Spatial Structure Preserving Feature Pyramid Network for Semantic Image Segmentation

YUAN YUAN, Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, China
JIE FANG, Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, China and University of Chinese Academy of Sciences, China
XIAOQIANG LU and YACHUANG FENG, Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, China

Recently, progress on semantic image segmentation is substantial, benefiting from the rapid development of Convolutional Neural Networks. Semantic image segmentation approaches proposed lately have been mostly based on Fully convolutional Networks (FCNs). However, these FCN-based methods use large receptive fields and too many pooling layers to depict the discriminative semantic information of the images. Specifically, on one hand, convolutional kernel with large receptive field smooth the detailed edges, since too much contexture information is used to depict the "center pixel." However, the pooling layer increases the receptive field through zooming out the latest feature maps, which loses many detailed information of the image, especially in the deeper layers of the network. These operations often cause low spatial resolution inside deep layers, which leads to spatially fragmented prediction. To address this problem, we exploit the inherent multi-scale and pyramidal hierarchy of deep convolutional networks to extract the feature maps with different resolutions and take full advantages of these feature maps via a gradually stacked fusing way. Specifically, for two adjacent convolutional layers, we upsample the features from deeper layer with stride of 2 and then stack them on the features from shallower layer. Then, a convolutional layer with kernels of $\mathbf{1} \times \mathbf{1}$ is followed to fuse these stacked features. The fused feature preserves the spatial structure information of the image; meanwhile, it owns strong discriminative capability for pixel classification. Additionally, to further preserve the spatial structure information and regional connectivity of the predicted category label map, we propose a novel loss term for the network. In detail, two graph model-based spatial affinity matrixes are proposed, which are used to depict the pixel-level relationships in the input image and predicted category label map respectively, and then their cosine distance is backward propagated to the network. The proposed architecture, called spatial structure preserving feature pyramid network, significantly improves the spatial resolution of the predicted

This work was supported in part by the National Natural Science Foundation of China under Grant 61772510 and 61702498, in part by the Key Research Program of Frontier Sciences, CAS under Grant QYZDY-SSW-JSC044, in part by the Young Top-notch Talent Program of Chinese Academy of Sciences under Grant QYZDB-SSW-JSC015, in part by the National Key R&D Program of China under Grant 2017YFB0502900, in part by the CAS "Light of West China" Program under Grant XAB2017B15.
Authors' addresses: Y. Yuan, Center for OPTical Imagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, China; email: y.yuan1.ieee@gmail.com; J. Fang, Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi, 710119, China, University of Chinese Academy of Sciences, Beijing, 100049, China; email: fangjie2015@opt.cn; X. Lu (corresponding author) and Y. Feng, Key Laboratory of Spectral Imaging Technology CAS, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi, 710119, China; emails: luxq666666@gmail.com, fengyachuang@opt.cn.

1551-6857/2019/08-ART73 $15.00
https://doi.org/10.1145/3321512

category label map for semantic image segmentation. The proposed method achieves state-of-the-art results on three public and challenging datasets for semantic image segmentation.

CCS Concepts: • **Computing methodologies** → **Neural networks**; **Feature selection**;

Additional Key Words and Phrases: Semantic image segmentation, spatial resolution, feature pyramid network, discriminative capability

ACM Reference format:
Yuan Yuan, Jie Fang, Xiaoqiang Lu, and Yachuang Feng. 2019. Spatial Structure Preserving Feature Pyramid Network for Semantic Image Segmentation. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 3, Article 73 (August 2019), 19 pages.
https://doi.org/10.1145/3321512

1 INTRODUCTION

Semantic image segmentation is an important task for understanding objects in a scene, which aims to parse images into several semantic regions. Specifically, it assigns each pixel or superpixel to one of the annotated semantic categories automatically. As a bridge toward high-level tasks, semantic image segmentation is adopted in various applications, such as food detection (Aguilar et al. 2018), human pose estimation (Xu et al. 2013), visual tracking (Hong et al. 2015), and so on. Even though remarkable efforts (He et al. 2017; Kang et al. 2018; Kemker et al. 2017; Shi et al. 2018; Wang et al. 2018) have been made for image semantic segmentation during the past decades, it is still a challenging problem.

Most recent semantic image segmentation methods (Dai et al. 2015; Lin et al. 2016; Souly et al. 2017; Yasrab 2017) to solve structured pixelwise labeling problems are based on *fully convolutional networks* (FCNs) (Long et al. 2015). Compared to the traditional *convolutional neural networks (CNNs)*, FCN ignores the fully connected layers, and its output is a two-dimensional matrix but not a vector. The main advantage of FCN-based methods is that the network accepts a whole image as an input, which makes it perform fast and accurate inference. Although many FCN-based methods have achieved state-of-the-art performance, these methods suffer from **a common limitation**: *low spatial resolution of the predicted category label map*. Because no explicit pixel grouping mechanism is employed, pixel-level relationships inside the deepest layer become inconsistent with the input image, which causes spatially fragment outputs.

Many existing methods are proposed to address the aforementioned limitation, such as FCN+CRF (Zhou et al. 2016), deconvolution network (Noh et al. 2016), U-Net (Ronneberger et al. 2015), DeepLab-based methods (Chen et al. 2017, 2018a, 2018b), and other embedding-based methods (Castillo et al. 2017; Kampffmeyer et al. 2016; Ravi et al. 2017; Sevillalara et al. 2016; Xiao et al. 2018). However, these methods suffer from another limitation: *the increment number of network parameters*, which leads to the increment of the computational complexity. Therefore, training these models is usually space and time cost. In these cases, *how to improve the spatial resolution of the predicted category label map with fewer model parameters* becomes a new challenging problem for semantic image segmentation.

As for CNNs, the outputs of low layers have more spatial structure information because of fewer convolution and pooling operations (Dechesne et al. 2017; Holliday et al. 2017). Conversely, the outputs of deep layers have stronger discriminative semantic information (Ghiasi and Fowlkes 2016; Hong et al. 2016; Paszke et al. 2016). According to these characteristics of the outputs from different layers, fusing the feature maps from CNNs is an appropriate way to address the limitation of low-spatial-resolution prediction. This is because the fused features contain both deep semantic and high-spatial-resolution information. Recently, *feature pyramid network* (FPN) (Lin

et al. 2017) has been proposed for object detection and achieved significant performance, especially for the small objects. The reason is that FPN fuses the feature maps from lower and deeper layers effectively and makes the final representation equip with strong discriminative capacity while remaining detailed spatial structure information. In this article, inspired by the successful application of FPN (Lin et al. 2017) in the task of object detection, we propose a method named *spatial structure preserving feature pyramid network* (SSPFPN) for semantic image segmentation. To exploit the spatial structure information of different layers, the proposed method adopts two strategies as follows: First, the proposed SSPFPN fuses feature maps with different resolutions from VGG19-Net to enhance spatial information of final prediction via a gradually stacked way. Specifically, for feature maps from adjacent convolutional layers, we upsample the features from deeper layer with stride of 2 and stack them on those from shallower layer. Besides, a convolutional layer with kernels of 1×1 is followed to fuse the stacked features. Second, to further exploit the spatial information and remain the reginal connectivity of the predicted category label map, the proposed method constructs a novel loss term. Specifically, two graph model-based spatial affinity matrices are proposed, which are used to depicts the pixel-level relationships in the input image and the pixel-level relationships in the predicted category label map, respectively. Then their cosine distance is added to the traditional softmax loss and backward propagated to the network. The proposed method can be described as follows:

(1) The VGG19-Net pre-trained on ImageNet is used to extract feature maps of images. Because both spatial structure and semantic discriminative information are vital for semantic image segmentation task, we extract multi-scale feature maps from different layers. Specifically, the deeper and lower layer feature maps consist semantic discriminative and spatial structure information of the image, respectively.
(2) The proposed SSPFPN is used to fuse the extracted feature maps with different size in a gradually stacked way. Importantly, a novel loss term is incorporated into our optimization function. In particular, reginal connectivity and spatial structure information can be remained through measuring the cosine distance of corresponding input image and predicted category label map spatial affinity matrixes.
(3) The semantic image segmentation task is completed by a pixel-level classifier. Differing from traditional FCN-based methods, as for ours, each pixel in image image has a independent feature vector. First, training samples for pixel-level classifier has higher degree of freedom, which avoids the overfitting to a large extent. Furthermore, pixel-level classifier in this task remains more detail spatial structure information of the image compared to the region-level ones.

Additionally, the proposed method outperforms state-of-the-art methods on three public and challenging datasets: PASCAL VOC 2012 (Hariharan et al. 2011), NYUDv2 (Silberman et al. 2012), and SIFT Flow (Tighe and Lazebnik 2010). In summary, the contributions of our work are listed as follows.

(1) The SSPFPN is proposed for the task of image semantic segmentation. It generates high-spatial-resolution features with deep semantic information by fusing lower and deeper features of pre-trained VGG19-Net.
(2) Two affinity matrices are proposed to depict the spatial structure information of the input image and the predicted category label map, respectively.
(3) The proposed novel loss term preserves pixel-level relationships in the input image to the predicted category label map. It enhances the spatial information of the predicted label map by measuring the similarity of two affinity matrices.

(4) The proposed SSPFPN is relatively small. It only contains 1×1 convolutional kernels, which results in few parameters and makes the network easy to train.

The article is organized as follows. In Section 2, we introduce some works related to our method. Section 3 describes the proposed method. We report the experiments in Section 4 and conclude the article in Section 5.

2 RELATED WORKS

This section details some related works for semantic image segmentation. First, most recent approaches for semantic image segmentation are based on classification architecture, so the representation of image is very important. Additionally, many existing methods combine features from different sources to improve the segmentation results. Therefore, feature fusion is also vital for this task. According to the reasons aforementioned above, this section includes three parts: (a) image features; (b) feature fusion; and (c) previous works.

2.1 Image Features

There are many feature extractors that can be applied to semantic image segmentation tasks, either handcrafted features or learning-based features. In early time, handcrafted features are in the lead, especially two popular ones: HOG (Dalal and Triggs 2005) and SIFT (Abdel-Hakim and Farag 2006). HOG features, and later SIFT features, were computed densely over entire pyramids. These HOG and SIFT pyramids have been used in numerous works for image classification, object detection, and so on. Recently, with the rapid progress of machine learning and artificial intelligence, CNNs equipped with significant capability in extracting image features are proposed, for instance, VGG-Net (Simonyan and Zisserman 2014), GoogleNet (Szegedy et al. 2015), and so on. These networks are successfully used in many computer visual tasks, such as semantic segmentation (Zhou et al. 2016) and visual tracking (Hong et al. 2015) and achieve satisfactory performances.

2.2 Feature Fusion

A number of recent approaches improve the accuracy of semantic segmentation by combining predictions from different layers in a ConvNet. FCN (Long et al. 2015) combines coarse-to-fine predictions from multiple layer by averaging segmentation probabilities. SSD (Liu et al. 2016) predicts objects at different layers of the feature hierarchy. Another category of approaches combine features from multiple layers before making the prediction. These methods include Hypercolumns (Hariharan et al. 2015), HyperNet (Kong et al. 2016), ParseNet (Liu et al. 2015), and ION (Bell et al. 2016).

There are recent methods exploiting lateral/skip connections that associate low-level feature maps across resolutions and semantic levels, including SharpMask (Pinheiro et al. 2016), Recombinator networks (Honari et al. 2016), and Stacked Hourglass networks (Newell et al. 2016) for segmentation and face detection, respectively. Ghiasi and Fowlkes (2016) present a Laplacian pyramid presentation for FCNs to progressively refine segmentation. Although these approaches implicitly utilize pyramidal architectures, they are not image pyramids (Felzenszwalb et al. 2010; Sermanet et al. 2013) where predictions are made independently as all levels.

2.3 Previous Works

In the past years, semantic image segmentation has attracted a lot of attention, because it is a basic problem in the computer vision filed. The recent introduced FCNs (Long et al. 2015) has led to remarkable results for semantic image segmentation task. However, due to the large receptive fields and many pooling layers, the FCN typically suffers from low-spatial-resolution predictions, which causes inconsistent relationships among the neighboring pixels inside the deep layers.

Recently, there have been some attempts to address these problems. These works can be divided into several groups.

The work in Donahue et al. (2013), Eigen et al. (2014), and Jia et al. (2014) used FCN to learn unary potentials, in the separate globalization framework to refine the original FCN results. One of the disadvantages related to these methods is that the learning unary potentials and globalization framework is completely disjoint and not integrated into the learning. To address this issue, several recent methods (Gupta et al. 2013, 2014; Krizhevsky et al. 2012) proposed to integrate a CRF inference procedure into their networks, which allowed to train such models in an end-to-end fashion and achieve satisfactory performance. However, one disadvantage of these methods is the dramatically increased computational complexity. For example, the method in Krizhevsky et al. (2012), cast the original FCN into a *recurrent neural network* (RNN), which is much more computationally expensive. The work in Gupta et al. (2014) proposed to use local convolutional layers, which lead to a significantly larger number of overall parameters. Similarly, the method in Gupta et al. (2013) proposed to model unary and pairwise potentials by the separate multi-scale network branches, which essentially doubles the number of parameters compared to the traditional FCN architectures. In addition, it is also worth mentioning the de-convolution networks (Ganin and Lempitsky 2014; Hariharan et al. 2015), which use de-convolution and unpooling layers to recover the fine object details from the coarse FCN predictions in an end-to-end fashion. However, to effectively recover the fine details, one must employ almost the same number of deconvolution layers as their convolution layers, which results in a largely increased number of parameters as well.

All of the above methods utilize traditional FCN (Long et al. 2015) in their architectures. While these methods have been able to circumvent the spatial inconsistency problem inside the deep layers of traditional FCNs, their solutions did not fix low spatial resolution of the predicted category feature map. Additionally, they lack consistent grouping mechanisms in the deep low-spatial-resolution layers inside the network. Our main goal, in this work, is to propose such a spatial grouping mechanism to improve the spatial resolution of the predicted category label map. Recently, FPN (Lin et al. 2017) has been proposed for object detection and achieved significant performance, especially for the small objects. The reason is that FPN fuses the feature maps from lower and deeper layers effectively and makes the final representation equip with strong discriminative capacity while remaining detailed spatial structure information. Consider object detection furthermore, if the objects in the image are all pixel-level ones in the image, this task is transformed to semantic image segmentation. In other words, semantic image segmentation is a special object detection task.

3 PROPOSED METHOD

Semantic image segmentation actually consists of two subtasks: segmenting the image to several regions and attaching category label to each region. Most recent successful methods are based on classification architecture, and they use the deepest features of the convolutional neural networks to classify each pixel directly. In other words, classification-based methods skip the segmentation process, and the detailed spatial structure information of the image is ignored. Additionally, there are not any regional connectivity and explicit pixel grouping mechanisms for their networks. All these cases lead to low spatial resolution of the predicted category label map and further limit the development of classification-based methods for semantic image segmentation.

In this work, a SSPFPN is proposed to enhance the feature representation for semantic image segmentation. The architecture of SSPFPN is shown in Figure 1. Additionally, the original feature maps extracted with the pre-trained VGG19-Net are introduced in Section 3.1. On one hand, SSPFPN generates high-spatial-resolution features with high-level semantic information by fusing low and deep features from VGG19-Net via a gradually stacked way, which is proposed in

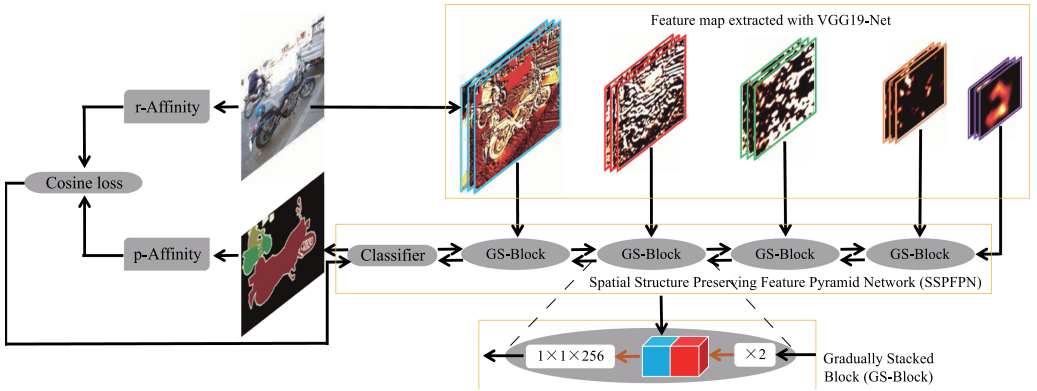

Fig. 1. The flowchart of proposed method for semantic image segmentation. It is based on the pre-trained VGG19-Net and the proposed SSPFPN. VGG19-Net is used to extract the feature maps of different convolutional layers, and the proposed SSPFPN is used to fuse these feature maps. Efficient and robust classifier is needed for the specific task. Additionally, combined with the classifier loss, a cosine loss term based on two constructed affinity matrices is also incorporated into the network, which is to keep pixel-level relationships in the predicted category label map. To simplify the flowchart, we use the "GS-Block" module to replace some operations, and all the blocks in the figure have the same architecture. The "GS-Block" is detailed in Section 3.2.2.

Section 3.2. On the other hand, pixel-paired-based similarity generated by graph models is introduced in the novel SSPFPN for making the predicted semantic category label map preserve the spatial structure information of the input image. This part is based on the loss function of SSPFPN, which is described as Equation (1),

$$L = \gamma \cdot L_c + (1 - \gamma) \cdot \ell, \tag{1}$$

where L_c is the classification loss, which mainly ensure the correct predicted category label for each pixel, and ℓ is the proposed term that emphasizes the pixel-pair-based similarity in the image. The details of the loss function is introduced in Section 3.3.

3.1 Multi-layer Feature Maps

Recently, with the development of machine learning, CNNs (Lin et al. 2016; Souly et al. 2017), such as GoogLeNet (Szegedy et al. 2015), and so on, they have achieved satisfactory performances on many computer vision tasks (Dai et al. 2015; Yasrab 2017) because of their significant capacity in extracting image's high-level semantic features. VGG19-Net is used to extract multi-layer features in this article. The deep features have high-level general semantic information, and superficial layer features have more detail local information of the image, and there is no doubt that fusing the deep and superficial layer features through an appropriate strategy can achieve better performance on the task of semantic image segmentation.

To simplify the model, only highest stage feature maps (conv1_2, conv2_2, conv3_4, conv4_4, conv5_4) of each size in VGG19-Net are chosen. The reason is that, for the feature maps with same size, deeper layer in ConvNet has higher semantic information and can depict the image more accurate.

3.2 Spatial Structure Preserving Feature Pyramid Network

The recent FPN (Lin et al. 2017) has achieved significant performance on object detection task, especially for the small objects. Additionally, if the "object" in an image is further smaller to pixel

level, then the object detection task is transformed to the semantic image segmentation task. Therefore, the FPN processed by structural adjustment can be applied to semantic image segmentation, appropriately. In this article, our goal is to fuse the feature maps from deep and low layers of VGG19-Net to generate high-spatial-resolution features with deep semantic information. Compared to the task of object detection, the "object" is much smaller in semantic segmentation; therefore, more lower feature maps with detailed spatial information should be considered, and the merged strategy between two stages should be enhanced.

The proposed SSPFPN takes a single-scale image of an arbitrary size as input, and outputs proportionally sized feature maps at multiple levels, in a fully convolutional fashion. This process is independent of the backbone convolutional architectures. The construction of the proposed pyramid involves a bottom-up pathway, a top-down pathway, and a gradually stack strategy, as introduced in the following.

3.2.1 Bottom-up Pathway. The bottom-up pathway is the feedforward computation of the backbone ConvNet, which computes a feature hierarchy consisting of feature maps at several scales with a scaling step of 2. There are often many layers producing output maps of the same size, and we define these layers are in the same network stage. For the proposed pyramid, we define one pyramid level for each stage. We choose the output of the last layer of each stage as our reference set of feature maps, which we use to enrich our pyramid. This choice is natural, since the deepest layer of each stage should have the highest semantic features. Specifically, for VGG19-Net, we use the feature activations output by each stage's last layer. We donate these last residual blocks as $\{C_1, C_2, C_3, C_4, C_5\}$ for conv1, conv2, conv3, conv4, and conv5 outputs, and note that they have strides of {2, 4, 8, 16, 32} pixels with respect to the input image.

3.2.2 Top-down Pathway and Gradually Stack Strategy. The top-down pathway hallucinates higher-resolution features by upsampling spatially coarser, but semantically stronger, feature maps from higher pyramid levels. These features are fused with features from the bottom-up pathway using the proposed GS-Block, which is to keep the discriminative information and the structure details of the image simultaneously.

Figure 2 shows the proposed GS-Block that constructs our top-down feature maps. Inspired by the connection approach of *densely connected convolutional networks* (Huang et al. 2017), the extracted feature cubes with different resolutions are processed by the followings steps: (a) With a coarser-resolution feature map cube, we upsample the spatial resolution by factor of 2 (using nearest neighbour upsampling for simplicity); (b) the upsampled feature map cube is then stacked on the corresponding bottom-up map cube; (c) the stacked feature map cube is merged further by a convolutional layer with 1×1 kernels. This process is iterated until the finest resolution map is generated. This final set of top-down feature cubes is called $\{P_1, P_2, P_3, P_4, P_5\}$, corresponding to $\{C_1, C_2, C_3, C_4, C_5\}$ that are respectively of the same spatial sizes on bottom-up map cubes.

Because all levels of the pyramid use shared classifier/regressors in a traditional featurized image pyramid, we fix the feature dimension (denoted as d) in all the feature maps. We set $d = 256$ in this article and thus all extra convolutional layers have 256-channel outputs. The inner architecture of the proposed SSPFPN is designed in Table 1, where N donates the class number of the dataset.

3.3 The Novel Loss Function for SSPFPN

To enhance the relationships among different pixels in predicted category label map, besides the traditional softmax loss, a new part based on the graph model is added to the final loss function of the SSPFPN. The new loss is based on the assumption that two pixels with similar color and location information are more likely belong to the same category. On the basis of this assumption, we build two affinity matrixes for original RGB image and the predicted category label map in

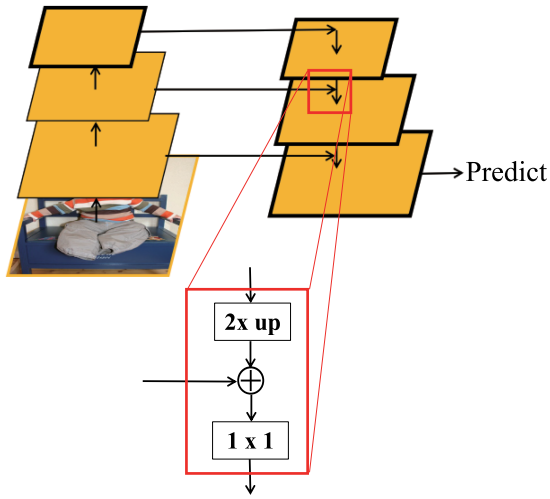

Fig. 2. A building "GS-Block" illustrating the gradually stacked feature cubes and top-down pathway, merged by addition. The top-down stack strategy is shown in the right branch in the figure, and the feature cube in this branch is called "P." Besides that, the left branch in the figure is the traditional bottom-up flow of CNNs, the feature cube in this branch is called "C." This strategy leads the stacked feature maps into a 1×1 convolution layer sized 256 to decrease the channels to the fixed number 256.

Table 1. The Inner Architecture of Spatial Structure Preserving Feature Pyramid Network

Layer	Input1_size (VGG19)	Input2_size (SSPFPN)	Kernel	stride	Output_size
Conv_1	$14 \times 14 \times 512$	/	$1 \times 1 \times 256$	2	$28 \times 28 \times 256$
ReLU	/	/	/	/	$28 \times 28 \times 256$
Conv_2	$28 \times 28 \times 512$	$28 \times 28 \times 256$	$1 \times 1 \times 256$	2	$56 \times 56 \times 256$
ReLU	/	/	/	/	$56 \times 56 \times 256$
Conv_3	$56 \times 56 \times 256$	$56 \times 56 \times 256$	$1 \times 1 \times 256$	2	$112 \times 112 \times 256$
ReLU	/	/	/	/	$112 \times 112 \times 256$
Conv_4	$112 \times 112 \times 128$	$112 \times 112 \times 256$	$1 \times 1 \times 256$	2	$224 \times 224 \times 256$
ReLU	/	/	/	/	$226 \times 224 \times 256$
Conv_5	$224 \times 224 \times 64$	$224 \times 224 \times 256$	$1 \times 1 \times 256$	1	$224 \times 224 \times 256$
SoftMax	/	$224 \times 224 \times 256$	/	/	$224 \times 224 \times (N + 1)$

each iteration, respectively, and then use cosine distance to measure the difference between two affinity matrixes, and both the measurement loss and the traditional softmax loss are feedback to train the network. The details of the optimization strategy are described as following subsections.

3.3.1 Spatial Affinity Matrix for Raw Images. First, we need to consider this question: How can we describe the spacial relationship of pixels in the image efficiently? As we all know, each pixel in an image includes the three-dimensional color intensity vector C = $[R, G, B]$ and the two-dimensional location coordinate vector L = [X, Y], so these two vectors are used to calculate the affinity matrix based on graph model.

Assume that the input images have the size of $m \times n \times 3$, the affinity matrix of the image will be the size of $mn \times mn$. First, to enhance the robustness of the affinity matrix, C and L for each pixel are normalized to c and l. The normalized functions are shown in Equation (2) and Equation (3),

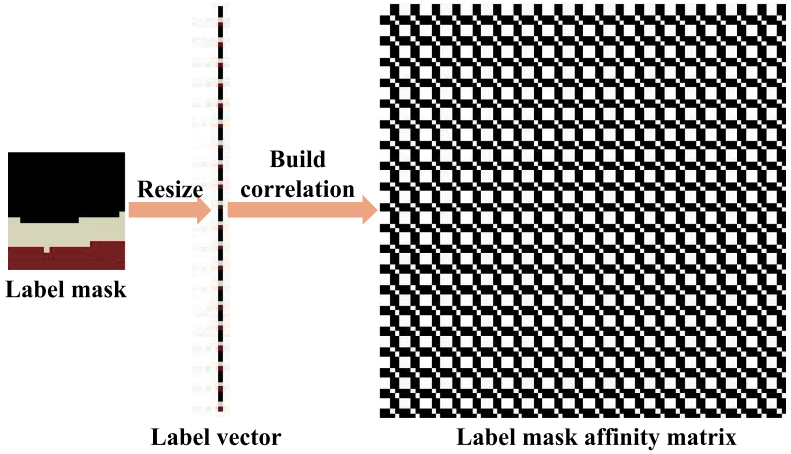

Fig. 3. The spacial affinity matrix for predicted label masks, and the process is shown in the figure. First, the label mask is resized to a vector. Second, an affinity matrix sized $mn \times mn$ is calculated by the proposed strategy.

respectively,

$$C \rightarrow c : [r, g, b] = [R, G, B]/255, \tag{2}$$

$$L \rightarrow l : [x, y] = \left[\frac{X}{m-1}, \frac{Y}{n-1}\right]. \tag{3}$$

Then, we concatenate the c and l vector of each pixel as a new vector $f = [c; l]$ and calculate the distance matrix $D^{mn \times mn \times 5}$ using Equation (4),

$$\begin{aligned} &\forall \quad 0 \le i, k \le m-1, 0 \le j, t \le n-1, \\ &D_{ni+j, nk+t, :} = \left| f_{i,j,:} - f_{k,t,:} \right|. \end{aligned} \tag{4}$$

Finally, the affinity matrix S_{rA} is calculated using Equation (5),

$$\begin{aligned} &\forall \quad 0 \le p, q \le mn-1, \\ &(S_{rA})_{p,q} = \exp\left(-\left|D_{p,q,0:2}\right|_1 - \lambda \left|D_{p,q,3:4}\right|_1\right), \end{aligned} \tag{5}$$

where λ is a hyperparameter to balance the the relationship of RGB intensity and the location information.

3.3.2 Spacial Affinity Matrix for Predicted Label Masks. Despite the spacial affinity matrix for raw images, we also need the affinity information of the predicted masks. As we all know, the outputs of network are images class label masks, we reshape the label mask to a vector $M^{mn \times 1}$ and define an affinity matrix $S_{pA}^{mn \times mn}$ as Equation (6),

$$S_{pA}^{(i,j)} = \begin{cases} 1, & if \quad M_i = M_j \\ 0, & if \quad M_i \neq M_j. \end{cases} \tag{6}$$

According to the invariability of images affinity, the matrix $S_{rA}^{mn \times mn}$ and the matrix $S_{pA}^{mn \times mn}$ should have the same spatial characteristic, so we can use the otherness of the two matrix to update the parameters of the whole networks and improve the predict accuracy. The spacial affinity matrix for predicted label mask is shown in Figure 3.

3.3.3 The Specific Loss Function. On the one hand, the high classification accuracy for each-pixel's label should be remained; on the other hand, the relationship among pixels in same image should be considered as well. Therefore, the overall loss functions of the proposed SSPFPN consists two terms. L_c in Equation (1) is the traditional sofmax loss of FCN, which is used to remain the classification accuracy, and it is defined as Equation (7),

$$L_c = -\frac{1}{mn}\left[\sum_{i=1}^{mn}\sum_{j=1}^{k} 1\{y^i = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right], \tag{7}$$

where θ represents the parameters of the proposed SSPFPN, k is the number of category in the dataset, mn is the number of pixels in each image, and $1\{\cdot\}$ is an indicator function. It equals to 1 when the condition in {} is satisfied, and, otherwise, it equals 0. The proposed term ℓ is used to emphasize the relationship among different pixels by enhancing the spatial information of the segmentation, and it is defined as Equation (8),

$$\ell = 1 - \cos\left(A_{rv}, A_{pv}\right). \tag{8}$$

In our work, when we calculate the similarity between the original and predicted annotated mask images, we reshape the affinity matrixe of predicted category label map $S_{pA}^{mn\times mn}$ and the affinity matrix of original RGB image $S_{rA}^{mn\times mn}$ to vectors A_{rv} and A_{pv}, respectively. The size of A_{rv} and A_{pv} are both $(mn)^2 \times 1$. $\cos\langle\cdot,\cdot\rangle$ is the cosine function, which is not sensitive to the specific values of the data, and its definition is as follows:

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} {x_i}^2}\sqrt{\sum_{i=1}^{t} {y_i}^2}}, \tag{9}$$

where t is the size of the vectors, x_i and y_i are the elements of vector $\mathbf{x}$ and $\mathbf{y}$, respectively.

Additionally, we utilize γ is to adjust the two aspects in Equation (1). The bigger the γ, the more important the pixel classification and the less important the spacial affinity.

3.3.4 Implementation. Because A_{rv} and A_{pv} are both high-dimensional feature vectors, the computational complexity of directly measuring the cosine distance of these two feature vectors is relatively high. To simplify the computation, we divide the image into several regions to calculate the proposed loss term. Specifically, both the original RGB image and the predicted category label map are divided into $N \times N$ regions, and then similarities of each block in RGB image and corresponding block in category label map are calculated. The final measurement is the average value of these N^2 similarities, and the specific formula is

$$\cos(A_{rv}, A_{pv}) \approx \frac{1}{N^2}\sum_{l=1}^{N^2} \cos\left(A_{rv}^l, A_{pv}^l\right), \tag{10}$$

where A_{rv}^l and A_{pv}^l are the vectorized information of the l_{th} block in original RGB image and the l_{th} block in predicted category label map, respectively. Of course, it is not the key point in this article, so we do not discuss the details here.

4 EXPERIMENTS

This section details the experiments, including the following five parts: Datasets, Experiment Settings, Metrics, Two Versions of the Proposed Method, and Experiment Results and Analysis.

Table 2. Classes in the PASCAL VOC 2012 Dataset

No.	Category	No.	Category	No.	Category	No.	Category	No.	Category
1	background	6	boat	11	cow	16	person	21	tv/screen
2	aeroplane	7	bottle	12	dining table	17	potted plant		
3	bicycle	8	car	13	dog	18	sheep		
4	bird	9	cat	14	horse	19	sofa		
5	boat	10	chair	15	motorbike	20	train		

Fig. 4. Samples from PASCAL VOC 2012 dataset.

4.1 Datasets

Three public and challenging datasets are used to verify the proposed method.

PASCAL VOC 2012 dataset (Hariharan et al. 2011) for semantic segmentation includes 2,913 label images for 21 semantic categories. The class information is shown in Table 2, and some samples of the dataset are shown in Figure 4.

NYUDv2 (Silberman et al. 2012) is an RGB-D dataset collected using the Microsoft Kinect, and it has 1,449 RGB-D images with pixelwise labels. This dataset is more challenging, because the additional depth information increases the structure complexity of the image. Some samples of the dataset are shown in Figure 5.

SIFT Flow (Tighe and Lazebnik 2010) is a dataset of 2,688 images with pixel labels for 33 semantic classes. The class information of the dataset is shown in Table 3, and some samples are shown in Figure 6.

4.2 Experiment Settings

In this article, we choose 80% images of the datasets to train the net, and use the others for testing. Specifically, for each dataset we test in this article, we consider the original training set, testing set, and validation set as an entire one. And then we randomly choose a certain percentage of the entire set as the training set and the others as the testing ones. We train the network using mini-batch SGD with patch size 224×224 and batch size 10 and the initial learning rate is set to 2.5×10^{-4}. Weight decay is set to 5×10^{-4}, and momentum is 0.9, the network is trained for 50 epochs. Besides the settings referred above, there is an additional important balance parameter in the proposed method to be chosen, γ in Equation (1), which is set to 0.85 in this article.

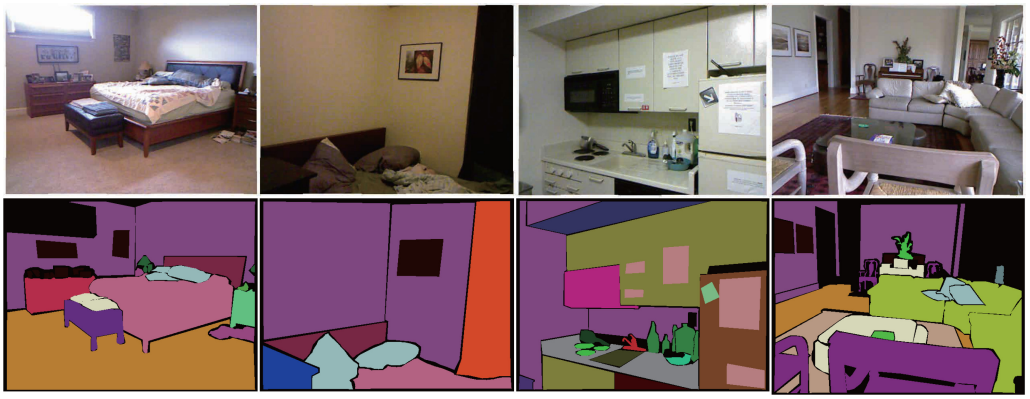

Fig. 5. Samples from NYUDv2 dataset.

Table 3. Classes in the SIFT Flow Dataset

No.	Category	No.	Category	No.	Category	No.	Category	No.	Category
1	awning	8	car	15	grass	22	road	29	staircase
2	balcony	9	cow	16	moon	23	rock	30	streetlight
3	bird	10	crosswalk	17	mountain	24	sand	31	sun
4	boat	11	desert	18	person	25	sea	32	tree
5	bridge	12	door	19	plant	26	sidewalk	33	window
6	building	13	fence	20	pole	27	sign		
7	bus	14	field	21	river	28	sky		

Fig. 6. Samples from SIFT Flow dataset.

4.3 Metrics

Inspired by Long et al. (2015), we also report four metrics from common semantic segmentation and scene parsing evaluations that are variations on pixel accuracy and region intersection over union (IU). Let n_{ij} be the number of pixels of class i predicted to belong to class j, where there are n_{c1} different classes, and let $t_i = \sum_j n_{ij}$ be the total number of pixels of class i. We compute the following:

- pixel accuracy (Pix acc.):

$$\sum_i n_{ii} / \sum_i t_i \tag{11}$$

- mean accuracy (Mean acc.):

$$(1/n_{c1}) \sum_i n_{ii}/t_i \tag{12}$$

- mean IU:

$$(1/n_{c1}) \sum_i n_{ii} / \left(t_t + \sum_j n_{ji} - n_{ii}\right) \tag{13}$$

- frequency weighted IU (*F.w.* IU):

$$\left(\sum_k t_k\right)^{-1} \sum_i t_i n_{ii} / \left(t_i + \sum_j n_{ji} - n_{ii}\right) \tag{14}$$

4.4 Contrasting Methods

To verify the effectiveness of the proposed method, we compare the experiment results of our method with FCN32s, FCN16s, FCN8s (Long et al. 2015), Deconvolution network (Deconv) (Noh et al. 2016), U-Net (Ronneberger et al. 2015), DeepLab-v2 (Chen et al. 2017), DeepLab-v3 (Chen et al. 2018a), and Hypercolumns (Hariharan et al. 2015).

(FCN32s), (FCN16s), and (FCN8s) are three versions of the original FCN architecture; the only difference among them is the unpooling stride step of the final prediction procedure.

(Deconv) is a method based on encoder-decoder architecture that achieves satisfactory performance for semantic image segmentation with double parameters compared to FCN methods.

(U-Net) is very similar to deconvolution network. The only difference is that U-Net utilizes skip-connection strategy to preserve the spatial structure information from the lower to deeper layers.

(DeepLab-v2) uses multi-scale receptive field domain to improve the segmentation performance, which considers more sufficient contexture information of the very pixel.

(DeepLab-v3) incorporates atrous spatial pyramid pooling operation into the network architecture, which significantly improve the segmentation performance. More importantly, DeepLab-v3 does not need the dense conditional random field as the postprocessing procedure, which simplifies the calculation to a large extent.

(Hypercolumns) defines the hypercolumn at a pixel as the vector of activations of all CNN units above that pixel, which obtains a more effective and robust representation of the very pixel.

4.5 Two Versions of the Proposed Method

There are different strategies to address the semantic segmentation task with CNN methods based on classification architecture: end-to-end fashion and separated-fashion. The former optimize the parameters of feature extractor and that of classier at the same time, while the latter only use CNN as a feature extractor, and the extracted features are used to train another independent classifier. Both strategies have their advantages and shortcomings: The former could gain the results directly without any extra medium operations, but the softmax classifier cannot achieve satisfactory performance sometimes. The latter could use any classifier equipped with significant capacity

Table 4. Results on Three Public and Challenging Datasets (%)

Dataset	Method	Pix acc.	Mean acc.	Mean IU	*F.w.* IU
PASCAL VOC 2012 Dataset	FCN32s	90.3	75.7	62.4	83.5
	FCN16s	91.9	77.1	64.3	85.6
	FCN8s	92.5	78.9	66.2	87.3
	Deconv	92.8	79.5	69.3	88.6
	U-Net	93.1	79.7	71.5	88.4
	DeepLab-v2	93.2	79.9	72.7	88.9
	DeepLab-v3	93.6	81.2	75.4	89.3
	Hypercolumns	91.8	79.4	67.2	87.5
	End-to-end(Ours)	93.6	80.5	74.1	89.9
	Separated(Ours)	**94.1**	**81.4**	**76.3**	**91.1**
NYUDv2 Dataset	FCN32s	64.2	45.3	33.8	49.1
	FCN16s	64.9	46.1	34.8	51.3
	FCN8s	66.8	52.7	40.4	55.8
	Deconv	68.3	54.9	42.7	59.1
	U-Net	68.5	55.3	43.1	59.7
	DeepLab-v2	68.7	56.1	43.3	60.8
	DeepLab-v3	70.5	58.4	46.6	63.9
	Hypercolumns	65.4	49.6	38.5	53.4
	End-to-end(Ours)	70.2	57.8	45.3	63.4
	Separated(Ours)	**72.3**	**60.1**	**47.0**	**64.7**
SIFT Flow Dataset	FCN32s	85.1	52.4	40.9	76.0
	FCN16s	85.4	53.6	42.3	78.4
	FCN8s	85.9	55.2	44.7	81.7
	Deconv	87.4	56.1	45.5	82.9
	U-Net	87.7	56.3	45.8	83.2
	DeepLab-v2	87.9	56.8	46.1	83.4
	DeepLab-v3	89.2	58.5	47.6	85.6
	Hypercolumns	86.5	54.2	43.9	79.8
	End-to-end(Ours)	88.3	57.4	47.3	84.5
	Separated(Ours)	**90.5**	**59.6**	**49.7**	**86.1**

to improve the classification accuracy, but besides the complex medium operations, image features need to be stored before the prediction, it is not as convenient as the former.

According to the descriptions above, the proposed method can be processed in two branches: end-to-end version and separation version. Specifically, the end-to-end version is finishing the segmentation with the proposed SSPFPN directly, and the separated version is extracting the features with the proposed SSPFPN (ignore its softmax layer) and completing the segmentation by support vector machine.

4.6 Experiment Results and Analysis

This section details the experiment results on three public and challenging datasets: PASCAL VOC 2012, NYUDv2, and SIFT Flow. The results on three datasets are shown in Table 4, and the visualized result of PASCAL VOC 2012 is shown in Figure 7. From Table 4 and Figure 7, the experiment results can be concluded as follows.

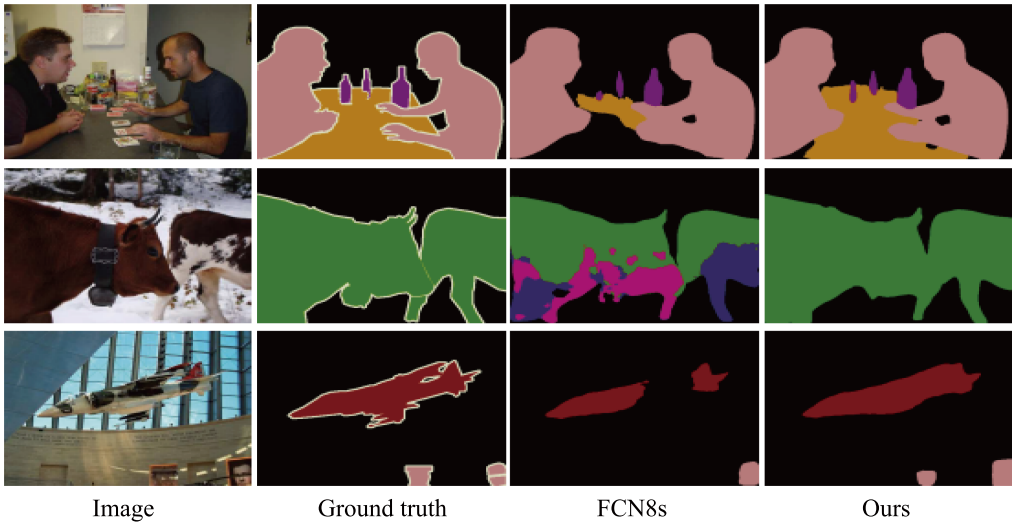

Fig. 7. Some visualized results on PASCAL VOC 2012 Dataset (separated version of the proposed method). Compared to FCN8s, the proposed method can partition the small objects more accurately, for example, the bottles on the desk.

4.6.1 Experiment Results. Table 4 shows the quantitative comparison results of 10 methods on three public and challenging datasets, measured by Pix acc, Mean acc, Mean IU, and *F.w.* IU, respectively. As can be seen, both the end-to-end and separated versions of the proposed method outperforms most comparison approaches in terms of Mean IU. For example, the separated version of the proposed method obtains 3.6% performance improvement on PSCAL VOC 2012 dataset, compared with DeepLab-v2. Besides that, interestingly, the improvements are also obvious on NYUDv2 dataset (3.7 points). Therefore, the proposed SSPFPN is also adapted to the RGB-D image for semantic segmentation, because it considers more structure information of the image, and the structure information of RGB-D image is more significant compared to the RGB image.

Additionally, the separated version of the proposed method achieves a significant performance improvement in terms of Mean acc. For instance, it obtains 1.7% and 1.1% improvements on NYUDv2 and SIFT Flow datasets, respectively, compared with the DeepLab-v3. The reason includes three aspects: The first is that the representation of the image by the proposed SSPFPN is effective; the second is the classification capacity of the support vector machine is significant; and the third is that the separated version of the proposed method considers semantic segmentation as a pixel-based classification task, which avoids the limitation of the down-/up-sampling.

Finally, both two versions of the proposed method improve the performance in terms of *F.w.* IU. For example, the end-to-end version achieves 1.8% improvement on the challenging PASCAL VOC 2012 dataset, compared with DeepLab-v3. The reason is that the relationships among different pixels are preserved to the predicted category label map, which considers the context information of the image adequately.

4.6.2 The Influence of Two Proposed Components. Because the proposed SSPFPN includes two important components: architecture and loss function. To verify that both components are useful for the improvement of the semantic segmentation accuracy, we design two contrast experiments: (1) Based on the proposed SSPFPN architecture, compare the performances of the methods with or without the proposed loss term. (2) Based on the traditional softmax loss function, compare the performances of the methods of FPN or the proposed SSPFPN architecture. The first and second

Table 5. The Influence of the Proposed Loss Term (%)

Method	Pix acc.	Mean acc.	Mean IU	*F.w.* IU
without(pLt)	93.2	79.8	71.4	89.2
with(pLt)	**93.6**	**80.5**	**74.1**	**89.9**

Table 6. The Influence of the Network Architecture (%)

Method	Pix acc.	Mean acc.	Mean IU	*F.w.* IU
FPN	91.6	77.3	69.5	87.9
SSPFPN	**93.2**	**79.8**	**71.4**	**89.2**

contrast experiments are used to verify the effectiveness of the proposed loss term and the proposed network architecture, respectively. These two experiments are tested on the PASCAL VOC 2012 dataset. The results of two contrast experiments are shown in Table 5 and Table 6. pLt in Table 5 represents the proposed loss term.

From Table 5, we can see that when the proposed loss term is added to the traditional softmax loss to optimized the SSPFPN, all evaluation metrics are improved. This is because the similarity of two proposed affinity matrixes enhances pixel-level relationships of the final predicted category label map and improves the segmentation accuracy furthermore.

As shown in Table 6, specific to semantic image segmentation task, the performance of the proposed SSPFPN surpasses that of FPN. The reason is that FPN uses a simple plus operation to process the feature cubes with same size, and this may leads to information loss. The proposed SSPFPN stacks the feature cubes with same size first and then use convolution kernels to fuse them, and this avoids the information loss to a certain extent through a data-driven strategy.

These two contrast experiments verify both the proposed architecture and the proposed loss term are useful for the semantic segmentation task.

4.6.3 Experiment Analysis. Generally speaking, the experiments on three datasets show similar results:

(1) **The experiment result with FCN16s method is better than that of FCN32s, while FCN8s has achieved more satisfactory performance than FCN16s.** In three methods, FCN32s uses the feature maps from the last convolutional layer to generate a label mask of image and then upsamples the label mask to 32× as the final segmentation result. While FCN16s and FCN8s use more feature maps from lower layers besides those of last convolutional layers and upsample their label masks to 16× and 8×, respectively. As is known, features from deep layers of CNNs have significant discriminative capacity while that from lower layers consist of abundant spacial information. Using more features from lower layers can enhance the spatial information of image and achieve more satisfactory performance.

(2) **The Deconv method has achieved more satisfactory performance than FCN8s.** Deconvolution network uses a encoder-decoder architecture to gradually enlarge the size of feature map, it has used more detail location information of pixels in the image. Through noting the location information when down sampling, Deconv uses them adequately when up sampling, and this remains the structure information of the image to a certain extent.

(3) **The end-to-end version of proposed method has achieved more satisfactory performance than Deconv.** Besides more lower feature maps with different sizes are fused by SSPFPN as the representation of the image, the proposed SSPFPN remains the spacial structure information of the predicted label map through a new loss function term. According to this proposed loss term, relationships among pixels in the input image are preserved in the predicted semantic category label map and further ensures the spatial structure of the predicted label map.
(4) **The features extracted with the proposed SSPFPN is good for semantic image segmentation.** Even though the features and classifier are trained separately, the proposed method also has achieved satisfactory performance. This is due to the effectiveness of the features; furthermore, it verifies the strong feature fused capacity of the proposed SSPFPN.

In general, the proposed method achieves satisfactory performance for semantic image segmentation. Additionally, the proposed SSPFPN has strong feature fused capability, and it can be generalized to other tasks referred pixel-level prediction conveniently.

5 CONCLUSION

In this work, SSPFPN is proposed to the task of semantic image segmentation, and achieves satisfactory performance on three public and challenging datasets. The key idea is to take full advantages of each layer's feature maps to depict the image accurately. Both the high-level semantic information and low-level spatial information are considered, and the spatial resolution of the final prediction is improved. Additionally, a novel loss term is incorporated into the framework, which is to remain the relationships among different pixels in the category label map. Finally, through comparing two versions of the proposed methods, we find that the proposed SSPFPN has strong capability in extracting the structural details of images. In general, the proposed method is effective.

REFERENCES

A. E. Abdel-Hakim and A. A Farag. 2006. CSIFT: A SIFT descriptor with color invariant characteristics. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1978–1983.

Eduardo Aguilar, Beatriz Remeseiro, Marc Bolaos, and Petia Radeva. 2018. Grab, pay and eat: Semantic food detection for smart restaurants. *IEEE Trans. Multimedia* 20, 12 (2018), 3266–3275.

Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick. 2016. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2874–2883.

Carlos Castillo, Soham De, Xintong Han, Bharat Singh, Abhay Kumar Yadav, and Tom Goldstein. 2017. Son of Zorn's lemma: Targeted style transfer using instance-aware semantic segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*. IEEE, 1348–1352.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2018a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 4 (2018), 834–848.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV'18)*. 801–818.

Jifeng Dai, Kaiming He, and Jian Sun. 2015. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. (2015), 1635–1643.

N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. 886–893.

Clément Dechesne, Clément Mallet, Arnaud Le Bris, and Valérie Gouet-Brunet. 2017. Semantic segmentation of forest stands of pure species combining airborne lidar data and very high resolution multispectral imagery. *ISPRS J. Photogram. Remote Sens.* 126 (2017), 129–145.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. DeCAF: A deep convolutional activation feature for generic visual recognition. *Comput. Sci.* 50, 1 (2013), 815–830.

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Comput. Sci.* (2014), 2366–2374.

Pedro F. Felzenszwalb, Ross B. Girshick, David Mcallester, and Deva Ramanan. 2010. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 9 (2010), 1627.

Yaroslav Ganin and Victor Lempitsky. 2014. N^4 *-Fields: Neural Network Nearest Neighbor Fields for Image Transforms.* Springer International Publishing. 536–551.

Golnaz Ghiasi and Charless C. Fowlkes. 2016. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *Proceedings of the European Conference on Computer Vision.* 519–534.

Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. 2013. Perceptual organization and recognition of indoor scenes from RGB-D images. In *Computer Vision and Pattern Recognition.* 564–571.

Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. 2014. Learning rich features from RGB-D images for object detection and segmentation. 8695 (2014), 345–360.

Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. 2011. Semantic contours from inverse detectors. In *Proceedings of the International Conference on Computer Vision.* 991–998.

B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. 2015. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 447–456.

Yang He, Wei Chen Chiu, Margret Keuper, and Mario Fritz. 2017. STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling. (2017).

Andrew Holliday, Mohammadamin Barekatain, Johannes Laurmaa, Chetak Kandaswamy, and Helmut Prendinger. 2017. Speedup of deep learning ensembles for semantic segmentation using a model compression technique. *Computer Vision and Image Understanding* (2017).

Sina Honari, Jason Yosinski, Pascal Vincent, and Christopher Pal. 2016. Recombinator networks: Learning coarse-to-fine feature aggregation. In *Computer Vision and Pattern Recognition.* 5743–5752.

Seunghoon Hong, Junhyuk Oh, Honglak Lee, and Bohyung Han. 2016. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3204–3212.

Seunghoon Hong, Tackgeun You, Suha Kwak, and Bohyung Han. 2015. Online tracking by learning discriminative saliency map with convolutional neural network. In *Proceedings of the International Conference on Machine Learning.* 597–606.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2261–2269.

Jia, Yangqing, Shelhamer, Evan, Donahue, Jeff, Karayev, Sergey, Long, and Jonathan. 2014. Caffe: Convolutional architecture for fast feature embedding. *eprint arxiv* (2014), 675–678.

Michael Kampffmeyer, Arnt Borre Salberg, and Robert Jenssen. 2016. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 680–688.

Byeongkeun Kang, Yeejin Lee, and Truong Q. Nguyen. 2018. Depth adaptive deep neural network for semantic segmentation. *IEEE Trans. Multimedia* 20, 9 (2018), 2478–2490.

Ronald Kemker, Carl Salvaggio, and Christopher Kanan. 2017. High-resolution multispectral dataset for semantic segmentation. (2017).

Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. 2016. HyperNet: Towards accurate region proposal generation and joint object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 845–853.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems.* 1097–1105.

Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. 2016. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 3194–3203.

Tsung Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference and Computer Vision and Pattern Recognition.* 2117–2125.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng Yang Fu, and Alexander C. Berg. 2016. SSD: Single shot MultiBox detector. In *Proceedings of the European Conference on Computer Vision.* Springer, 21–37.

Wei Liu, Andrew Rabinovich, and Alexander C. Berg. 2015. ParseNet: Looking wider to see better. *Comput. Sci.* (2015).

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 79, 10 (2015), 1337–1342.

Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*. Springer, 483–499.

Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. 2016. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1520–1528.

Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. 2016. ENet: A deep neural network architecture for real-time semantic segmentation. (2016).

Pedro O. Pinheiro, Tsung Yi Lin, Ronan Collobert, and Piotr Dollár. 2016. Learning to refine object segments. In *Proceedings of the European Conference on Computer Vision*. Springer, 75–91.

D. Ravi, H. Fabelo, G. M. Callico, and G. Yang. 2017. Manifold embedding and semantic segmentation for intraoperative guidance with hyperspectral brain imaging. *IEEE Trans. Medical Imag.* 36, 9 (2017), 1845–1857.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 234–241.

Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann Lecun. 2013. OverFeat: Integrated recognition, localization and detection using convolutional networks. *eprint Arxiv* (2013).

Laura Sevillalara, Deqing Sun, Varun Jampani, and Michael J. Black. 2016. Optical flow with semantic segmentation and localized layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3889–3898.

Hengcan Shi, Hongliang Li, Fanman Meng, Qingbo Wu, Linfeng Xu, and King N. Ngan. 2018. Hierarchical parsing net: Semantic scene parsing from global scene to objects. *IEEE Trans. Multimedia* 20, 10 (2018), 2670–2682.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from RGBD images. In *Proceedings of the European Conference on Computer Vision*. 746–760.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *Comput. Sci.* (2014).

Nasim Souly, Concetto Spampinato, and Mubarak Shah. 2017. Semi and weakly supervised semantic segmentation using generative adversarial network. (2017).

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Computer Vision and Pattern Recognition*. 1–9.

Joseph Tighe and Svetlana Lazebnik. 2010. SuperParsing: Scalable nonparametric image parsing with superpixels. *Int. J. Comput. Vis.* 101, 2 (2010), 352–365.

Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. 2018. Understanding convolution for semantic segmentation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV'18)*. IEEE, 1451–1460.

Huaxin Xiao, Jiashi Feng, Yunchao Wei, Maojun Zhang, and Shuicheng Yan. 2018. Deep salient object detection with dense connections and distraction diagnosis. *IEEE Trans. Multimedia* 20, 12 (2018), 3239–3251.

Wei Xu, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1 (2013), 221–231.

Robail Yasrab. 2017. DCSeg: Decoupled CNN for classification and semantic segmentation. In *Proceedings of the IEEE Sponsored International Conference on Knowledge and Smart Technologies*.

Hao Zhou, Jun Zhang, Shuohao Lei, Jun, and Dan Tu. 2016. Image semantic segmentation based on FCN-CRF model. In *Proceedings of the International Conference on Image, Vision and Computing*. 9–14.

Received August 2018; revised January 2019; accepted March 2019