that when $\boldsymbol{E_s}$ is close to $\mathbf{0}$, $\boldsymbol{P}$ is close to $\boldsymbol{F_s}$, and when $\boldsymbol{E_s}$ is close to $\mathbf{1}$, P is close to $\boldsymbol{P_D}$.

Fusing the features extracted by the two networks can naturally improve segmentation accuracy. In this way, the fusion strategy we designed is a good way to backpropagate losses back to individual components for end-to-end training purpose.

D. Composite Loss Function

During the training process, we define a composite loss function corresponding to our fusion strategy

$$L = \lambda L_E + (1 - \lambda) L_P \tag{2}$$

where L_E represents the loss function of E-Net, L_P represents the loss function of fusion module, and λ is the regularization parameter, which controls the tradeoff between the two loss functions. In the experiment, to keep E-Net stable and not to affect the fusion module, we set λ to a small value to constrain it. By cross validation, we set $\lambda = 0.05$. Since D-Net has obtained

more refined results, we do not hope that the fusion strategy will have a great impact on the optimization of the two subnetworks. We define it as a regression task. The loss function of fusion module L_P is constructed based on the mean squared loss function

$$L_P = \|\boldsymbol{P} - \boldsymbol{G}\|_2^2 \tag{3}$$

where $\boldsymbol{G}$ denotes the label of building ground truth provided by the dataset.

As for the loss function of E-Net, it is defined as a three-category segmentation task, we apply cross-entropy (CE loss) [24] for classification

$$L_E = -\sum_{i}^{c} y^* \log(\tilde{y}) \tag{4}$$

where L_E is the CE loss function, where y^* represents the ground truth of the pixel, $\tilde{y}$ represents the predicted probability of the pixel, and c represents the category of the pixel.

E. Implementation Details

In this section, we will detail the setup of hyperparameters and the implement of training, including the training of the two subnets and the end-to-end training.

1) Preprocessing: In theory, our network can take images with any size as input, but it requires a lot of GPU memories to store feature maps when the image size is large. Since the sizes of the Aerial Image Labeling Dataset are 5000×5000, we randomly crop a 384×384 square image from the input image to reduce the memory usage. In order to significantly improve the performance of the network in terms of accuracy, we use random rotations in the range of [0360] with reflection padding, random horizontal flip and mean subtraction for data augmentation.

2) Hyperparameters Setting: In the training process, the batch size is set as 8, and the weight decay is set as 0.995. We chose cosine annealing [25] with an initial learning rate as 0.001 and a minimum learning rate as 0.0001 for training. All network parameters are optimized with Adam.

3) Training: Pretraining has proven to play an important role in deep CNNs [26]. Following this idea, we first pretrain the two subnets, E-Net and D-Net, and then fine-tune the entire network in an end-to-end manner.

First, pretrain E-Net. As mentioned above, we take a three-channel image as input and produce foreground, background, and edge probability maps of the same size as the original image. The encoder portion of E-Net is based on VGG11, which can be initialized using models trained in the ImageNet classification contest [12]. During the training process of E-Net, the weighted sum loss for classification in (4) is employed.

Second, pretrain D-Net. As shown in Fig. 3, D-Net takes a six-channel array of E-Net output and the original image as input. During the training process of D-Net, we applied the original images with E-Net's ground truth as input to train D-Net alone. D-Net is focused on detail recovery and the regression loss (3) is employed.

Finally, the entire model is initialized with pretrained E-Net and D-Net during end-to-end training process. We utilize three-channel remote sensing images as input, and output edge probability map and final probability map.

4) Postprocessing: In the postprocessing, the output of the network is a one-channel image with the same size as the input image, where the probability score of each pixel is in the range of [0, 1]. We set the threshold as 0.45 to convert the threshold graph into a binary graph through experiments on the validation set. The output patches are assembled into tiles of the original size of the dataset and overlapping areas near the edges are down-weighted.

III. Experiments

In this section, some experimental results based on INRIA aerial image labeling dataset are provided to demonstrate the performance of the proposed E-D-Net. All experiments are carried out on computer servers with one GPU card (NVIDIA GeForce GTX TITAN XP).

A. Experimental Setups

1) Dataset: The Inria aerial image labeling dataset is a benchmark database of labeled imagery that covers varied urban landscapes, ranging from highly dense metropolitan financial districts to alpine resorts [27]. This dataset covers 810 km^2 area in 10 different cities with a spatial resolution of 0.3 m and is divided into two equal sets (each 405 km^2) for training and testing. The dataset consists of 3 band ortho-RGB images, and the training labels consist of ground truth data in two semantic categories: buildings and nonbuildings. The training set covers some cities in Austin, Vienna, Chicago, Kitsap County, and western Tyrol. The test set covers some cities in Innsbruck, San Francisco, Bellingham, Bloomington, and Eastern Tyrol. Each city has 36 images with a resolution of 5000×5000 pixels. Each image covers an area of 1500×1500 m^2 on the ground, and the images do not overlap. We divide the training set into the following two parts according to the suggestions in [27]: 1) the first five pictures of each city are composed into a validation set; 2) the remaining pictures are composed into a training set.

In the experiments, we choose the INRIA aerial image labeling dataset for the following two reasons: 1) the training and the test datasets are from different cities without overlap, that is, all images and labels of 5 cities (Austin, Western-Tyrol, Kitsap, Chicago, Vienna) are provided for training, and all images of the other five cities (Innsbruck, Eastern-Tyrol, San Francisco, Bellingham, Bloomington) as test data are not provided with labels. All of our test results will submit to the official evaluation server; 2) this dataset covers different urban settlements, such as Bloomington and San Francisco. The density and overall characteristics of these cities have large deviations, and their buildings have different shapes (such as flat roof and cupola). For these reasons, we believe that the INRIA aerial image labeling dataset is an ideal choice for evaluating the generalization capabilities of the network.

We also conduct experimental evaluations on the ISPRS Vaihingen 2-D semantic labeling challenge dataset [28]. Vaihingen is a relatively small village with many detached buildings and small multistory buildings. It contains 3-band IRRG (Infrared,

Red, and Green) image data, corresponding digital surface model (DSM) and normalized DSM data. Overall, there are 33 images of about 2500×2000 pixels at a ground surface distance of about 9 cm. All images have corresponding ground truth images. There are the following five labeled categories: building, low vegetation, tree, car, and impervious surface. In this article, we focus on building extraction, so we only apply the building categories of ISPRS Vaihingen 2-D dataset to validate our method. For corresponding benchmark evaluation, we follow the data partition way, which uses 24 images as the training set and 14 as the test set.

For E-Net training, it classifies the inputs in the following three categories: foreground, background, edges. We define ten pixels inside the edge of the building's truth label as the edges of the image, because this part is more likely to produce poor prediction when segmenting. As we mentioned in Section II-A, the E-Net can be defined as a three-class segmentation task. We did not put a certain probability threshold for this decision. E-Net directly outputs three feature maps of the same size as the input, and then fed into a softmax layer to get the probability estimate of the pixel category. We use eroded operations to make labels for E-Net. Dilated and eroded operations are morphological operations in image processing. Fig. 4(b) shows an example of erosion. The gray lines in the image are the building edges defined by us. From this example, we can see that erosion shrinks or thins buildings in a binary label. We use this newly designed label as the ground truth label of E-Net. To make the result more robust, we accept different kernel sizes of dilating between 5 and 10. For more details about eroded operation, see [29].

2) Evaluation Metrics: Two different quality metrics are taken to evaluate the performance of the proposed E-D-Net in the INRIA aerial image labeling dataset, i.e., pixel accuracy (Acc) and region IoU [30]. The "Acc" is the percentage of correctly classified positive pixels among all pixels that are predicted to be positive. The "IoU" is equal to the number of pixels marked as buildings in both the prediction and the reference, divided by the number of pixels marked as buildings in the prediction or the reference. The calculation formulas of the two metrics can be written as follows:

$$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+F P+FN+TN}} \tag{5}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP+FP+FN}} \tag{6}$$

where TP denotes true positives (correctly extracted building pixels), FP denotes false positives (pixels mislabeled as buildings in results), TN denotes true negatives (correctly identified nonbuilding pixels), and FN denotes false negatives (pixels incorrectly labeled as nonbuildings or can be interpreted as missed building pixels).

Since our work focuses on boundary improvement, we provide another quality metric to evaluate the performance of edge saving. It is the edge saving index (ESI) [31], which is used to evaluate the edge retaining capability of the methods. It is defined as follows:

$$\text{ESI} = \frac{\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\sqrt{(\hat{\boldsymbol{u}}_{i,j}-\hat{\boldsymbol{u}}_{i+1,j})^2+(\hat{\boldsymbol{u}}_{i,j}-\hat{\boldsymbol{u}}_{i,j+1})^2}}{\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\sqrt{(\boldsymbol{u}_{i,j}-\boldsymbol{u}_{i+1,j})^2+(\boldsymbol{u}_{i,j}-\boldsymbol{u}_{i,j+1})^2}} \tag{7}$$

where u denotes the pixel values of the ground truth label, and $\hat{u}$ denotes the pixel values of the predict result. Larger ESI values indicate stronger edge retaining capability.

Furthermore, the evaluation indexes also include the time needed by each model to complete one training epoch, the size of network parameters.

B. Main Results

In order to show the performance and effectiveness of the proposed E-D-Net, the following state-of-the-art methods are selected as the comparison.

1) INRIA [32], benchmark provided by the open source of the INRIA aerial image labeling dataset. It derived a so-called MLP network on top of the base FCN (which convey information at different resolutions and with different receptive fields).
2) AMLL [32], the winner of the first INRIA dataset competition. It used the original U-Net architecture from [33], with a single major modification by only applying half filters of [33] at each layer.
3) FCN [9], a classic image segmentation algorithm based on the neural network. In this article, we used the VGG network as its encoder.
4) SegNet [34], an improved image segmentation algorithm based on FCN, by making modifications in the encoder and decoder parts of FCN.
5) LinkNet [35], an efficient semantic segmentation neural network, which takes the advantages of skip connections, residual blocks, and encoder–decoder architecture;
6) U-Net (VGG11) [21], a classical U-Net architectures with a pretrained encoder.
7) DeeplabV3 [36], an off-the-shelf state-of-the-art network for semantic segmentation, it achieves the best results on multiple semantic segmentation tasks.
8) Multitask learning [16], a recent state-of-the-art aerial image building segmentation algorithm that introduces a novel multitask loss to address the problem of preserving semantic segmentation boundaries in aerial imagery.
9) Building-A-Nets [37], a novel deep adversarial network joining with the deep convolutional network and an adversarial discriminator network to segment building rooftops.

To make a fair comparison among these methods, all networks use the default hyperparameter settings and optimizer suggested in their respective papers. Recently, it is well known that the performance of neural networks based on deep learning has a great relationship with training strategy, such as learning rate, optimizer, selection of hyperparameters, and so on. AMLL followed the training strategy in [38] at that time. In this article, we do not use U-Net to reduce the parameters. In addition, we follow the training strategy proposed in [25]. The difference

TABLE I
NUMERICAL EVALUATION ON INRIA TEST SET

Method	BellingHam		Bloomington		Innsbruck		San Francisco		EastTyrol		Overall				
	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	ESI	Time(h)	Params(M)
Inria	56.11	95.37	50.40	95.27	61.03	95.37	61.38	87.00	62.51	96.61	59.31	93.93			
AMLL	67.14	96.64	65.43	96.73	72.27	96.66	75.72	91.80	75.67	97.70	72.55	95.91			
FCN	69.02	96.88	58.69	96.06	67.06	96.07	65.55	88.54	72.74	97.49	66.28	95.01	68.4	2.4	57.00
SegNet	70.60	97.03	64.67	96.58	72.33	96.65	69.36	89.55	76.70	97.90	70.29	95.54	70.1	2.2	56.20
LinkNet	69.59	96.97	66.51	96.82	75.25	97.07	73.27	91.10	78.39	98.08	72.83	96.01	74.4	**0.9**	**22.00**
U-Net	70.14	96.91	69.31	97.06	74.98	96.99	74.55	91.52	78.71	98.09	73.83	96.12	76.4	1.7	26.97
DeepLabV3	72.36	97.01	75.45	97.53	76.39	97.05	78.38	92.48	80.11	98.16	77.04	96.45	83.2	2.5	56.72
E-D-Net	**73.12**	**97.22**	**75.58**	**97.64**	**77.66**	**97.31**	**79.81**	**93.26**	**80.61**	**98.25**	**78.08**	**96.73**	**85.3**	3.2	54.42

Note: The best result in each column is in boldface.

TABLE II
NUMERICAL EVALUATION ON INRIA VALIDATION SET

Method	Austin		Chicago		Kitsao Co.		West Tyrol		Vienna		Overall	
	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.
Inria	61.20	94.20	61.30	90.43	51.50	98.92	57.95	96.66	72.13	91.87	64.67	94.42
FCN	69.37	95.02	74.08	94.80	68.36	95.11	67.36	98.79	72.69	93.55	71.27	95.45
LinkNet	71.62	95.63	77.13	95.47	74.32	95.69	69.24	**98.80**	77.33	94.28	75.78	95.97
U-Net	77.29	96.69	68.52	92.40	72.84	**99.25**	**75.38**	98.11	78.72	93.79	76.55	96.05
Multi-Task	72.43	95.71	77.68	95.60	72.28	95.81	64.34	98.76	76.15	94.48	74.49	96.07
Building-A-Nets	80.14	**96.91**	**79.31**	97.06	72.77	96.99	74.55	93.52	75.71	98.09	78.73	**96.71**
E-D-Net	**81.85**	94.78	78.46	**98.23**	**77.64**	98.01	73.76	93.25	**79.89**	**98.68**	**79.78**	96.66

Note: The best result in each column is in boldface.

TABLE III
NUMERICAL EVALUATION ON ISPRS TEST SET

Method	Iou	Acc.	ESI	Params (M)	Time(h)
LinkNet	83.0	90.7	76.4	**22.00**	**0.4**
U-Net	85.4	91.4	78.4	26.97	0.6
DeepLabV3	87.4	93.3	83.67	56.72	0.8
E-D-Net	**88.1**	**94.4**	**87.30**	54.42	1.2

Note: The best result in each column is in boldface.

between them is necessary, and different training strategies will have a great impact on the performance of the model.

With the best results indicated in boldface, Tables I and II show the performance of the building segmentation in the test set and validation set of the INRIA aerial image labeling dataset, respectively. From these two tables, we can see that in most cases the performance of the proposed E-D-Net outperformes the other methods, including AMML [32], the winners of the IAIL competition 2018. By solving the problem of producing predictions with poor boundaries, E-D-Net outperforms other state-of-the-art methods in overall IoU, accuracy, and ESI.

Table III shows the performance of the building segmentation in the test set of ISPRS Vaihingen 2-D semantic labeling dataset, our model also outperforms other methods, including DeepLabV3 [36].

However, E-D-Net is a combination of two subnetworks. It has double number of network parameters and learning capacity, which leads to its worse parameters and training time than other methods. Tables I and III show the number of network parameters and the time to complete one training epoch of each method.

Several visual examples are shown in Figs. 5 and 6. From those figures, we can see that E-D-Net can better distinguish boundaries between buildings, compared to the other methods. The prediction results of E-D-Net are almost always regular geometry and rarely have the loss of edge information. It not only captures more "sharp" details, but also has much less semantic errors.

C. Ablation Experiments

In this section, we analyze and verify the effectiveness of each component of E-D-Net by conducting experiments with different ablation settings in Table IV. All experiments are performed on the validation set of the INRIA aerial image labeling dataset.

1) E-Net Architecture: Table IV(a) shows the performance of our network with various E-Net backbones. We compare our designed E-Net (with additional dilated convolutional layers) with other advanced networks. From this table, we can see that different frameworks have a tremendous impact on the results and the improved E-Net backbone can get better results. It benefits from the innovative design of U-Net and the powerful performance of the dilated convolution.

2) D-Net Architecture: Table IV(b) shows the performance of our network with various D-Net backbones. We have compared U-Net with other networks, such as DeeplabV3, LinkNet, and FCN(8 s). The network structures of LinkNet and FCN(8 s) are simpler than U-Net. The results are shown in Table IV(b). From this table, we can see that in some cases DeeplabV3 and FCN(8 s) has better results than U-Net. However, DeeplabV3 and FCN(8 s) have a large number of parameters to train. To make a tradeoff between training time and segmentation performance, we choose U-Net as the backbone of D-Net.

3) End-to-End Versus No End-to-End: E-D-Net uses end-to-end training to obtain the segmentation results. In training stage, we train E-Net first, then D-Net. And the model is initialized

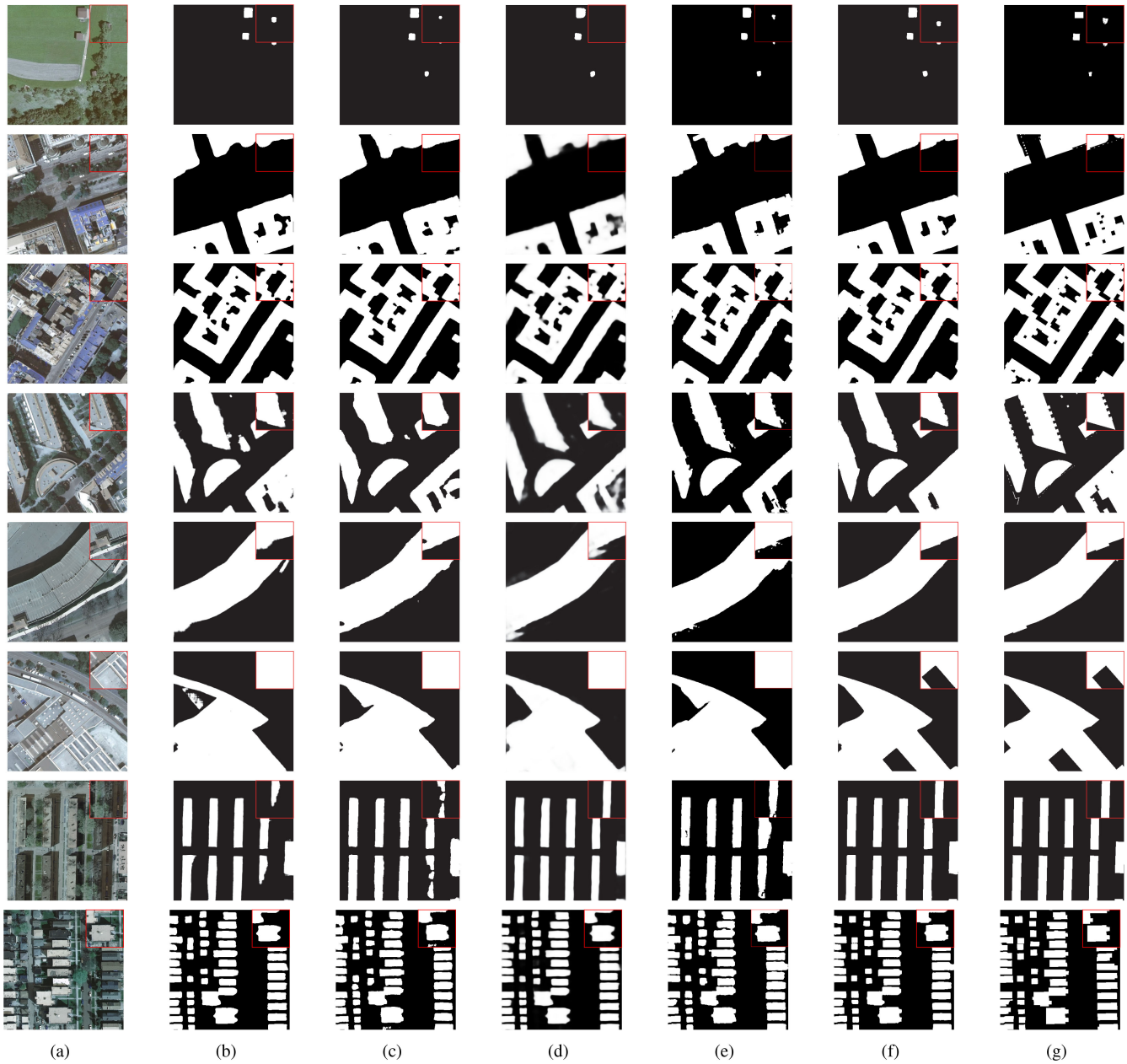


Fig. 5. Classified fragments of the aerial image labeling dataset validation image. (a) Color image. (b) U-Net. (c) LinkNet. (d) Multitask. (e) Building-A-Net. (f) E-D-Net. (g) Ground Truth.

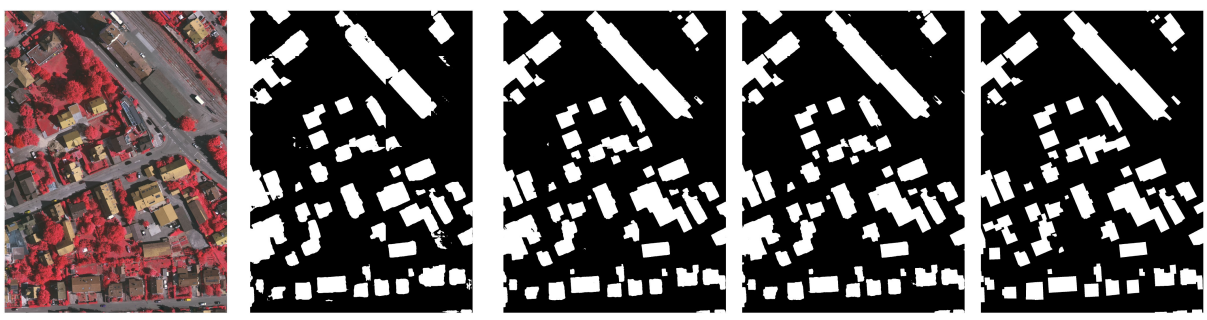


Fig. 6. Classified fragments of the ISPRS image labeling dataset test image. (a) Color image. (b) U-Net. (c) DeepLabV3. (d) E-D-Net. (e) Ground Truth.

TABLE IV
ABLATION EXPERIMENTS ON INRIA TEST SET

E-Net backbone	Austin		Chicago		KITSAP Co.		West Tyool		Vienna		Overall		
	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Params (M)
U-Net	79.32	94.98	75.66	**98.84**	75.32	96.96	68.32	89.84	75.64	95.72	76.33	95.72	26.97
PSPNet-50	81.33	94.21	**79.03**	97.86	**78.83**	**98.81**	71.23	90.35	79.33	96.81	78.21	95.68	62.43
LinkNet	81.02	**96.93**	78.53	97.96	77.52	96.63	72.06	86.77	78.02	96.75	77.56	95.01	**22.00**
U-Net(with dilated layers)	**82.05**	95.02	78.96	98.55	77.76	98.05	**72.76**	**93.45**	**79.76**	**98.66**	**79.98**	**96.77**	28.32

(a)

D-Net backbone	Austin		Chicago		KITSAP Co.		West Tyool		Vienna		Overall		
	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Params (M)
FCN(8s)	82.01	94.54	78.39	98.08	**77.89**	**98.52**	72.47	93.22	79.46	98.54	79.33	96.54	57.00
DeeplabV3	**82.24**	**96.04**	78.93	98.32	78.02	98.33	**72.76**	**93.45**	**79.97**	98.42	**80.24**	**97.22**	56.72
LinkNet	81.88	94.33	77.54	96.32	77.33	97.65	72.04	92.89	78.99	97.32	79.54	95.33	**22.00**
U-Net	82.05	95.02	**78.96**	**98.55**	77.76	98.05	72.73	93.12	79.76	**98.66**	79.98	96.77	26.97

(b)

	Austin		Chicago		KITSAP Co.		West Tyool		Vienna		Overall	
	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.
No End-to-End	78.34	94.32	73.55	96.37	72.35	98.32	68.33	93.02	75.53	97.53	74.37	96.03
End-to-End	82.05	95.02	78.96	98.55	77.76	98.05	72.76	93.45	79.76	98.66	79.98	96.77
	+3.71	+0.70	+5.41	+2.18	+5.41	-0.27	+4.43	+0.43	+4.23	+1.13	+5.61	+0.74

(c)

	Austin		Chicago		KITSAP Co.		West Tyool		Vienna		Overall	
	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.
The E-Net output(foreground)	72.04	93.32	69.39	96.56	68.19	95.02	64.66	90.03	67.14	93.23	66.12	95.78
The E-Net output(foreground + edges)	75.32	94.22	72.55	95.65	72.77	96.03	68.54	91.01	71.84	94.73	72.45	96.68
The D-Net output	80.03	95.58	76.33	98.32	76.02	98.32	70.78	92.98	78.22	98.22	76.76	95.08
The final result	82.05	95.02	78.96	98.55	77.76	98.05	72.76	93.45	79.76	98.66	79.98	96.77

(d)

	Austin		Chicago		KITSAP Co.		West Tyool		Vienna		Overall	
	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.
No L	80.25	94.68	77.32	98.52	76.33	97.58	71.32	93.03	78.88	97.32	77.39	96.58
L	82.05	95.02	78.96	98.55	77.76	98.05	72.76	93.45	79.76	98.66	79.98	96.77
	+1.80	+0.34	+1.64	+0.03	+1.43	+0.47	+1.44	+0.42	+0.88	+1.34	+2.59	+0.19

(e)

	Austin		Chicago		KITSAP Co.		West Tyool		Vienna		Overall		
	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Iou	Acc.	Time(h)	Params(M)
Edge map	82.32	95.32	77.35	98.32	77.35	98.32	72.33	93.02	79.53	98.53	79.77	96.53	54.41	3.22
Fusion map	82.05	95.02	78.96	98.55	77.76	98.05	72.76	93.45	79.76	98.66	79.98	96.77	54.42	3.24

(f)

Note: The best result in each column is in boldface.
(a) **E-Net Architecture:** Better backbones bring expected gains. The U-Net with dilated layers outperforms U-Net, PSPNet-50, and LinkNet.
(b) **D-Net Archutecture:** Comparison of the different D-Net backbones.
(c) **End-to-end versus No End-to-end:** The end-to-end training approach gives better results. It brings a big boost to the model.
(d) **The output of each component versus The final result:** Comparison of the output by E-Net and D-Net with the final result.
(e) L **versus No** L**:** The implicit constraint L brings a certain improvement effect to the model.
(f) **The input of D-Net:** The comparison between the results obtained by using only edge map as the input of D-Net and the results obtained by using fused map as the input of D-Net.

by pretrained E-Net and D-Net. No end-to-end means that we directly use the fusion of two pretraining models as the output of the whole framework (1). End-to-end means that we will use the composite loss function (2) to train the whole model. To verify the effectiveness of end-to-end training, we compare end-to-end training and no end-to-end training approach and list the results in Table IV(c). From this table, we can see that the end-to-end training is more effective than no end-to-end training.

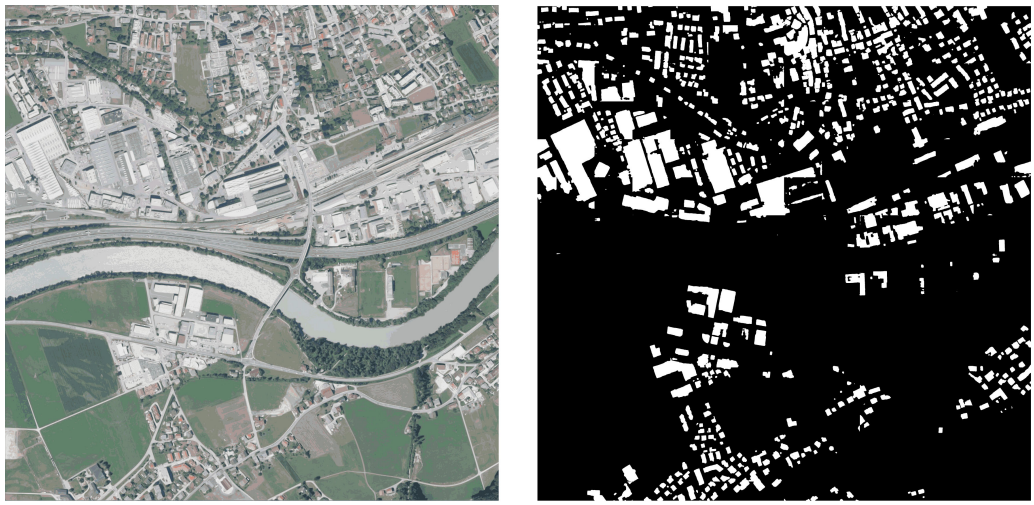
4) Output of Each Component Versus the Final Result: To verify the role of D-Net, we have added ablation experiments for each component of E-D-Net. The results are shown in Table IV(d). From this table, we can see that compared with the foreground layer of E-Net output and the foreground layer with edge layer of E-Net output, the final output result has much higher IoU and Acc values. Compared with the result of E-Net, the result of D-Net is better, but it is still slightly lower than the final result of fusion strategy.

5) L Versus No L: During the training process, we set an implicit constraint (2) to keep E-Net stable. To verify the effect of this operation, we eliminated this constraint in the ablation experiment, i.e., only use L_P as a loss function. The results are presented in Table IV(e). It can be seen from this table that the impact of L on the entire frame is minimal, but it is also indispensable.

6) Input of D-Net: We take E-Net's three-channel output and original image as the input of D-Net. In fact, we can only use edge mapping and original image as the input of D-Net. From Table IV(f), we can see that different maps for the following network have nearly equal segmentation performance. Moreover, the parameter amount and training time of the model are almost

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

Fig. 7. Intermediate result visualization. (a) Input remote sensing image. (b) Output of the E-Net. (c) Result after fusion. (d) Input image with the superimposed predict mask.



(a) (b)

Fig. 8. Example of E-D-Net's segmentation of test set image. (a) Original image. (b) E-D-Net result.

the same. Considering that six-channel input conveniently fits the output of E-Net, and it does improve the accuracy of the model to a certain extent, we employ six-channel as the input of the following network.

D. Visualization of Results

To better understand our approach, we present the visualization results in Figs. 7 and 8.

As shown in Fig. 7(b), we can see that E-Net can extract edge information and texture structure from aerial images. In this figure, the green pixels represent the edge information, the red pixels represent the background, and the blue pixels represent the foreground. D-Net combines the output of E-Net, emphasizing small targets, and detailed segmentation. Finally, we combine the advantages of E-Net and D-Net and achieve excellent results through our fusion strategy. Fig. 7(d) shows the binary masks with green pixels indicating class membership (buildings).

Fig. 8 shows E-D-Net's prediction of a 5000×5000 image from the INRIA aerial image labeling test set. The entire image is cut by 384×384 patches and fed into the network, the output patches are then assembled into tiles of the original size of the entire image. We can see that the proposed E-D-Net can extract the buildings correctly.

IV. Conclusion

Accurate and automatic building segmentation from remote sensing imagery is essential for application areas such as urban planning and disaster management. In this article, a new method (E-D-Net) for extracting buildings in aerial imagery acquired over urban areas is proposed. Considering the existing FCN-based methods have many limitations, such as tend to produce predictions with poor boundaries. In this article, we address the problem of preserving semantic segmentation boundaries in high-resolution aerial imagery by introducing a new cascaded network. The significant contributions of this article can be concluded as follows.

1) By saving the edge information through the E-Net, E-D-Net can alleviate the problem of losing detailed information.

2) By feeding more features to D-Net, D-Net can restore finer texture details.
3) By fusing the output of two components, E-D-Net alleviates the poor boundaries problem to some extent.

The ablation experiments demonstrate the effectiveness of the E-D-Net. In this case, our method achieves outstanding performance consistently on the INRIA aerial image labeling dataset and ISPRS Vaihingen 2-D semantic labeling dataset. Those findings show the practicality of the E-D-Net and its ability to perform effective building segmentation from aerial images.

However, E-D-Net still has the problems of high memory occupation and more time-consuming training compared to other models (e.g., LinkNet). We plan to investigate the strategies for model compression, which can be used to reduce memory usage. In addition, some training strategies aiming to reduce the time consumption will also be investigated.

ACKNOWLEDGMENT

The authors would like to thank @INRIA and @ISPRS for kindly providing the image labeling datasets.

REFERENCES

[1] J. R. Jensen and D. Cowen, "Remote sensing of urban/suburban infrastructure and socio-economic attributes," *Photogramm. Eng. Remote Sens.*, vol. 65, no. 5, pp. 153–163, 2011.
[2] H. L. Yang, J. Yuan, D. Lunga, M. Laverdiere, A. N. Rose, and B. L. Bhaduri, "Building extraction at scale using convolutional neural network: Mapping of the united states," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2600–2614, Aug. 2018.
[3] W. Boonpook, Y. Tan, Y. Ye, P. Torteeka, K. Torsri, and S. Dong, "A deep learning approach on building detection from unmanned aerial vehicle-based images in riverbank monitoring," *Sensors*, vol. 18, no. 11, 2018, Art. no. 3921.
[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[5] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.
[6] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.
[7] I. Sevo and A. Avramovic, "Convolutional neural network based automatic object detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 5, pp. 740–744, May 2016.
[8] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 210–223.
[9] T. Zuo, J. Feng, and X. Chen, "HF-FCN: Hierarchically fused fully convolutional network for robust building extraction," in *Proc. Asian Conf. Comput. Vis.*, 2016, pp. 291–302.
[10] S. Saito, T. Yamashita, and Y. Aoki, "Multiple object extraction from aerial imagery with convolutional neural networks," *Electron. Imag.*, vol. 60, no. 10, pp. 1–9, 2016.
[11] Z. Zhong, J. Li, W. Cui, and H. Jiang, "Fully convolutional networks for building and road extraction: Preliminary results," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 1591–1594.
[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
[13] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3193–3202.
[14] X. Jiang, X. Zhang, Q. Xin, X. Xi, and P. Zhang, "Arbitrary-shaped building boundary-aware detection with pixel aggregation network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, to be published, doi: 10.1109/JSTARS.2020.3017934.
[15] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 158–172, 2018.
[16] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Conf. Image Process.*, 2019, pp. 1480–1484.
[17] J. Tu, F. Gao, J. Sun, A. Hussain, and H. Zhou, "Airport detection in SAR images via salient line segment detector and edge-oriented region growing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 314–326, Nov. 2020, doi: 10.1109/JSTARS.2020.3036052.
[18] S. Chen, W. Shi, M. Zhou, M. Zhang, and P. Chen, "Automatic building extraction via adaptive iterative segmentation with lidar data and high spatial resolution imagery fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2081–2095, May. 2020, doi: 10.1109/JSTARS.2020.2992298.
[19] E. Li, S. Xu, W. Meng, and X. Zhang, "Building extraction from remotely sensed images by integrating saliency cue," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 3, pp. 906–919, Mar. 2017.
[20] Y. Xie *et al.*, "Refined extraction of building outlines from high-resolution remote sensing imagery based on a multifeature convolutional neural network and morphological filtering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1842–1855, Apr. 2020.
[21] V. Iglovikov and A. Shvets, "Ternausnet: U-net with VGG11 encoder pre-trained on imagenet for image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit (CVPR)*, 2018.
[22] D. H. Lee, K. M. Lee, and S. U. Lee, "Fusion of lidar and imagery for reliable building extraction," *Photogrammetric Eng. Remote Sens.*, vol. 74, no. 2, pp. 215–225, 2008.
[23] Z. Deng *et al.*, "R-net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 684–690.
[24] C. H. Li and C. Lee, "Minimum cross entropy thresholding," *Pattern Recognit.*, vol. 26, no. 4, pp. 617–625, 1993.
[25] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016.
[26] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[27] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Can semantic labeling methods generalize to any city? The Inria aerial image labeling benchmark," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2017, pp. 3226–3229.
[28] M. Gerke, *Use of the Stair Vision Library Within the Isprs 2D Semantic Labeling Benchmark*. Berlin, Germany: ResearcheGate, 2014.
[29] P. Soille, *Morphological Image Analysis: Principles and Applications*. Berlin, Germany: Springer, 2013.
[30] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6409–6418.
[31] W. Zhang, F. Liu, and L. Jiao, "SAR image despeckling via bilateral filtering," *Electron. Lett.*, vol. 45, no. 15, pp. 781–783, 2009.
[32] B. Huang *et al.*, "Large-scale semantic classification: Outcome of the first year of Inria aerial image labeling benchmark," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2018, pp. 6947–6950.
[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.
[34] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[35] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.
[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[37] L. Xiang, Y. Xiaojing, and F. Yi, "Building-A-Nets: Robust building extraction from high-resolution remote sensing images with adversarial networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 10, pp. 3680–3687, Oct. 2018.

[38] B. Huang, D. Reichman, L. Collins, K. Bradbury, and J. Malof, "Sampling training images from a uniform grid improves the performance and learning speed of deep convolutional segmentation networks on large aerial imagery," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2018.

Yuting Zhu received the B.S. degree in electronic science and technology from the School of Hainan University, Hainan, China, in 2015. He is currently working toward the Ph.D. degree in electronic and communication engineering with Sun Yat sen University, Shenzheng, China.

His remote sensing research interests include building extraction, change detection, and deep learning.

Zili Liang received the B.E. degree in biomedical engineering from the School of Information Engineering, Guangdong Pharmaceutical University, Canton, China, in 2020. He is currently working toward the master's degree in electronic information with Shantou University, Shantou, China.

His research interests include image segmentation and deep learning.

Jingwen Yan received the M.S. degree in cartography and remote sensing from Changchun Institute of Geography, Chinease Academy of Sciences, Changchun, China, in 1992, and the Ph.D degree in optics from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 1997.

He is currently a Professor with the Electronic Information, College of Engineering, and a Deputy Director with the Key Lab of Digital Signal and Image Processing of Guangdong Province, Shantou University. His research interests include wavelet analysis and application, compressed sensing, signal sparse representation, and remote sensing image processing.

Gao Chen received the master's degree in control theory and control engineering from Xiamen University, Xiamen, China, in 2009, and the Ph.D. degree in communications and information systems from Southwest Jiaotong University, Chengdu, China, in 2016.

From October 2016 to October 2018, he was a Postdoctoral Researcher with the Department of Electronic Engineering, Tsinghua University, Beijing, China. He is currently a Lecturer with the School of Electrical Engineering and Intelligentization, Dongguan University of Technology, Dongguan, China. His current research interests include remote sensing image processing and machine learning.

Xiaoqing Wang received the B.S. degree in electronics engineering from Xiamen University, Xiamen, China, in 2000 and the Ph.D. degree in communication and information system from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2005.

He is currently a Professor with the School of Electronics and Communication Engineering, Sun Yat-sen University, Guangzhou, China. His research is mainly focused on SAR ocean remote sensing, SAR agriculture remote sensing, and SAR image processing.