Fine Building Segmentation in High-Resolution SAR Images Via Selective Pyramid Dilated Network

Hao Jing, Xian Sun, Zhirui Wang, Kaiqiang Chen, Wenhui Diao, and Kun Fu

Abstract—**The building extraction from synthetic aperture radar (SAR) images has always been a challenging research topic. Recently, the deep convolution neural network brings excellent improvements in SAR segmentation. The fully convolutional network and other variants are widely transferred to the SAR studies because of their high precision in optical images. They are still limited by their processing in terms of the geometric distortion of buildings, the variability of building structures, and scattering interference between adjacent targets in the SAR images. In this article, a unified framework called selective spatial pyramid dilated (SSPD) network is proposed for the fine building segmentation in SAR images. First, we propose a novel encoder–decoder structure for the fine building feature reconstruction. The enhanced encoder and the dual-stage decoder, composed of the CBM and the SSPD module, extract and recover the crucial multiscale information better. Second, we design the multilayer SSPD module based on the selective spatial attention. The multiscale building information with different attention on multiple branches is combined, optimized, and adaptively selected for adaptive filtering and extracting features of complex multiscale building targets in SAR images. Third, according to the building features and SAR imaging mechanism, a new loss function called L-shape weighting loss (LWloss) is proposed to heighten the attention on the L-shape footprint characteristics of the buildings and reduce the missing detection of line buildings. Besides, LWloss can also alleviate the class imbalance problem in the optimization stage. Finally, the experiments on a large-scene SAR image dataset demonstrate the effectiveness of the proposed method and verify its superiority over other approaches, such as the region-based Markov random field, U-net, and DeepLabv3+.**

Index Terms—**Automatic fine segmentation of buildings, L-shape weighting loss (LWloss), selective spatial pyramid dilated (SSPD) network, synthetic aperture radar (SAR).**

I. Introduction

THE building is a significant topographic object class in the city and a momentous data layer in the geographic information system. Building segmentation in geographic remote sensing images plays a vital role in the geographic information system application, which is also a challenging question of great interest in remote sensing. Automatic extraction of buildings from aerial remote sensing images is frequently used for surveying and mapping of ground objects, detection of illegal buildings, urban ecological planning, and regional development.

The synthetic aperture radar (SAR) images are obtained from all-day and all-weather remote sensing sensors free of the atmosphere variation. A great number of advanced works have been launched depending on the superiority of SAR [1]–[7]. The apparent scattering features of buildings theoretically ensure the good extraction effect in SAR images. Recent developments in building segmentation have heightened the need for fine extraction. However, some works [8], [9] are limited by low-resolution images, resulting in the chaotic extraction effects. As the technology develops, the resolution of the SAR images is higher, and the richer details emerge. To obtain the precise boundary, positions, and scales of the buildings, it is of great need to fulfill the fine segmentation of buildings based on the high-resolution SAR images.

Generally, before the large-scale application of deep learning, most of the building extraction for SAR images adopt the methods of designing features and establishing statistical models, such as the gray level cooccurrence matrix (GLCM) method and the Markov random field (MRF) model. The existing researches suggest that the conventional methods are fast and straightforward to be implemented without the large datasets. Nevertheless, for areas with substantial feature changes, the effect of extracting buildings is coarse. The higher level semantic information is not expressed, which cannot be adapted to the increasingly changeable complex SAR scenes. Recently, benefited from the wide application of deep convolutional neural networks in remote sensing [10]–[15], most deep learning-based methods have been applied to enhance the accuracy and efficiency of extracting buildings in SAR images [16], [17]. They form an end-to-end approach that raises the level of the feature extraction compared to the manual design. The fully convolutional network (FCN) models or their plain variants are generally relied on to extract buildings' features in complex SAR scenes. FCNs receive images of any size and finally output the classification score map of the same size by extracting features through several convolution layers and fusing multiple feature maps. Simply transplanting FCNs to the SAR building extraction leads to insufficient learning ability in accurately determining the shape, size, and location of buildings. Meanwhile, the severe multiscale characteristic problem can affect the actual extraction capability due to the complex diversity of building structures. Recent trends in spatial pyramid have led to a proliferation of studies [18]–[20]

Manuscript received September 20, 2020; revised January 26, 2021; accepted April 24, 2021. Date of publication April 27, 2021; date of current version July 14, 2021. This work was supported in part by the National Science Foundation of China (NSFC) under Grant 61725105 and Grant 62076241. *(Corresponding author: Xian Sun.)*

Hao Jing, Xian Sun, and Kun Fu are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jinghao17@mails.ucas.edu.cn; sunxian@mail.ie.ac.cn; fukun@mail.ie.ac.cn).

Zhirui Wang, Kaiqiang Chen, and Wenhui Diao are with the Key Laboratory of Network Information System Technology (NIST), Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China (e-mail: zhirui1990@126.com; chenkaiqiang14@mails.ucas.ac.cn; dwh1031@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2021.3076085

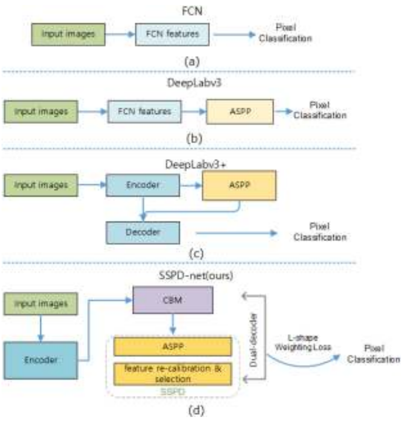


Fig. 1. (a) Scheme of building extraction in SAR images with FCNs. In previous feature extraction networks, only convolutional layers are used to extract features. (b) Spatial pyramid structures such as atrous spatial pyramid pooling (ASPP) [25], are widely used to solve the multiscale characteristic problem in FCN improvement schemes. (c) Encoder and decoder structures for segmentation become a general network. (d) Our method, the spatial pyramid with adaptive selection and the LWloss strategy are added. The proposed structure is organized with a novel encoder–decoder based on the dual-stage decoder. CBM: Context balancing module. SSPD: Selective spatial pyramid dilated net.

that the spatial pyramid structure is rapidly becoming a key instrument in the multiscale characteristic problem. However, the multiscale building extraction performs poorly with the fragments of large buildings and the absences of tiny buildings under the actual large remote sensing scenes. It is accepted that the critical calibration and the selection of multiscale features are imperative for mitigating the multiscale characteristic problem.

The past years have seen the rapid development of deep learning in a wide range of fields. Plenty of deep learning methods are simply transferred to the building extraction [21]–[24] in the SAR images. As we all know, the double bounce scattering formed by the grounds and the walls is considered a major indicator of buildings in high-resolution SAR images. L-shape features are major areas of interest within the field of building extraction. In the achievements for the FCN and its variants, the L-shape features are missing more seriously, which results in the rough and mixed building boundaries and the high missing detection for some line objectives of buildings. In addition, the problem of data imbalance in the process of building extraction is often ignored. Generally, the building class pixels are much less than the background class pixels. The cross-entropy loss function used in the traditional training process can easily cause that the background pixels occupy the dominant position, which makes the attention of the network training shift to the dominant pixels under the condition of data imbalance. In this case, the model generalization ability is reduced and overfitting. In light of the deficiencies in the above methods, it is becoming extremely difficult to develop the fine building segmentation.

Aiming at these issues and considering the building features and mechanism in SAR images, we propose a unified framework called selective spatial pyramid dilated (SSPD) net for the fine building segmentation in SAR images. As shown in Fig. 1(d) specifically, we improve the multiscale context fusion and reconstruction by the instrumental SSPD module and the novel dual-decoder. The L-shape weighting loss (LWloss) is employed to give more attention to the L-shape footprint elements and their nearby elements. We demonstrate the effectiveness of our model in the fine building segmentation on a Gaofen-3 satellite SAR dataset, and achieve the 91.2% accuracy performance on the test set without any postprocessing.

In brief, our principal contributions are summarized as follows.

1) For the multiscale characteristic issue, we design a multilayer SSPD module combining the channel selection and the branch selection, which offers a comprehensive feature representation of adaptive nonlinear aggregation. The SSPD module enhances the spatial pyramid's multiscale feature selection and reconstructs the spatial feature relationship, which promotes the adaptive fine extraction of the SAR buildings.
2) A novel encoder–decoder structure is proposed based on the dual-stage decoder. The context balancing module (CBM) and the SSPD module are involved in the multilevel semantic information fusion and construction, which is conducive to the restoration of the complete shape and location of buildings.
3) The newly LWloss function is designed to focus on the ignored L-shape footprint and line objectives. Compared with the cross-entropy loss function that treats all the pixels equally, our loss function gives higher weight to the more crucial L-shape footprint pixels with dynamically distance adjustment. It also alleviates the class imbalance problem.
4) Compared with other optical transplanted models, our proposed unified framework focuses on the characteristic SAR building features and amelioration that are not valued in ordinary networks. It is more competitive for fine segmentation of variable-scale buildings in large SAR scenes, whether in terms of visual effects or quantitative metrics.

The experiments on a large-scene SAR image dataset indicate that SSPD-net achieves the better building extraction accuracy and visual effects compared with the popular semantic segmentation methods and the conventional methods, which proves the advantages of the proposed method.

The remainder of our work is organized as follows. In Section II, we briefly illustrate the related tasks, including feature-based methods, model-based methods, and deep convolutional neural methods in building segmentation. Next, we pay attention to the proposed framework in Section III, including the SSPD module, the dual-stage decoder, and the LWloss. The details and conclusions of the experiments are discussed in Section IV and Section V. Finally, Section VI concludes this article.

II. Related Work

Since our work mainly refers to the fine building segmentation in SAR images, we briefly review the related studies in this field.

It has previously been observed that the conventional building extraction approaches for SAR images can be divided into two categories, i.e., the feature-based and the model-based. The feature-based group is to extract the brightness, texture, border, and mixing characteristics from the SAR images. The Fourier power spectrum [26], the Gabor filter analysis [27], the MRF model texture description [28], [29], and the GLCM texture measure [30], [31] are commonly used. In [31], the GLCM is applied to extract texture features for building area segmentation in SAR images. To obtain the texture images similarly, a set of heuristic Gabor filter sets [27] are designed. The scale-invariant feature transform (SIFT) algorithm for SAR images (SAR-SIFT) [32] is proposed to solve the image registration problems with different incident angles. These features are frequently combined with methods such as the nonsupervised cluster analysis. Commonly, the complicated objects in SAR images cannot be fully described by the low-level features, and an enormous gap between them and the high-level semantic expressions exists. Furthermore, the low-level features only represent the pixel-level information, which is greatly affected by the multiplicative noise.

On the other hand, the model-based category is to establish the statistical distribution model of the SAR images, which combines the spatial background information for segmentation, including MRF [29], Fisher distribution [33], logarithmic normal distribution [34], and generalized Gaussian distribution [35]. In the early k-means [36] and gamma models, the feature spatial representation is only considered, but the spatial interaction is not taken into account. The primary spatial context constraint works in the region-based MRF model [37], but it will lead to oversegmentation, especially in building areas. Tison *et al.* [33] employ Fisher distribution to model the ground objects of SAR images (especially for buildings) and combine the statistical model with MRF to realize the classification of ground objects. However, the detailed information is constantly lost in the results, and some attached blocks emerge. Xia *et al.* [38] propose an MRF model on region adjacent graph (MRF-RAG) to correct the edge error resulting from the oversegmentation algorithm. Plenty of semantic inconsistencies in the building areas exist yet. When it comes to the highresolution SAR images, these models can no longer accurately describe the detailed and bright spot texture structure exhibited by the buildings. Moreover, the predefined statistical distribution model may not be suitable for broad regions with different characteristics.

In addition to designing the features and the models, how to design a good classifier is also the focus. The classification methods are summarily divided into three categories: Unsupervised learning, semisupervised learning, and supervised learning. There exists no labeled samples in unsupervised learning with the goal of inferring the internal structure in a set of data. The common algorithms for SAR building segmentation or other object classification consist of k-means clustering and principal component analysis, etc. There is a general lack of a current method to compare the performance of the algorithms. The semisupervised learning understands the remaining large amount of unlabeled data by learning a small amount of labeled data. Many commonly used semisupervised methods have been applied to SAR object classification, such as transductive support vector machine, graph-based anchor graph regularization [39], and squared-loss mutual information regularization for multiclass probabilistic classification based on manifold assumption [40]. In the case of very small labeled data, Protopapadakis *et al.* [41] use the semisupervised learning approaches as the loss function throughout the training of neural networks, which can be beneficial to pixel level segmentation tasks on a limited dataset. Compared with the semisupervised learning, the supervised approaches applies sufficient labeled samples to fit the relationship between input and output. The logistic regression, naive Bayes, support vector machine, random forest, neural network, and other methods are also drawn into the SAR image segmentation. Significantly, the dimension reduction can make the data features dense and eliminate data redundancy when faced with excessive input dimensions. Makantasis *et al.* [42] propose tensor-based linear and nonlinear models for hyperspectral image classification, which is also an available solution.

In recent years, deep neural networks are gradually being applied to the image interpretation of natural scenes and remote sensing scenes. The deep learning methods have gradually replaced the traditional SAR segmentation methods, on account of meeting the requirements of the fine extraction of buildings and other objects in high-resolution SAR images. It is worth mentioning that the application of deep learning heightens the ability of the feature extraction and makes the precision of the ground objects extraction significantly improved in the SAR images [16], [43]–[46]. Yao *et al.* [47] successfully apply FCNs to the semantic segmentation in the SAR images and classify the landuse, water, buildings, and natural areas. Although the FCN can accept the input images in any size, lots of spatial information is lost, leading to a coarse segmentation result. Considering the multiscale feature of the SAR images, Duan *et al.* [48] present a multiscale convolutional neural network for the SAR semantic segmentation, and the labeling consistency is obtained in most of the terrains. Nevertheless, the model with shallow network structure merely makes the simple scale transformation of the input information, which leads to insufficient extraction of the practical information. In contrast, as one of the most advanced neural networks, DeepLabv3+ [49] utilizes several parallel atrous convolutions at different rates called atrous spatial pyramid pooling (ASPP) to capture more sufficient context information. Compared with some previous convolution structures, ASPP which has emerged as a powerful tool, can mainly extract multi-scale buildings accurately and efficiently. There is a growing body of researchers that recognize the importance of exploring the spatial dependence [50], [51] and representing the spatial feature correlation with integrating learning mechanism [52]–[54]. In particular, the spatially dependent guidance for the multiscale features performs crucially for the adaptive feature selection in the complex context information and large variations. Besides, the encoder–decoder structure [55], [56] has always been an advanced model in the field of image segmentation, which can extract and restore the features wholly and quickly. In U-net [55], the features are concatenated in the channel layers on the equal level of the encoder and the decoder. This is effective for preserving the semantic information of SAR image extraction, but the feature fusion is still inadequate.

III. Methods

A. Framework

Building segmentation in complex SAR scenes is interfered with other complex backgrounds and multiscale characteristics

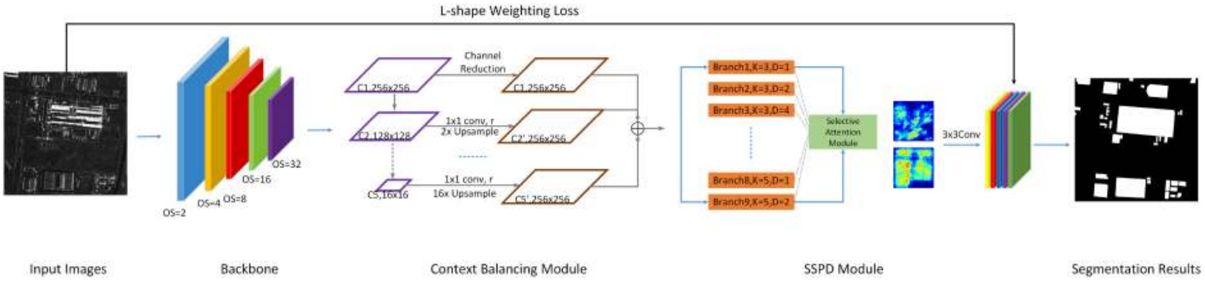


Fig. 2. SSPD-net structure. K: Kernel size. D: Dilation rates. OS: Output stride.

of buildings. Traditional classifiers or artificial features are gradually tired of coping with the high-resolution complex scenes. However, some general advanced neural networks have been verified to be effective for target detection and scene classification of remote sensing scenes, although the utilization and mining of specific features in SAR images are not sufficient. Therefore, our unified framework is designed with popular encoder–decoder structure. The overall structure of SSPD-net for SAR building segmentation is illustrated in Fig. 2, which consists of two parts: The encoder and dual-decoder. As an encoder with moderate parameters, ResNet34 [57] has the outstanding feature extraction ability and computation speed. The convolution unit is based on the residual blocks composed of 3 × 3 kernel convolution layers, which have a low computation cost. The output stride is 32. The output spatial resolution of the last convolution layer of the encoder is 32 times smaller than that of the input image, which contributes to extracting the denser features. To utilize a dual-decoder to recover the characteristics of the target in stages is an innovative approach. The basic SAR features obtained from the backbone network are sent to the dual-decoder consisting of the CBM and the SSPD module. In the Decoder1, the extracted SAR image features by the CBM are comprehensively characterized and reconstructed to close up to the adaptive feature balancing, thus supporting the feature recovery in the Decoder2. A multiscale building attention mechanism is established by the SSPD module, which fully integrates and adaptively selects multiscale building features to recover building details more finely in the Decoder2. To increase the guidance of the buildings' double bounce scattering characteristics to the network, a weight mask is added to the common loss by employing the LWloss. Our method extracts the location and profile of potential buildings in SAR images in the inference stage directly and accurately.

B. Dual-Stage Decoder Based on the CBM and SSPD Module

In DeepLabv3 [25], the decoder is a 16 times upsampling module, which is directly amplified by the last feature map. In this case, the decoder is not very effective in restoring the details of objects. Considering the semantic information contained in different output layers of the encoder, we propose a capable two-stage decoder, as shown in Fig. 2. First, the output feature maps of the five layers are in different scales with the corresponding output stride, 2, 4, 8, 16, and 32, respectively. They are resized to the same spatial resolution with bilinear interpolation upsampling and concatenated, as shown in Fig. 3. We adopt an average compression strategy to balance the high-level and the low-level semantic information. The number of channels per layer is reduced to r by 1 × 1 convolution. The above operation is named as the CBM, while the hyperparameter r and the average compression strategy will be discussed in Section IV-D3. The CBM is equivalent to five skip connections between the encoder and the decoder structures at the same time. We reckon that the CBM is better than the behavior of gradually upsampling from high-level semantic information. It creates more information paragraphs that preserve the high-frequency information in the smoothing process.

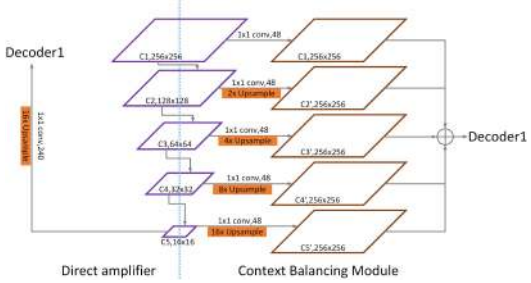


Fig. 3. Direct amplifier and the CBM structure of the Decoder1.

The second stage of the decoder is the SSPD module, which can further extract and refine the multiscale semantic information. In DeepLabv3 [25], ASPP uses the atrous convolution at different rates for multiscale probing features and aggregates contextual information. It shows that the features extracted at each rate are processed in a separate branch and then are merged to generate the final result. However, the fusion lacks the guiding information, which leads to the independent process of resampling the features extracted from each scale. Incorporating the spatial and the channel attention into the network is necessary to enhance the multiscale feature selection and expression. The squeeze and excitation operations are utilized to reconstruct the interdependence and suppression in the feature channels of the spatial pyramid module. As Fig. 4 exhibits, the branches of the spatial pyramid module are amplified to nine, and their receptive fields are nested. With the selection operations, the spatial pyramid branches are recalibrated following the multiscale information of the targets. And all the ordinary convolutions are displaced by the dilated depthwise separable convolutions [58]. Finally, the two 3 × 3 convolutions are employed to refine the features. These operations (in the particular SSPD module) play a crucial role in the fine segmentation of SAR buildings.

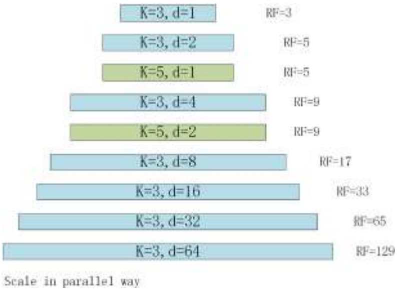


Fig. 4. Illustration of the receptive fields with respect to different scale diversity for different branches of spatial pyramid dilated convolution. K: Kernel size. D: Dilation rates. RF: Size of the receptive field of the corresponding branch.

C. Multilayer Spatial Pyramid Dilated Convolution Module Based on Channel Selection and Branch Selection

The SSPD is the central unit, mapping the input $X \in R^{H\times W\times C}$ to feature maps $Y \in R^{H\times W\times C}$. We propose fusion, squeeze, excitation, and selection operations, especially in fusion pyramid modules, which provide global information access and calibration feature responses at the channel level with the appropriate receptive fields. Consequently, the SSPD is a feature refinement network that is sensitive to multiscale information, during which the feature learning of channel selection and branch selection constraints is carried out according to the multiscale information of the target.

1) Channel and Branch Selection: In Fig. 5, the merged feature map is denoted as $X \in R^{\widetilde{H}\times\widetilde{W}\times\widetilde{C}}$ in the Decoder1. Each parallel dilated convolution layer is regarded as a unit d, and all the pyramid convolution units are combined as a super module $D = [d_1, d_2, \ldots d_9]$. D has the kernel convolution with different sizes and various dilated rates. In order to adjust the size of the receptive field, different branches of D are squeezed after fusion to generate the channel statistics. Then, different branches are finally selected through the established relational model, as shown in Fig. 5.

Fusion: In SSPD, for any input $X \in R^{\widetilde{H}\times\widetilde{W}\times\widetilde{C}}$, we first conduct nine transformations using nine kernel convolution $F_1 - F_9 : X_1 - X_9 \rightarrow U_1 - U_9 \in R^{H\times W\times C}$ with different sizes. The whole F is composed of atrous depthwise separable convolutions, and their kernel convolution size and dilated rates are shown in Fig. 4. The information flow is first merged from the multiple branches. Note that the number of output feature channels per branch is 48. The fusion result of multiple branches is obtained by channelwise summation, as

$$U = U_1 \cup U_2 \cdots \cup U_9. \tag{1}$$

Squeeze: Subsequently, we symbolize each channel layer with a channel descriptor $z \in R^c$. That is, the information in the $H \times W$ spatial dimension on each channel is compressed into a number to generate the channel statistics. As shown in (2), the global average pooling (ap) is used to calculate the cth element of z

$$z_c = F_{ap}(u_c) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{2}$$

where $u_c \in R^{H\times W}$.

Excitation: A concise feature s is created based on two full-connected layers [50]. The relationship between the channels is modeled with the feature s, so as to adaptively establish the expression of the inhibition or the promotion of channels. The expression relationship of the modeling feature s can be used for soft selection of the pyramid branches. The conduct of such modeling is flexible and nonlinear

$$s = F_{fc}(z) = \sigma\left(W_2\delta\left(W_1 z\right)\right) \tag{3}$$

where δ stands for ReLU function [59], $W_1 \in R^{\frac{C}{r}\times C}$, $W_2 \in R^{C\times\frac{C}{r}}$, and σ stands for sigmoid function. The default r value is set as 16 with the purpose of forming a general dimensionality reduction layer in the fully connected (fc) layers.

Selection: The concise feature s obtained by the excitation operation can be considered as a set of mapping channel weights. As shown in (4), it recalibrates the rich semantic information in different scales to obtain the final output $Y(Y = [y_1, y_2 \ldots y_c])$ in the SSPD module, which achieves the selection for the nine branches. The coexistence of inhibition and promotion works in the 48 channels inside the branches, which can also be regarded as the soft self-attention mechanism of convolution response on channels

$$y_c = s_c \cdot u_c. \tag{4}$$

The concise feature s and the feature mapping U are channelwisely multiplied.

In addition to the channelwise selection above, the branch selection can also be performed in another way. In the fusion operation, provided that elementwise is summed, the fusion results of multiple branches will be expressed as follows:

$$\widetilde{U} = U_1 + U_2 \cdots + U_9. \tag{5}$$

Furthermore, $\widetilde{U}$ is squeezed and excited according to the equivalent operation by (2) and (3), respectively

$$\widetilde{z}_c = F_{ap}\left(\widetilde{u}_c\right) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} \widetilde{u}_c(i,j) \tag{6}$$

$$\widetilde{s} = F_{fc}(\widetilde{z}) = \sigma\left(W_2\delta\left(W_1\widetilde{z}\right)\right). \tag{7}$$

The concise feature $\widetilde{s}$ obtained by (6) and (7) guides the selection of multiscale information on different branches, which uses the softmax operation

$$a_i = \frac{e^{\widetilde{s}_i}}{\sum_{i=1}^{i=9} e^{\widetilde{s}_i}}, \sum_i a_i = 1 \tag{8}$$

where a_i represents the branch weight of $\tilde{U}_i$. The final output feature map is weighted by the attention weight on different branches, which is

$$y_c = \sum_{i=1}^{i=9} a_i \widetilde{u}_i. \tag{9}$$

The comparison of the experimental results brought by two different fusions of branches is shown in Section IV-D1.

2) Dilated Depthwise Separable Convolution in SSPD: It is observed that the atrous depthwise separable convolution [60], [61] is applied to the SSPD module. Dilated convolution [62] is a powerful convolution tool with the exponential growth receptive

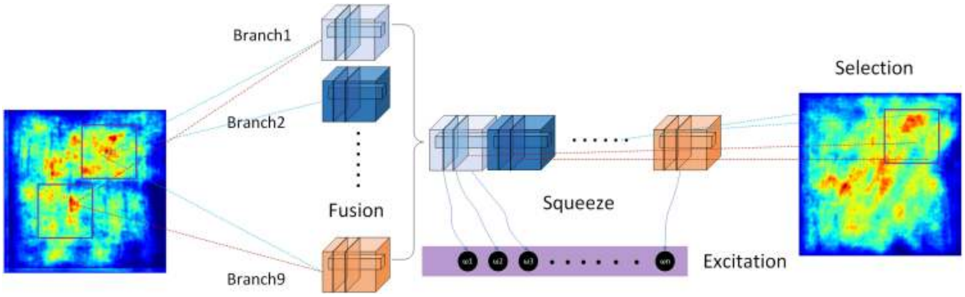


Fig. 5. Multilayer pyramid dilated convolution module based on channel selection and branch selection.

field and the linear increase of parameter number. It can process the input feature maps with a higher precision while the size is maintained by the dilated convolution, and the dense prediction of target details is implemented. For 2-D signals, the dilated convolution is brought to the input feature maps P, and the output feature mappings Q are

$$Q[i] = \sum_{k} P[i + r \cdot k]F[k] \tag{10}$$

where F stands for the convolutional filters, r is the dilated rate, and k is the skip stride of the dilated convolution. As shown in Fig. 4, the combined pyramid convolutional filter fields are obtained. Thus, multiscale context information can be captured for multiscale information for buildings from an adaptive selection of a wider input perspective.

Unlike conventional convolution, the depthwise separable convolution [63]–[65], consisting of the depthwise convolution and the pointwise convolution, has a lower parameter quantity and operation cost. The depthwise convolution performs independent convolutions on each channel of the input layers, and the quantity of output channels is the same as that of the input. And then, the pointwise convolution makes a weighted combination in depth. The computational complexity of the model is significantly reduced by combining the depthwise convolution and the pointwise convolution. The dilated depthwise separable convolution in SSPD can make the model lightweight accordingly. Simultaneously, the high model property is maintained.

D. L-Shape Weighting Loss

The backscattering signal intensity of the buildings is the superposition of the backscattering signals from various parts of the building. Among them, corner reflector, layover, and shadow are important features of SAR buildings. Many segmentation errors in SAR images are resulted by the scattering and imaging mechanism. For instance, large buildings usually appear as strong linear or L-shaped echoes in SAR images due to the strong double bounce reflection toward the direction of radar incidence angle, which means that in some cases, only two edges of the buildings can be clearly observed by the SAR. Nevertheless, the segmentation results of many convolutional models still have some defects, such as boundary deletion and roughness, especially when the large buildings are extracted. According to the different representations of L-shape, we introduce visual interpretation assistance to classify three types of building targets on the Gaofen-3 SAR images. As illustrated in Fig. 6, there are three representations for buildings in SAR images: Surface objectives, line objectives, and complex objectives. Generally, complex objectives have complex structures and high-backscattering intensity. Correspondingly, both surface objectives and line objectives, with the L-shape footprint, have a low proportion in all building backscattering areas. Based on the U-net experiments, the extraction results for three building objectives are shown in Table I.

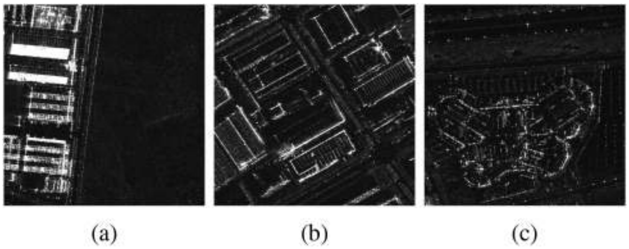


Fig. 6. (a) Surface objectives. (b) Line objectives. (c) Complex objectives.

TABLE I
EXTRACTION EFFECTS OF U-NET IN THREE BUILDING OBJECTIVES

Objectives	OA(%)	IoU(%)	Missing alarm(%)
Surface	83.14	34.53	13.28
Line	80.68	31.75	16.83
Complex	**83.87**	**34.76**	**13.23**
Overall	82.70	33.38	14.75

The bold entities are the highest scores in each comparison.

The accuracy and missing alarm rate of surface and line objectives are worse than that of complex objectives. We hold that the complex objectives cause strong attention in neural network training. Simultaneously, the surface and line objectives are easily ignored in feature learning, resulting in a high-missing alarm rate and the unfinished edge of buildings. A novel loss function is proposed to extract the surface and line targets of SAR buildings effectively. In the process of network training, the L-shape footprint terrain pixels are given a high weight mask to enhance the training attention of the surface and line objectives. First, the edge detection algorithm is applied to the SAR image to detect a strong echo similar to the angular reflector. Then the global Hough transform and the local Hough transform with the moving window are performed to extract the indicator to buildings such as the L-shape and the linear features. Next, the intersections of L-shape scattering are determined by utilizing the orthogonal line structures. And the appropriate intersections

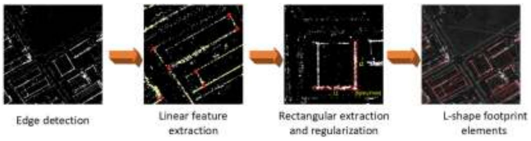

Edge detection
Linear feature extraction
Rectangular extraction and regularization
L-shape footprint elements

Fig. 7. Extraction of the L-shape footprint pixels.

are selected for subregion clipping. For each cropped subregion, the rectangular-package method [66] is applied to extract the L-shape footprint accurately. Finally, the L-shape footprint is mapped to the original image space according to the slope angles θ and the lengths L of the detected lines, and the original location for the subregion. At this point, the set of all the L-shape terrain elements is obtained, as shown in Fig. 7.

The Euclidean distance from this pixel to the set of L-shape footprint terrain pixels is then calculated for each pixel. The final weight is calculated based on the Gaussian weight contributed by all the L-shape footprint pixels to this pixel, as follows:

$$w_p = 1 + w_0 \sum_{q \in M} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_p - x_q)^2 + (y_p - y_q)^2}{2\sigma^2}} \tag{11}$$

where w_0 is the penalty factor for L footprints. M is the collection of L footprint terrain elements and (x_p, y_p) is the coordinates of the pth pixel. σ can affect the size of M region.

Furthermore, the issue of class imbalance is common in building extraction. The cross-entropy loss, generally used in image segmentation, is defined as

$$L = -\sum_{N} y_{\text{true}} \log\left(y_{\text{pred}}\right) \tag{12}$$

where y_{true} represents the true label of each pixel, and y_{pred} represents the prediction probability of each pixel. It evaluates the class prediction for each pixel vector equally. Considering the dominant background pixels guiding the training attention, we add the soft dice coefficient loss [67] to restrain the class imbalance. The final defined loss function is

$$L = -\sum_{N} w_p y_{\text{true}} \log\left(y_{\text{pred}}\right) + w_p - \frac{2 w_p \sum_{N} y_{\text{true}} y_{\text{pred}}}{\sum_{N} y_{\text{true}}^2 + \sum_{N} y_{\text{pred}}^2}. \tag{13}$$

The soft dice coefficient loss is multiplied by the penalty factor separately for each category, and the final result is then averaged to normalize the loss.

IV. Experiments

A. Dataset

Due to the low-resolution and small scales, the data used in the previous related work [47], [48] lack accurate labels, which is not suitable for our fine extraction method. Researchers also lack the publicly available SAR datasets for the building extraction. Hence, to facilitate the research for the fine building extraction on high-resolution SAR images, we build a new dataset to evaluate the effectiveness of the proposed method. We collect the urban images acquired from the Gaofen-3 satellite by the spotlight mode. All the images are single-band and single-polarized. The spatial resolution is 1 m, which ensures the quality of the annotation. The SAR images are labeled referring to the corresponding optical remote sensing images and verified by experts. The positive annotations are buildings, and other pixels are the background class. The images are cropped into 512 × 512 pixels with a total of 279. 80% images of the dataset are used for training and the rest are for testing. The portion of the dataset is shown in Fig. 8.

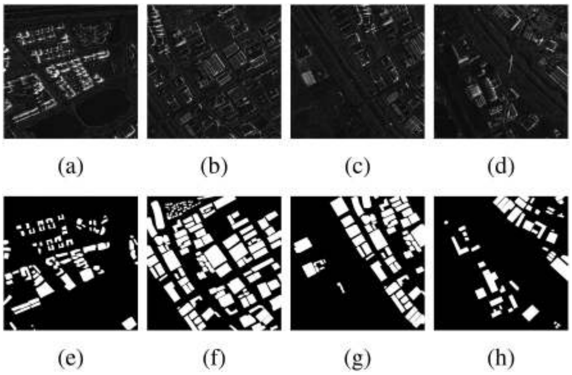


Fig. 8. Partial visualization of high-resolution SAR images of GaoFen-3 satellite. The typicality of these groups of pictures is that small and large buildings exist concurrently, and their respective dense and sparse states exist. (a)–(d) SAR images. (e)–(h) Corresponding ground truth.

B. Implementation

The proposed network is operated on the NVIDIA p100 GPU based on PyTorch [68]. The amount of training data is expanded to 1674 by the image morphological transformation, including random horizontal and vertical folding, rotating, arbitrary scaling, random migration, and accidental lifting. These operations effectively increase the amount of data and weaken the tendency of overfitting caused by insufficient raw data. The expanded slices are disorganized and randomly fed into the network. The size of each batch is 5 for 250 epochs. The pretraining ResNet34 on ImageNet [57] is adapted to the encoder. The initialization weight of the decoder follows the default uniform distribution of PyTorch [68]. The momentum is 0. The adaptive learning rate optimization algorithm named Adam [69] is applied to train the network. The betas are (0.9, 0.999), and the weight decay is 0. The initialization learning rate is 0.001. When the training loss is stable, the learning rate is reduced five times manually. Our model adopts end-to-end training.

C. Evaluating Metrics

The experimental results are evaluated based on several widely used indicators, as follows:

$$\text{Overall Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
$$\left(\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}\right) \tag{15}$$

TABLE II
ABLATION RESULTS OF DUAL-STAGE DECODER

Decoder1		TT(s)	Params(M)	OA(%)	F1(%)	mIoU(%)	fwIoU(%)
DA	CBM						
✓	×	**35**	**52.6**	88.08	78.66	67.11	78.67
×	✓	38	54.1	**90.28**	**83.16**	**72.70**	**82.27**

TT denotes the training time.

$$\begin{aligned}\text{mIoU} = &\frac{1}{2} \times \frac{TP}{TP + FP + FN}\\ &+ \frac{1}{2} \times \frac{TN}{TN + FP + FN}\end{aligned} \tag{16}$$

$$\begin{aligned}\text{fwIoU} = &P(P) \cdot \frac{TP}{TP + FP + FN}\\ &+ P(N) \cdot \frac{TN}{TN + FP + FN}\end{aligned} \tag{17}$$

where TP refers to all the correctly classified building pixels. TN is all the correctly classified background pixels. FP denotes all the building pixels that do not have the correct classification, and FN represents all the background pixels that do not have correct classification. OA indicates the ratio of all the correctly classified buildings and background pixels to all the classified pixels. F1 and mIoU can reflect and evaluate the effect based on the above metrics. FwIoU sets the weight in line with the frequency of the building areas and the background areas on the basis of IoU, which enhances the impact on the category frequency.

D. Ablation Studies

In this section, we successively focus on the decoder design, the different improvements in SSPD, and the channel compression strategy. A series of ablation experiments are carried out to study the effectiveness of SSPD-net. Both the training strategies and the data enhancement way are the same as the methods described in Section IV-B.

1) Design and Selection of Dual-Stage Decoder: For Decoder1, we have two designs. As shown in Fig. 3, the direct amplifier (DA) receives the output from the last layer of the encoder (output stride = 32). The CBM is employed to receive the output of five encoder layers (output stride=2, 4, 8, 16, 32). In Decoder2, we set ASPP as a baseline.

Baseline: The first-row block in Table II includes the results of a simple 16-time upsampling using bilinear interpolation (DA). In the CBM, five outputs are concurrently normalized to the identical size (output stride = 2) using bilinear interpolation. Both the DA and the CBM connect two 3 × 3 kernel convolutional layers for feature refinement. The experimental results reveal that the CBM significantly advances the performance, but the running time is not obviously increased.

Adding ASPP: We verify the feasibility of a dual-stage decoder using ASPP as the Decoder2. The comparison of rows 3 and 4 in Table III exhibits that ASPP helps to improve OA and mIoU by 0.4% and 1%, respectively, when the CBM is used as Decoder1. Similarly, when the DA is applied as Decoder1, OA and mIoU, respectively, increase by 2% and 5% due to ASPP.

Adding SSPD: We evaluate the two scenes proposed in Section III-C at the last row block in Table IV. SSPD (U) represents the fusion of 9 branches in the channel concatenation and the channelwise selection of branches. SSPD (+) embodies the fusion of 9 branches in element summation and the selection of branches in softmax operators. The differences in the experimental effects resulting from SSPD (U) and SSPD (+) are shown in Table IV. The results of this study indicate that SSPD (U) improves capability by about 2% with a slight increase in running time and complexity. The partial segmentation results are shown in Fig. 9. The results of the DA are jagged and incoherent. The large sawtooth of buildings in the CBM segmentation is gone, but the small buildings are partially missing, and they tend to stick together. After adding the ASPP, the segmentation result is evidently advanced but is slightly inferior to SSPD owing to the occasional breakup inside the buildings. It is interesting to note that SSPD (+) and SSPD (U) have a less visual difference, but the latter has smoother segmentation and lower missing alarm of small buildings.

2) Effectiveness of Internal Improvements in SSPD: In this section, we further explore the design of the Decoder2. The baseline is the selective pyramid convolution module that contains the first seven branches (two branches with kernel=5 are removed) without dilated convolution. The contrast experiments are based on Decoder1 = CBM and SSPD (U). As illustrated in Table V, the experimental results imply that the added larger kernel convolution (kernel = 5) promotes the segmentation property. The dilated convolution is found to cause better behavior than ordinary convolution. Compared with the baseline, the mIoU brought by the larger kernel and the dilated convolution increases by 0.7% and 1%, respectively. The current important finding is that under the condition of keeping the performance, replacing the dilated convolution with dilated depthwise separable convolution significantly reduces the model complexity and lowers the computing load on the hardware.

3) Channel Compression: The compression quantity r introduced in the Decoder1 is a hyperparameter, which compresses the number of channels of output feature maps in the encoder. To find the optimal balance between the capability and the complexity, we implement the experiments in different r values for SSPD-net. Table VI shows the influences for a range of different r values on the experimental effect. The experiments demonstrate that the monotone increase of r cannot lead to the linear growth of performance. According to the accuracy and calculation cost, $r = 48$ is an optimal choice. In summary, Table VII shows the statistics of the final segmentation results after compression of different proportions of high-level and low-level channel information. The different levels are represented by feature maps with different sizes in Table VII. The feature map of the higher level is smaller. These experimental results suggest that the balanced high-level and low-level information fusion has the best segmentation effect. In our structure, the equal compression of high-level and low-level semantic information is the final choice.

4) Improvement of the L-Shape Weighting Loss Function: In the experimental study, comparing LWloss with cross-entropy loss (CEloss) indicates that the former can heighten the precision of building extraction, especially for the optimized boundary of linear objectives. This result may be explained by the fact

TABLE III
ABLATION RESULTS OF DUAL-STAGE DECODER

Decoder1		**Decoder2**		**TT(s)**	**Params(M)**	**OA(%)**	**Recall(%)**	**Precision(%)**	**F1(%)**	**mIoU(%)**	**fwIoU(%)**
DA	CBM	ASPP	SSPD								
✓	×	×	×	**35**	**52.6**	88.08	86.28	75.47	78.66	67.11	78.67
✓	×	✓	×	37	92.3	89.95	87.93	77.85	82.95	72.37	81.90
×	✓	×	×	38	54.1	90.28	88.48	79.90	83.16	72.70	82.27
×	✓	✓	×	92	99.2	**90.65**	**88.76**	**80.53**	**83.93**	**73.70**	**82.91**

TABLE IV
ABLATION RESULTS OF DIFFERENT SSPDS

Decoder1		**Decoder2**		**TT(s)**	**Params(M)**	**OA(%)**	**Recall(%)**	**Precision(%)**	**F1(%)**	**mIoU(%)**	**fwIoU(%)**
DA	CBM	ASPP	SSPD								
×	✓	×	✓(+)	**96**	**82.4**	90.50	88.51	80.47	83.62	73.30	82.65
×	✓	×	✓(U)	111	85.6	**91.16**	**89.35**	**81.91**	**85.09**	**75.28**	**83.87**

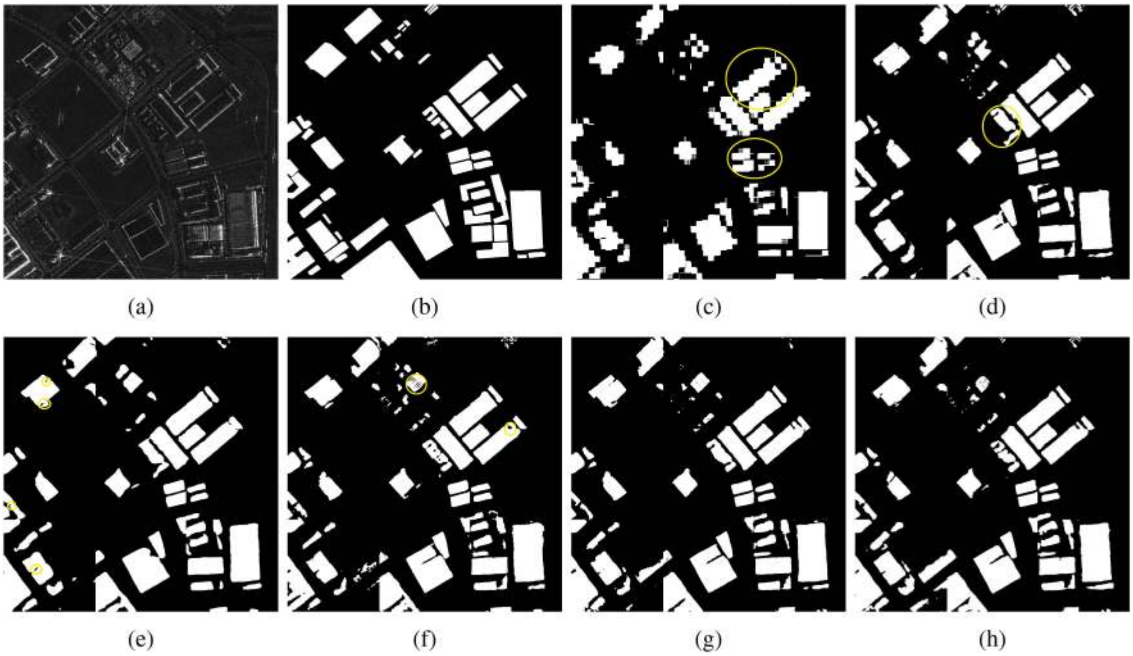


Fig. 9. Segmentation results of different combinations in dual-stage decoder over the urban areas. (a) SAR image. (b) Ground truth. (c) DA. (d) CBM. (e) DA + ASPP. (f) CBM + ASPP. (g) CBM + SSPD(+). (h) CBM + SSPD(U). The yellow circles in the diagram represent obvious defects in the segmentation results of some methods, such as (c) jaggies, (f) breakage, (e) deletion, (d) adhesion, and etc.

TABLE V
EFFECTS OF DIFFERENT SSPD DESIGNS. K: KERNEL

K=5	**Dilated**	**TT(s)**	**Params(M)**	**OA(%)**	**F1(%)**	**mIoU(%)**	**fwIoU(%)**
×	×	**86**	**66.7**	90.70	83.51	73.82	82.99
✓	×	90	84.3	90.31	83.60	74.58	82.87
✓	✓	111	85.6	**91.16**	**85.09**	**75.28**	**83.87**

TABLE VI
EFFECTS IN DIFFERENT r VALUES

r	**OA(%)**	**mIoU(%)**	**Params(M)**
16	90.12	72.47	**82.7**
32	90.40	73.31	84.1
48	**91.16**	**75.28**	85.6
64	90.29	73.08	87.2
80	90.34	72.99	88.9

that the LWloss can guide the training attention of the L-shape features and raise the effectiveness of building segmentation. In Fig. 10, we intuitively see that the extraction boundary of the building is more definite, and some of the missing building edges are supplemented in the LWloss results. For experiments with the cross-entropy loss function, the L-shape objectives' extraction boundary is confused and unclear with both deletions and dilations. The prediction results with LWloss, by contrast, are more sensitive to L-shape features, and the missing detection of some small L-shape targets decreases. At the same time, the boundary determination is much clearer, which reduces the boundary adhesion of the side-by-side buildings. Further, as shown in Table VIII, one interesting finding is that the IoU of the building class is higher than the mIoU of that, which confirms that the LWloss has a specific inhibition for the data imbalance. Meanwhile, LWloss has good portability for other networks.

TABLE VII
EFFECTS OF DIFFERENT RATIO OF HIGH AND LOW LAYER CHANNEL INFORMATION FUSION

Size(FP)	**256**	**128**	**64**	**32**	**16**	**OA(%)**	**mIoU(%)**	**Params(M)**
	6.7%	13.3%	20%	26.7%	33.3%	90.29	72.68	87.3
	20%	20%	20%	20%	20%	**91.16**	**75.28**	**85.6**
	33.3%	26.7%	20%	13.3%	6.7%	89.07	70.19	86.9

The level of the layer is negatively correlated with the size of the feature map (FP).

TABLE VIII
PERFORMANCE COMPARISON BETWEEN DIFFERENT LOSS FUNCTIONS

	Metrics	**OA(%)**			**IoU(%)**			**mIoU(%)**			**Missing alarm(%)**		
	Objectives	U-net	Deeplabv3+	SSPD-net	U-net	Deeplabv3+	SSPD-net	U-net	Deeplabv3+	SSPD	U-net	Deeplabv3+	SSPD-net
CEloss	Surface	83.14	85.23	91.38	34.53	53.08	61.04	-	-	-	13.28	11.88	11.65
	Line	80.68	82.78	87.79	31.75	49.78	58.13	-	-	-	16.83	18.03	17.37
	Complex	83.87	86.43	92.05	34.76	52.76	61.30	-	-	-	13.23	10.96	9.74
	Overall	82.70	86.06	89.92	33.38	51.24	60.27	58.17	69.27	75.15	14.75	12.83	12.28
LWloss	Surface	85.89	87.89	92.63	35.27	53.54	60.75	-	-	-	12.43	12.05	11.67
	Line	82.26	85.07	88.92	32.81	50.52	58.98	-	-	-	15.91	17.11	16.41
	Complex	85.71	88.52	92.35	35.92	53.42	61.74	-	-	-	12.38	11.87	9.52
	Overall	84.37	87.25	91.16	34.05	52.10	60.80	58.52	69.84	75.28	13.69	12.24	12.05

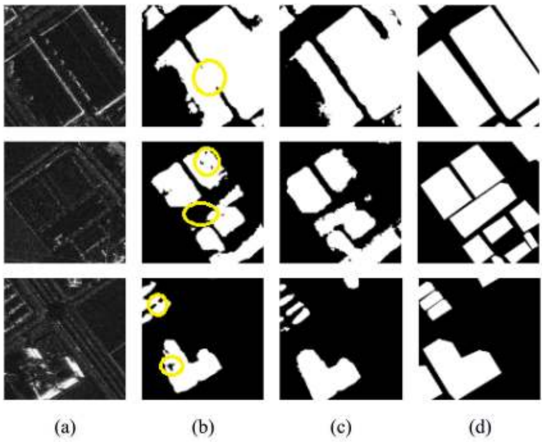


Fig. 10. Extraction effects of partial linear objectives with different loss functions in SSPD-net. Each column from left to right. (a) Input images. (b) Extraction results with cross-entropy loss. (c) Extraction results with LWloss. (d) Ground truths.

TABLE IX
PERFORMANCE COMPARISON BETWEEN DIFFERENT EXTRACTION OF THE L-SHAPE FOOTPRINT SETS

Methods	**OA(%)**	**Recall(%)**	**Precision(%)**	**mIoU(%)**
CEloss	88.23	84.35	72.10	68.99
CFAR+LN	88.68	85.78	76.60	68.92
CFAR+PR	**91.07**	86.17	77.01	70.13
MP+DMP	87.96	84.05	73.38	66.09
LWloss(-)	89.98	88.83	79.52	73.14
LWloss(+)	90.65	88.46	**79.82**	74.28
LW-Eloss	88.12	86.34	79.60	72.66
LWloss	91.02	**88.91**	79.77	**74.79**

The overall accuracy and the mIoU with the L-shape weighting loss for several networks are lifted, as shown in Table VIII. The optimal building extraction performance is achieved in SSPD-net with the LWloss. It should be noted that the convergence time of model training is increased by 22%. On the whole, the proposed loss function alleviates the data imbalance problem in the SAR building segmentation and improves the extraction effect of linear objectives at the cost of extra time, which is acceptable.

In addition, the proposed LWloss function is based on the distance calculation and the weighted superposition of the set of L-shape footprint terrain pixels extracted from SAR images. The ablation experiments are conducted on the accuracy of extracting the L-shape footprint pixels. The comparing methods of extracting the L-shape footprint set are discussed, including utilizing the constant false alarm rate (CFAR) algorithm combined with the log-normal (LN) distribution probability density function to extract linear features of buildings, utilizing the CFAR detector and power ratio (PR) method to extract buildings, utilizing the morphological profiles (MP) and the difference morphological profiles (DMP) [70] to extract the morphological information. Meanwhile, the addition and subtraction sets of the L-shape footprint pixels extracted by the proposed method are compared. This process is carried out by the corrosion and expansion operations. Regarding the weighting factor of the L-shape footprints in (11), the weighted Euclidean distance based on the exponential weight function is also used to compare the performance of the different feasible LWloss functions. We define it as LW-Eloss, as shown in (18), where λ is set to 1

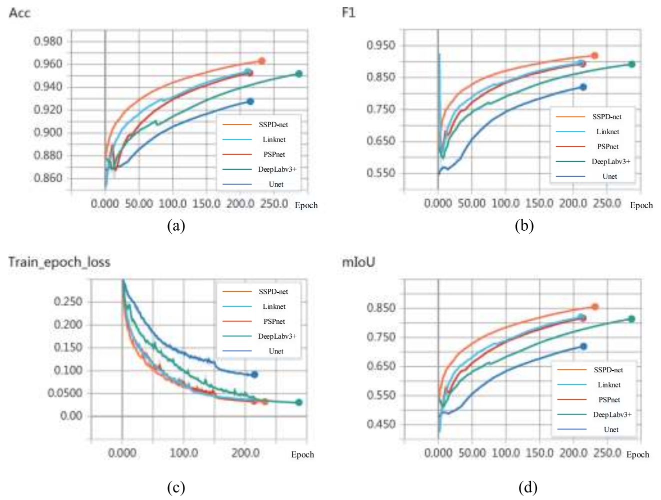
$$\begin{aligned} L &= -\sum_{N} w_p y_{\text{true}} log\left(y_{\text{pred}}\right) + w_p - \frac{2w_p \sum\limits_{N} y_{\text{true}} y_{\text{pred}}}{\sum\limits_{N} y_{\text{true}}^2 + \sum\limits_{N} y_{\text{pred}}^2} \\ w_p &= 1 + w_0 \sum_{q\in M} \lambda e^{-\lambda\left[(x_p-x_q)^2+(y_p-y_q)^2\right]}. \end{aligned} \tag{18}$$

The experiment results are shown in Table IX. The first-row block in Table IX shows that different traditional methods of extracting the L-shape footprint set perform well in overall accuracy and other evaluating metrics than the CEloss function. However, the change of each method is about 1%, which is not obvious enough. In addition, the results of the increase or

TABLE X
COMPARISON OF SEGMENTATION RESULTS OF DIFFERENT MODELS

Methods	TT(s)	Params(M)	GFLOPs	OA(%)	Recall(%)	Precision(%)	F1(%)	mIoU(%)	fwIoU(%)
2-Mode [71]	-	-	-	76.06	75.76	69.27	61.76	48.94	63.85
OTSU [72]	-	-	-	81.24	80.90	75.02	57.11	47.44	67.09
Threshold-histogram [73]	-	-	-	44.60	42.55	38.78	41.85	27.28	33.30
K-means [36]	-	-	-	70.25	69.49	62.54	53.65	41.75	57.66
MRF [74]	-	-	-	80.32	78.18	73.85	63.99	51.87	67.91
PMRF [75]	-	-	-	81.16	80.64	74.36	61.51	50.35	68.02
U-net [55]	**41**	150.0	642.87	84.37	83.06	76.45	70.75	58.52	73.01
Linknet [76]	42	**82.7**	**339.14**	90.53	88.77	80.72	84.01	73.78	82.83
DeepLabv3+ [49]	108	159	667.83	89.06	87.35	78.88	80.94	69.84	80.35
PSPnet [77]	247	186	773.76	90.01	88.42	80.19	82.60	71.96	81.81
SSPD-net(our)	111	85.6	363.81	**91.16**	**89.35**	**81.91**	**85.09**	**75.28**	**83.87**

The computational cost is evaluated with FLOPs, i.e. floating point operations [78]. 1 GFLOPs $= 10^9$ FLOPs.



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

Fig. 11. Example training curves of several methods on the Gaofen-3 dataset. The SSPD-net exhibits the stable optimization characteristics and gains the best training performance.

decrease operations on the L-shape footprint set generated by the proposed method are revealed in the second-row block. Though some metrics show a slight decrease, the effects are better than the CEloss function. Generally, extracting the L-shape footprint set and generating the distance loss function make a relatively large contribution to the building extraction, while the slight increase, decrease, and morphological change of the L-shape footprint set have a little impact on the building extraction. Besides, LW-Eloss generates a loss function based on the exponential weight distance, whose test metrics are 1% lower than those of the proposed loss function based on the Gaussian weight distance. Finally, LWloss is considered to be used to guide the network training attention.

E. Comparison With Other Methods

In comparison, other methods are tested based on the same Gaofen-3 satellite SAR dataset, including the 2-Mode [71], OTSU [72], Threshold-histogram [73], K-means [36], the conventional MRF [74], the improved MRF (PMRF) [75], U-net [55], Linknet [76], DeepLabv3+ [49], and PSPnet [77]. Considering the apparent contrast between targets and backgrounds, two global single threshold segmentation methods, the OTSU [72] and the 2-Mode [71], are added to observe the optimal solution under the customary criterion, such as the maximum intraclass variance and gray histogram. The PMRF [75] mainly adds a multiscale MRF image pyramid model based on the MRF [74]. What matters is the case that the conventional MRF and the PMRF both follow the experimental settings in [75]. Moreover, the lightweight Linknet [76] and the PSPnet [77] with capable global context aggregation are augmented for the sake of fully verifying our advantages over other advanced deep convolutional models.

The example training curves for different algorithms are depicted in Fig. 11. It can be observed that the proposed method yields the most stable improvement throughout the whole optimization process. Their experimental results are listed in Table X, and the corresponding segmentation examples are shown in Figs. 12 and 13. Although the MRF model does not require training and is extremely fast in the test process, it contains little semantic information of the SAR buildings, and the segmentation results are rough. In contrast, the abundant semantic information is contained in the SSPD-net with 14% higher accuracy. For the manifestation of the U-net method, the large holes exist in the interior of the buildings. In the Linknet segmentation results, some small buildings adhere and the boundaries are difficult to distinguish. Certain medium-sized buildings are missing in DeepLabv3+ segmentation images. A flow of PSPnet is the inability to extract a large number of dense small buildings. Compared with the current advanced U-net, Linknet, DeepLabv3+, and PSPnet methods, our method (SSPD-net) obtains the best accuracy, F1, mIoU, and fwIoU. Furthermore, the results of this study show that SSPD-net achieves the effect of fine building segmentation. Simultaneously, it does not increase the parameter complexity and not lower the running speed. Finally, the segmentation effect of our method on the large-scene SAR image is shown in Fig. 14.

V. DETAILED ANALYSIS

A. Design and Selection of Dual-Stage Decoder

We evaluate the performance of the dual-stage decoder in Table IV. The results show that the combination of CBM and SSPD achieves the finest segmentation with the ResNet-based encoder. Compared with the direct amplifier, the CBM combines more full semantic information of both high and low layers simply and directly. The model capability will be better with

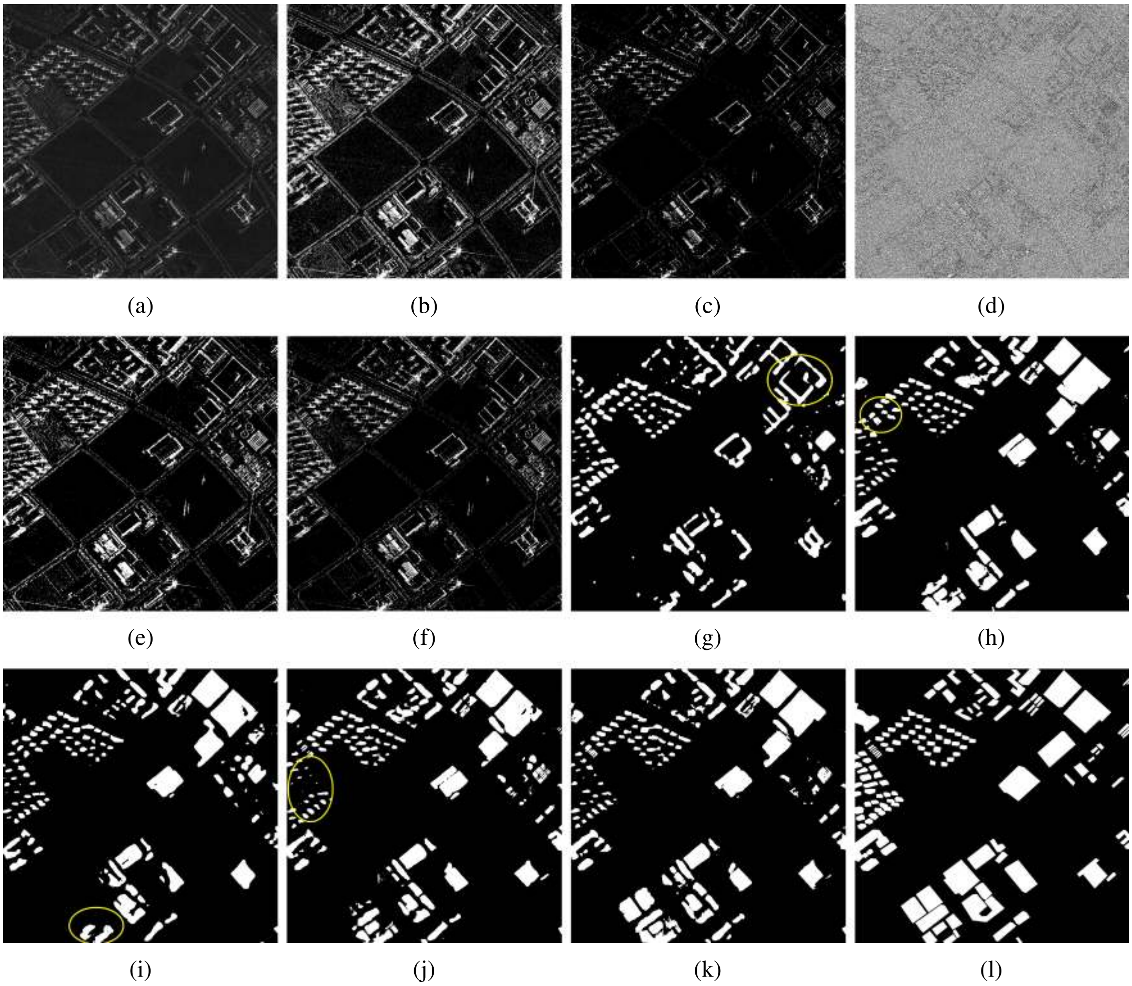


Fig. 12. Segmentation results of different methods. From top to bottom, from left to right, the images in turn are as follows. (a) SAR. (b) 2-mode. (c) OTSU. (d) Threshold-histogram. (e) MRF. (f) PMRF. (g) U-net. (h) Linknet. (i) DeepLabv3+. (j) PSPnet. (k) SSPD-net. (l) Ground truth. The defects of some deep learning methods are marked in yellow circles in the graph.

an equal compression strategy. For Decoder2, the experiments further corroborate the advantages of SSPD over ASPP. This discrepancy could be attributed to the channel and the branch selection attention for SSPD instead of the spatial pyramid structure. The obvious finding to emerge from Fig. 9 is that SSPD can better divide the building boundary between small and large buildings than the general spatial pyramid pooling module. Another important finding is that the channel-based soft branch selection [SSPD(U)] is more effective than the hard-selection branch mode [SSPD(+)]. A possible explanation for this might be due to the channel attention concentration of the former to the target.

B. *Effectiveness of Internal Improvements in SSPD*

For the details of SSPD, we add the branches with large kernel convolution and dilated depthwise separable convolution, which increases the mIoU by 0.8% and 0.7%, respectively. The branches with large kernel convolution effectively supplement the convolution probing fields of the spatial pyramid module. Each complementary nesting combination of convolution branches exerts a pivotal part in multiscale information extraction. The dilated depthwise separable convolution calculates the feature mapping with higher sampling density to restore the full resolution feature maps so that the computed feature mapping is denser. This improvement gives the whole network a more profitable receptive field. Besides, from Table V, the increase of model complexity mainly lies in the addition of large kernel branches, while the dilated depthwise separable convolution has a small effect on the model parameters. Compared with other models [49], [55], our network has also achieved superior model lightweight, which effectively improves its extensibility.

C. *Channel Attention Analysis*

In order to prove the effects of the proposed SSPD, we observe the attention weight of SSPD under different building scales. Fig. 15 shows the visualization effects of the feature maps for some samples containing small buildings and large buildings in the output layers. The first four patches are samples dominated by small buildings and the last four dominated by large buildings. When the input is an SAR image with the small building dominant, the network attention is mostly focused on

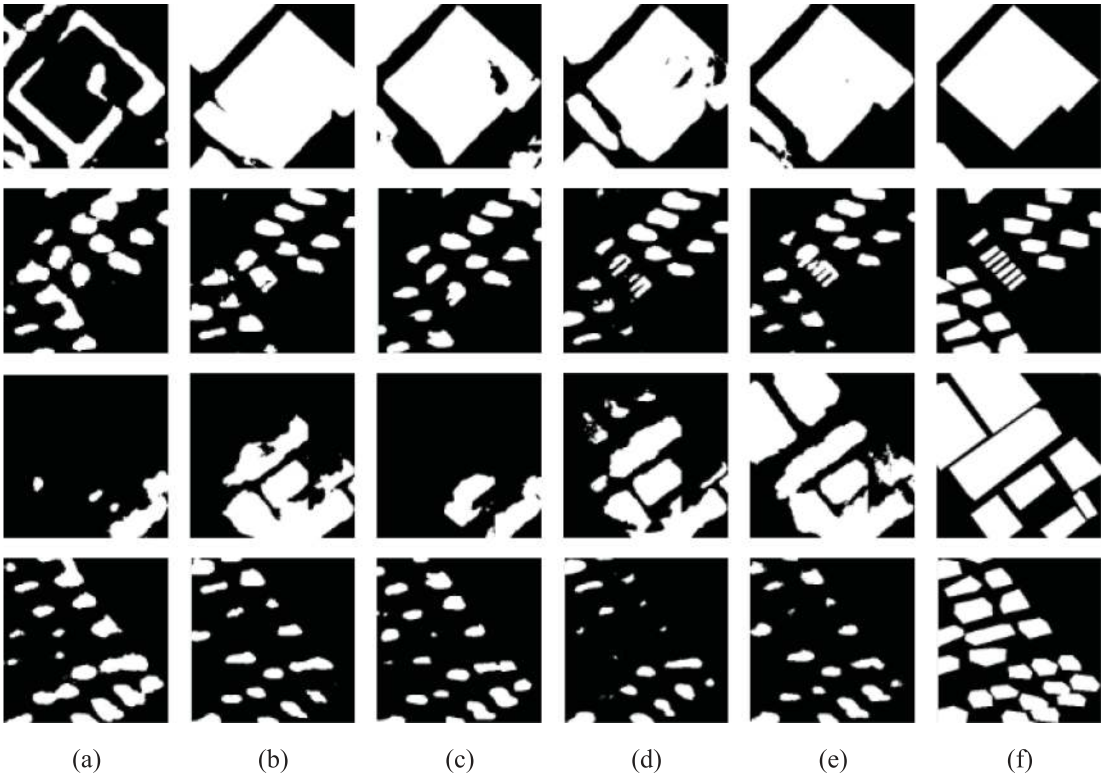


Fig. 13. Local area presentation of segmentation results (see the yellow circle in Fig. 12) for different methods. Each column from left to right belongs to the following methods. (a) U-net. (b) Linknet. (c) DeepLabv3+. (d) PSPnet. (e) SSPD-net. (f) ground truth.

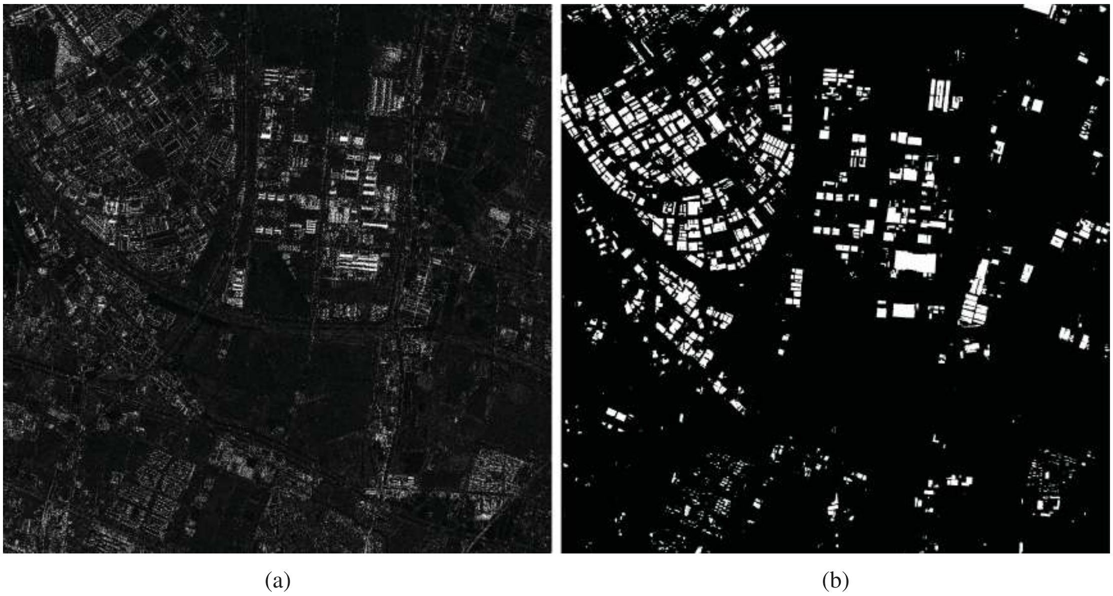


Fig. 14. Fine building segmentation on the large-scene SAR image. The left is the original image and the right is the segmentation image by SSPD-net.

the small buildings. Conversely, attention reverses. The attention value distributions of two random samples on all branches in SSPD are shown in Figs. 16 and 17, where the two samples are patches including some small buildings and large buildings, respectively. The channel activation value for most of the small receptive fields in SSPD is high for the minor targets. As the target object size increases, the channel activation of the large receptive fields rises, which seems to be consistent with our expected network selectivity.

D. Interpretation of the L-Shape Weighting Loss

As shown in Fig. 18, we plot the attention maps for the SSPD-net with the CEloss and the LWloss. The LWloss results

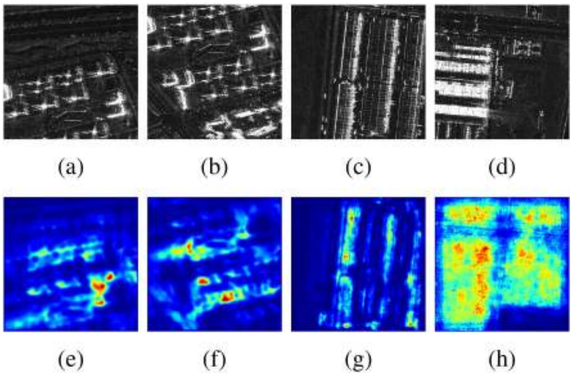


Fig. 15. Feature maps for small and large buildings. (a) and (b) SAR images with small buildings. (c) and (d) SAR images with large buildings. (e) and (f) Feature maps corresponding to the (a) and (b). (g) and (h) Feature maps corresponding to the (c) and (d).

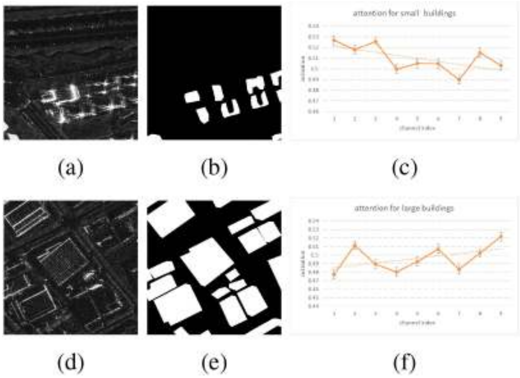


Fig. 16. Attention value of the channels in small and large buildings. (a) and (b) Image of small buildings and its groudtruth. (c) Attention distribution of small buildings samples on all branches. (d) and (e) Image of large buildings and its groudtruth. (f) Attention distribution of large buildings samples on all branches.

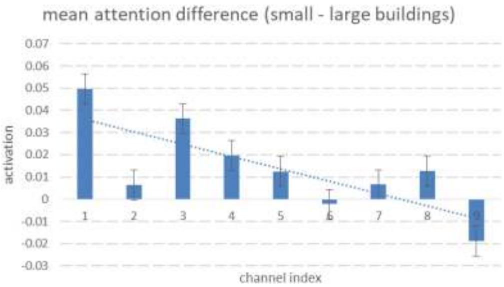


Fig. 17. Mean attention difference in small and large buildings.

show that the attention division between the buildings and the background areas is distinct, and especially the attention to the linear objectives rises in the network. In the CEloss results, some line objectives and small buildings are easily affected by other complex structures and thus missing the sensitivity to some easily neglected structures, although the boundary extraction of some large buildings is acceptable. Table VIII illustrates that the LWloss is suitable for some current segmentation networks and has a guaranteed improvement in extraction precision and other metrics. The LWloss has good potential for the fine building segmentation in SAR images.

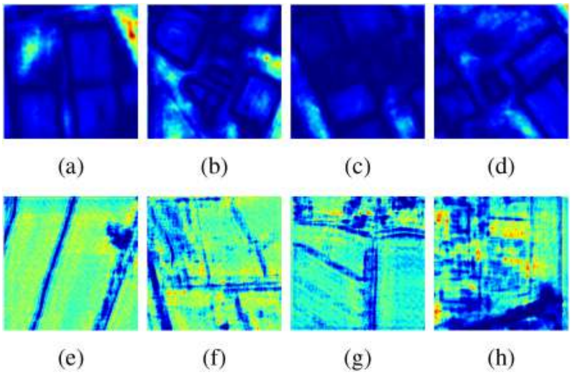


Fig. 18. Comparison of the CEloss and the LWloss. Each row from top to bottom: (a)–(d) CEloss results. (e)–(h) LWloss results.

VI. Conclusion

In this article, a unified framework named SSPD-net is proposed for the fine building segmentation in SAR images based on the selective attention mechanisms. We design the dual-decoder, the CBM, and the advanced SSPD convolution module. The multibranch information is fused and reselected to conform to the multiscale extraction with the specific building attention. Additionally, in light of the building features and SAR imaging mechanism, the LWloss for the fine building extraction is established to promote the attention on the L-shape footprint characteristics of buildings. The extraction effects of linear targets are enhanced, and the class imbalance problem in the training process is restrained with the LWloss. The experimental results on a high-resolution SAR dataset demonstrate the superiority of our approach.

Acknowledgment

The authors would like to thank all colleagues in the lab for generously providing their raw data and helping to annotate the images. The authors were grateful for the helpful suggestions of anonymous Reviewers.

References

[1] Z. Xu *et al.*, "Effect analysis and spectral weighting optimization of sidelobe reduction on SAR image understanding," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 9, pp. 3434–3444, Sep. 2019.

[2] K. Fu, F.-Z. Dou, H.-C. Li, W.-H. Diao, X. Sun, and G.-L. Xu, "Aircraft recognition in SAR images based on scattering structure feature and template matching," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4206–4217, Nov. 2018.

[3] L. Li, L. Du, and Z. Wang, "Target detection based on dual-domain sparse reconstruction saliency in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4230–4243, Nov. 2018.

[4] Z. Wang, L. Du, and H. Su, "Superpixel-level target discrimination for high-resolution SAR images in complex scenes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3127–3143, Sep. 2018.

[5] L. Huang *et al.*, "OpenSARShip: A dataset dedicated to Sentinel-1 ship interpretation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 1, pp. 195–208, Jan. 2018.

[6] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 12, pp. 7177–7188, Dec. 2017.

[7] S. Xian, W. Zhirui, S. Yuanrui, D. Wenhui, Z. Yue, and F. Kun, "Air-sarship-1.0: High resolution SAR ship detection dataset," *J. Radars*, vol. 8, no. 6, pp. 852–862, 2019.

[8] C. He, B. Shi, Y. Zhang, X. Su, W. Yang, and X. Xu, "The algorithm of building area extraction based on boundary prior and conditional random field for SAR image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 1321–1324.

[9] M. Sun, L. Pang, H. Liu, X. Zhang, L. Ai, and S. He, "Urban extraction based on multi-scale building information extra-segmentation and SAR coherence image," in *Proc. Int. Conf. Geo-Informat. Resource Manage. Sustain. Ecosystem*, 2015, pp. 471–479.

[10] Z. Zhang, L. Chen, W. Yu *et al.* "Super-resolution for MIMO array SAR 3-D imaging based on compressive sensing and deep neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3109–3124, 2020.

[11] G. Yang, H.-C. Li, W.-Y. Wang, W. Yang, and W. J. Emery, "Unsupervised change detection based on a unified framework for weighted collaborative representation with RDDL and fuzzy clustering," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8890–8903, Nov. 2019.

[12] Z. Wang, L. Du, J. Mao, B. Liu, and D. Yang, "SAR target detection based on SSD with data augmentation and transfer learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 150–154, Jan. 2019.

[13] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[14] Y. Feng, W. Diao, X. Sun, M. Yan, and X. Gao, "Towards automated ship detection and category recognition from high-resolution aerial images," *Remote Sens.*, vol. 11, no. 16, 2019, Art. no. 1901.

[15] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 161, pp. 294–308, 2020.

[16] M. Shahzad, M. Maurer, F. Fraundorfer, Y. Wang, and X. X. Zhu, "Buildings detection in VHR SAR images using fully convolution neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1100–1116, Feb. 2019.

[17] Z. Zhang, W. Guo, W. Yu, and W. Yu, "Multi-task fully convolutional networks for building segmentation on SAR image," *J. Eng.*, vol. 2019, no. 20, pp. 7074–7077, 2019.

[18] Y. Zhao, L. Zhao, C. Li, and G. Kuang, "Pyramid attention dilated network for aircraft detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 662–666, Apr. 2021.

[19] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020.

[20] Z. Cui, Q. Li, Z. Cao, and N. Liu, "Dense attention pyramid networks for multi-scale ship detection in SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8983–8997, Nov. 2019.

[21] G. Priyanka, R. Gandhirai, and K. Soman, "Target recognition in SAR using deep learning," in *Proc. 2nd Int. Conf. on Intell. Comput., Instrum. and Control Technol.*, vol. 1, 2019, pp. 1442–1447.

[22] L. Li, C. Wang, H. Zhang, and B. Zhang, "Residual Unet for urban building change detection with Sentinel-1 SAR data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 1498–1501.

[23] R. Jaturapitpornchai, M. Matsuoka, N. Kanemoto, S. Kuzuoka, R. Ito, and R. Nakamura, "SAR-image based urban change detection in Bangkok, Thailand using deep learning," in *Proc. IGARSS 2019-2019 IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 7403–7406.

[24] H. Dong, L. Zhang, and B. Zou, "Densely connected convolutional neural network based polarimetric SAR image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 3764–3767.

[25] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[26] P. M. Mather, B. Tso, and M. Koch, "An evaluation of landsat TM spectral data and SAR-derived textural information for lithological discrimination in the Red sea hills, Sudan," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 587–604, 1998.

[27] H. Yu, X. Zhang, S. Wang, and B. Hou, "Context-based hierarchical unequal merging for SAR image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 2, pp. 995–1009, Feb. 2013.

[28] H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using Gibbs random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 1, pp. 39–55, Jan. 1987.

[29] Q. Yu and D. A. Clausi, "IRGS: Image segmentation using edge penalties and region growing," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 30, no. 12, pp. 2126–2139, Dec. 2008.

[30] L. K. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 2, pp. 780–795, Mar. 1999.

[31] A. Voisin, V. A. Krylov, G. Moser, S. B. Serpico, and J. Zerubia, "Classification of very high resolution SAR images of urban areas using copulas and texture in a hierarchical Markov random field model," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 1, pp. 96–100, Jan. 2013.

[32] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: a SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.

[33] C. Tison, J.-M. Nicolas, F. Tupin, and H. Maĺtre, "A new statistical model for Markovian classification of urban areas in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 10, pp. 2046–2057, Oct. 2004.

[34] A. A. Amory, A. O. Rokabi, A. El Zaart, H. Mathkour, and R. Sammouda, "Fast optimal thresholding based on between-class variance using mixture of log-normal distribution," in *Proc. Int. Conf. Inf. Technol. e-Serv.*, 2012, pp. 1–12.

[35] A. Dutta and K. K. Sarma, "SAR image segmentation using wavelets and Gaussian mixture model," in *Proc. Int. Conf. Signal Process. Integr. Netw.*, 2014, pp. 466–770.

[36] K. Venkateswaran, N. Kasthuri, K. Balakrishnan, and K. Prakash, "Performance analysis of k-means clustering for remotely sensed images," *Int. J. Comput. Appl.*, vol. 84, no. 12, pp. 23–27, 2013.

[37] X. Yang, C. Liu, K. Wu, and W. Lang, "SAR sea ice image segmentation using SRRG-MRF," *J. Remote Sens.*, vol. 18, no. 6, pp. 1993–2002, 2014.

[38] G. S. Xia, C. He, and H. Sun, "Integration of synthetic aperture radar image segmentation method using Markov random field on region adjacency graph," *IET Radar Sonar Navigation*, vol. 1, no. 5, pp. 348–353, 2007.

[39] M. Wang, W. Fu, S. Hao, D. Tao, and X. Wu, "Scalable semi-supervised learning by efficient anchor graph regularization," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1864–1877, Jul. 2016.

[40] G. Niu, W. Jitkrittum, B. Dai, H. Hachiya, and M. Sugiyama, "Squared-loss mutual information regularization: A novel information-theoretic approach to semi-supervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 10–18.

[41] E. Protopapadakis, A. Doulamis, N. Doulamis, and E. Maltezos, "Semi-supervised fine-tuning for deep learning models in remote sensing applications," 2020, *arXiv:2006.00345*.

[42] K. Makantasis, A. D. Doulamis, N. D. Doulamis, and A. Nikitakis, "Tensor-based classification models for hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6884–6898, Dec. 2018.

[43] C. Henry, S. M. Azimi, and N. Merkle, "Road segmentation in SAR satellite images with deep fully convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 12, pp. 1867–1871, Dec. 2018.

[44] X. Yang, X. Li, Y. Ye, R. Y. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.

[45] J. Geng, H. Wang, J. Fan, and X. Ma, "SAR image classification via deep recurrent encoding neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2255–2269, Apr. 2018.

[46] J. Geng, H. Wang, J. Fan, and X. Ma, "Deep supervised and contractive neural network for SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2442–2459, Apr. 2017.

[47] W. Yao, D. Marmanis, and M. Datcu, "Semantic segmentation using deep neural networks for SAR and optical image pairs," in *Proc. Big Data Space*, 2017.

[48] Y. Duan, X. Tao, C. Han, X. Qin, and J. Lu, "Multi-scale convolutional neural network for SAR image semantic segmentation," in *Proc. IEEE Global Commun. Conf.*, 2018, pp. 1–6.

[49] 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. Vision.*, 2018, pp. 801–818.

[50] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf, Computer Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[51] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 510–519.

[52] S. Bell, C. Lawrence Zitnick, K. Bala, and R. Girshick, “Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks,” in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2016, pp. 2874–2883.

[53] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in *Proc. Eur. Conf. Computer Vis.*, 2016, pp. 483–499.

[54] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu , “Spatial transformer networks,” in *Proc. Adv. Neural Inf. Process, Sys.*, 2015, pp. 2017–2025.

[55] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[56] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[57] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2016, pp. 770–778.

[58] M. Abadi *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” 2016, *arXiv:1603.04467*.

[59] V. Nair and G. E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[60] L. Sifre and S. Mallat, “Rigid-motion scattering for image classification,” Ph D. dissertation, CMAP, Ecole Polytechnique, Palaiseau, France, 2014.

[61] V. Vanhoucke, “Learning visual representations at scale,” *ICLR invited talk*, vol. 1, no. 2, 2014.

[62] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” 2015, *arXiv:1511.07122*.

[63] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2017, pp. 1251–1258.

[64] A. G. Howard *et al.*, “MobileNets: Efficient convolutional neural networks for mobile vision applications,” 2017, *arXiv:1704.04861*.

[65] M. Wang, B. Liu, and H. Foroosh, “Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial” bottleneck” structure,” 2016, *arXiv:1608.04337*.

[66] F. Zhang, Y. Shao, X. Zhang, and T. Balz, “Building L-shape footprint extraction from high resolution SAR image,” in *Proc. Joint Urban Remote Sens. Event*, 2011, pp. 273–276.

[67] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proc. Fourth Int. Conf. 3D Vision*, 2016, pp. 565–571.

[68] A. Paszke *et al.*, “Automatic differentiation in Pytorch,” in *Proc. 31st Conf. Neural Inf. Process. Syst.*, 2017.

[69] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.

[70] S. Adelipour and H. Ghassemian, “The fusion of morphological and contextual information for building detection from very high-resolution SAR images,” in *Proc. Elect. Eng. (ICEE), Iranian Conf.*, 2018, pp. 389–393.

[71] J. M. Prewitt and M. L. Mendelsohn, “The analysis of cell images,” *Ann. the New York Acad. Sci.*, vol. 128, no. 3, pp. 1035–1053, 1966.

[72] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Trans. Syst., Man, Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[73] O. J. Tobias and R. Seara, “Image segmentation by histogram thresholding using fuzzy sets,” *IEEE Trans. Image Process.*, vol. 11, no. 12, pp. 1457–1465, Dec. 2002.

[74] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. London, U.K.: Springer, 2009.

[75] X.-Y. Fu, H.-J. You, and K. Fu, “Building segmentation from high-resolution SAR images based on improved Markov random field,” *Dianzi Xuebao(Acta Electronica Sinica)*, vol. 40, no. 6, pp. 1141–1147, 2012.

[76] A. Chaurasia and E. Culurciello, “LinkNet: Exploiting encoder representations for efficient semantic segmentation,” in *Proc. IEEE Vis. Commun. Image Process.*, 2017, pp. 1–4.

[77] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vision. Pattern Recognit.*, 2017, pp. 2881–2890.

[78] X. Zhang, X. Zhou, M. Lin, and J. Sun, “ShuffleNet: An extremely efficient convolutional neural network for mobile devices,” in *Proc. IEEE Conf. Computer Vis. Pattern Recognit.*, 2018, pp. 6848–6856.

Hao Jing received the B.Sc. degree in electronic information engineering from Dalian University of Technology, Dalian, China, in 2017. He is currently working toward the Ph.D. degree in signal and information processing with the University of Chinese Academy of Sciences, Beijing, China, and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China .

His research interests include synthetic aperture radar (SAR) image segmentation, geospatial data mining, and remote sensing image understanding.

Xian Sun received the B.Sc. degree from the Beijing University of Aeronautic and Astronautics, Beijing, China, in 2004, and the M.Sc. and Ph.D. degrees in electronic information engineering from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2009.

He is a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Zhirui Wang received the B.Sc. degree in electronic information engineering from the Harbin Institute of Technology, Harbin, China, in 2013, and the Ph.D. degree from Tsinghua University, Beijing, China, in 2018.

He is an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include synthetic aperture radar (SAR) terrain classification, and SAR target detection and recognition.

Kaiqiang Chen received the Ph.D. in information and communication engineering from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2019.

He was a Visiting Scholar with the Karlsruhe Institute of Technology, Karlsruhe, Germany, from November 2017 to October 2018. He is currently an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.

Wenhui Diao received the B.Sc. degree from Xidian University, Xi’an, China, in 2011, and the M.Sc. and Ph.D. degrees in electronic information engineering from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2016.

He is currently an Assistant Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision and remote sensing image analysis.

Kun Fu received the B.Sc., M.Sc., and Ph.D. degrees in electronic information engineering from the National University of Defense Technology, Changsha, China, in 1995, 1999, and 2002, respectively.

He is a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, remote sensing image understanding, and geospatial data mining and visualization.