Neurocomputing 284 (2018) 27–35

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Contour-aware network for semantic segmentation via adaptive depth

Zhiyu Jiang[a,b], Yuan Yuan[a], Qi Wang[c,*]

[a] Center for OPTical IMagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi 710119, PR China
[b] University of Chinese Academy of Sciences, Beijing 100049, PR China
[c] School of Computer Science, and Center for OPTical IMagery Analysis and Learning (OPTIMAL), and Unmanned System Research Institute (USRI), Northwestern Polytechnical University, Xi'an 710072, PR China

ARTICLE INFO

Article history:
Received 6 December 2017
Revised 25 December 2017
Accepted 4 January 2018

Communicated by Dr. Nianyin Zeng

Keywords:
Semantic segmentation
Scene parsing
Contour
CRF
Adaptive depth

ABSTRACT

Semantic segmentation has been widely investigated for its important role in computer vision. However, some challenges still exist. The first challenge is how to perceive semantic regions with various attributes, which can result in unbalanced distribution of training samples. Another challenge is accurate semantic boundary determination. In this paper, a contour-aware network for semantic segmentation via adaptive depth is proposed which particularly exploits the power of adaptive-depth neural network and contour-aware neural network on pixel-level semantic segmentation. Specifically, an adaptive-depth model, which can adaptively determine the feedback and forward procedure of neural network, is constructed. Moreover, a contour-aware neural network is respectively built to enhance the coherence and the localization accuracy of semantic regions. By formulating the contour information and coarse semantic segmentation results in a unified manner, global inference is proposed to obtain the final segmentation results. Three contributions are claimed: (1) semantic segmentation via adaptive depth neural network; (2) contour-aware neural network for semantic segmentation; and (3) global inference for final decision. Experiments on three popular datasets are conducted and experimental results have verified the superiority of the proposed method compared with the state-of-the-art methods.

1. Introduction

Semantic segmentation, which can be applied to still images, videos, or even 3D hyperspectral data, has been widely investigated in computer vision and machine learning areas for it can help achieve deep understanding of regions, objects, and scenes. Concretely, semantic segmentation tends to make dense predictions so that each pixel can be labeled with the class of enclosing region or object [1]. Semantic segmentation is highlighted by the fact that it can provide abundant semantic information for mid-level and high-level tasks, such as behavior analysis, abnormal detection, scene understanding, and autonomous driving [2–7]. Semantic segmentation has been addressed by traditional models in the past decades and great progress has been made. Traditional approaches, which do not take Convolutional Neural Networks (CNNs) into consideration, mainly focus on domain knowledge and decision strategy. Generally, the choice of features plays an important role in traditional approaches, including local and global features. Pixel colors in different images spaces [8–10] and gradient features [11–13] are widely considered for their intuitive and straightforward properties. Besides, segmentation methods [10,14–18] are another way to utilize domain knowledge and they tend to detect consistent regions or region boundaries. What's more, decision models considering contextual information result in significant improvements, such as Markov Random Fields (MRFs) [19] and Conditional Random Fields (CRFs) [20].

Despite the high popularity of those traditional models, the deep architectures, which can be usually regraded as CNNs, are showing distinct superiority for the ability of learning representations in an end-to-end manner instead of using hand-crafted features that require domain expertise [1]. Currently, the most successful state-of-the-art deep learning techniques for semantic segmentation is Fully Convolutional Network (FCN) [21] and its varieties based on famous classification models, including AlexNet, VGG, GoogLeNet, and ResNet [21–23]. Recently, inspired by the FCN architecture, other deep models are developed to make it suitable for segmentation, such as SegNet [24], dilated convolution net [23], DeepLab model [25], CRF as RNN [26], skip connections net [27], and ParseNet [28].

* Corresponding author.
E-mail address: crabwq@gmail.com (Q. Wang).

https://doi.org/10.1016/j.neucom.2018.01.022
0925-2312/

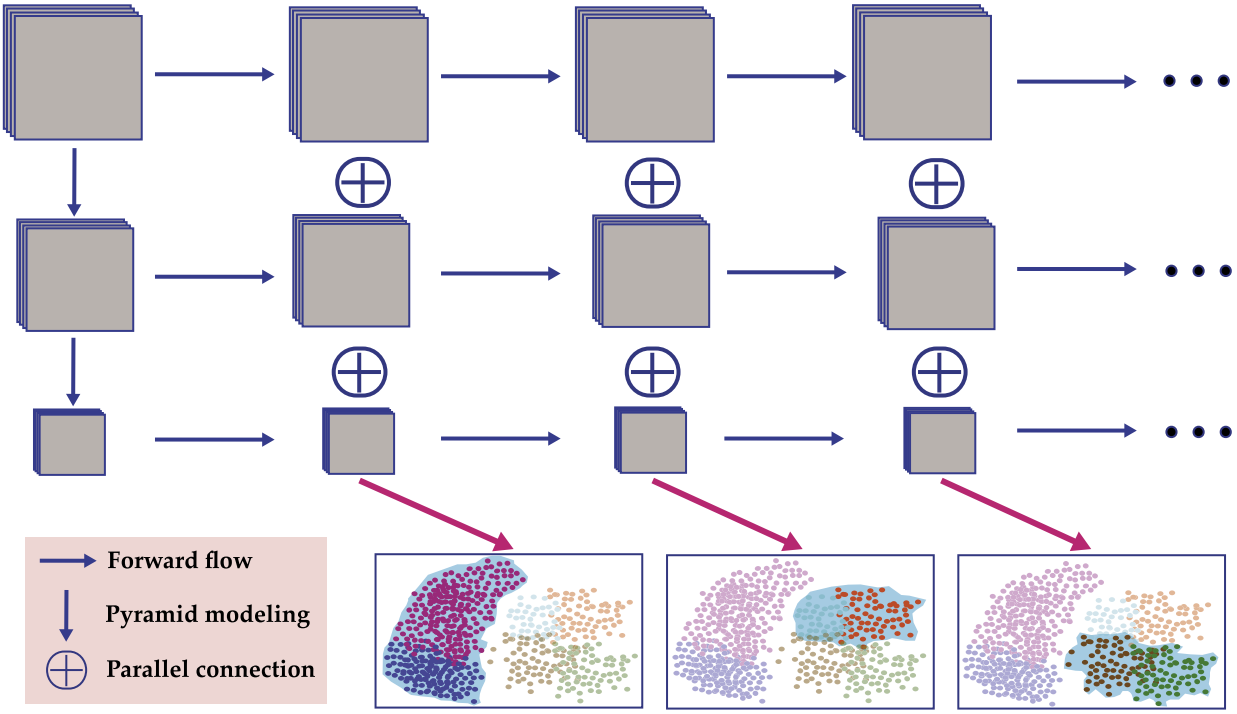


Fig. 1. Adaptive-depth neural network for semantic segmentation. The horizontal arrow indicates the forward flow of the neural network, the vertical arrow indicates the pyramid modeling of the input image, and the symbol ⊕ represents upsampling and parallel-connected operations. Under the pyramid modeling, each vertical feature maps are parallel-connected after upsampling. Moreover, each semantic class can adaptively determine the specific decision layer where only small number of samples are classified.

Although lots works have been done and great progress has been made on semantic segmentation, some challenges still exist. The first one challenge which makes semantic segmentation difficult is the different perceptual complexities of semantic regions with various attributes. This phenomenon can be explained in two aspects. The first aspect lies in the *thing* and *stuff*. The *thing* is a semantic object with specific size and shape which is also known as foreground object, such as car, pedestrian, and traffic sign. Contrarily, the *stuff* is semantic region which is defined by a homogeneous or repetitive pattern with no specific spatial shape and it can be regarded as background object, such as road, building, and sky. These two kinds of semantic objects make the semantic segmentation model ambiguous for the various attributes. Another aspect is the unbalanced distribution of training samples, especially for background and foreground objects. Generally, background region tends to occupy large area and large number of training samples will be obtained in the manner of pixel-wise labeling. While the foreground object is on the contrary. These facts all make the simultaneous perception difficult and different perceptual procedures are essential for background and foreground object segmentation due to various attributes.

Another challenge is the accurate boundary determination of semantic regions. Due to the power of the deep neural networks, high-level features show a strong representation for semantic region. However, spatial information is largely eliminated due to pooling operation especially for deep CNNs, such as GoogleNet [29] and ResNet [30]. Since pooling is the inborn defect of CNN models for semantic segmentation [31] and it weakens the details of image features which are useful for accurate semantic segmentation, one intuitive method to overcome this shortage is removing pooling in deep network. Actually this intuitive way would result in shrinking receptive field for each neuron and the context information would also be lost. Another drawback of eliminating pooling is the dramatical increase of time complexity when the feature maps get large. Edges and contours are important for accurate segmentation since they give detail information. Therefore edge extractors obtaining the detail information is necessary. We propose to address these challenges by means of contour-aware network for scene parsing via adaptive depth, which takes contour detection neural network into consideration to determine the boundary of semantic regions. And a segmentation network through adaptive depth is also employed to address the different perceptual difficulties on semantic regions with various attributes as shown in Fig. 1. Experimental results indicate that even though the coarse segmentation net is not effective enough, contour detection net can increase the accuracy for better semantic segmentation performance. Overall, the main contributions of this work are summarized and explained as follows:

(1) *Semantic segmentation via adaptive depth network.* Traditional methods tend to tackle samples with different semantic labels in the same procedure. However, different semantic regions with various attribute will make a simplex model confusion. This phenomenon is especially obvious for background and foreground regions. In this work, an adaptive-depth semantic segmentation model is proposed which can adaptively determine the feedback and forward neural network layer.

(2) *Contour-aware neural network for semantic segmentation.* For semantic segmentation, little previous works take the contour information into consideration. In this work, a simple and efficient contour detection model is proposed and the contour information is formulated as similarity value through an intuitive method. The semantic segment coherence is enhanced and the localization of semantic regions is also improved through contour detection.

(3) *Global inference for final decision.* It is difficult to transform contour information and coarse semantic segmentation results into a unified viewpoint for the contour line is not closed. In this work,

both contour information and coarse semantic segmentation results are transformed into similarity values and CRF is served as a global inference model for the final decision.

The remainder of this paper is organized as follows. Firstly, Section 2 introduces the related semantic segmentation methods in recent years. The formulation of the proposed method is described in detail in Section 3. Section 4 demonstrates the experimental setup and experimental results are also analyzed in this part. Finally, conclusions are drawn in Section 5.

2. Related works

Before presenting the proposed method, we first review the traditional methods for semantic segmentation and the recent CNN based methods are also been discussed in more detail.

Traditional semantic segmentation methods adopt domain knowledge to learn the representation. Owing to different emphases, three aspects are discussed in the following. Firstly, the choice of features is very important for traditional approaches. Colors in different image spaces are mostly considered. Kasson and Plouffe [9] measured the performance of different color spaces and Cheng et al. [8] summarized some major color representations. Moreover, the statistic color information is also considered in [10,32]. Another import feature is the gradient feature for its illumination invariance, such as SIFT [11] and HOG [12]. Secondly, clustering or segmentation is another way to take domain information into consideration. Chen et al. [14] applied *k*-means for medical image segmentation and mean shift proposed by Comaniciu and Meer [33] was also utilized for segmentation by Zhang et al. [15]. Furthermore, Carreira and Sminchisescu [34] proposed a graph-based method, which typically interpreted pixels as vertices and an edge weights as measure of dissimilarity, and some other methods, such as active contour [16] and watershed segmentation [17], were also considered for automatic object segmentation. Thirdly, decision models were widely analyzed as well. Markov Random Fields (MRFs) are wide-spread model in computer vision and Liu et al. [19] associated undirected graph of an MRF with semantic segmentation problem. Moreover, Vemulapalli et al. [20] proposed Conditional Random Field (CRF) for segmentation considering the contextual constraints.

The recent successful methods for semantic image segmentation are mostly based on CNNs. CNNs are Artificial Neural Networks (ANNs), which are inspired by biologic neurons, and they can drastically reduce the number of parameters while being still general enough for image processing. At first, CNN-based methods are region-based methods. The region proposals are first generated and then assign semantic labels to each of them. Girshick et al. [35] utilized bottom-up region proposals and domain-specific fine-tuning for semantic segmentation. However, region-based methods would result in inaccuracy for pixel-level semantic segmentation. Recently, FCNs have become popular for end-to-end training. Long et al. [21] utilized fully convolution operation instead of fully connected network for semantic segmentation by fine-tuning the classification network, such as VGG-16 model [36] which takes advantages of the large ImageNet dataset. Nevertheless, the resolution of the output feature map is down-sampled due to convolution and pooling layers. One naive method is directly reducing the strides for all layers. Although this strategy can alleviate the problem in a certain aspect, it will dramatically increase the computational complexity and the receptive field is also reduced which makes the model unable to capture high-level semantic information. To address this down-sampling problem, a variety of FCNs methods were proposed recently which focused on obtaining high resolution even pixel-level semantic segmentation results. Chen et al. [25] first utilized atrous convolution to enlarge the receptive field without increasing the computation complexity and dense CRF was also considered to refine the object boundary. Zheng et al. [26] regraded the mean field CRF inference as recurrent layers for end-to-end learning of the dense CRF and FCN network. Noh et al. [37] learned a multi-layer deconvolution network to explore the shape information and detail structures by reconstructing the original size segmentation maps from deep and small feature maps step by step.

Although significant progress has been made for semantic segmentation, some challenging problems still exist. Firstly, accurate boundary determination can improve the performance of semantic segmentation a lot and how to tackle this problem is still challenging. Previous methods tend to increase the resolution of the output feature map [25,37] to obtain accurate semantic boundary. However, this strategy will result in high computation complexity and intentional boundary detection is one way to alleviate this challenge. Secondly, different semantic regions show various attributes which makes the simplex model confusion. This phenomenon is extremely obvious for background and foreground regions for the difference of the sample numbers. Previous works are likely to utilize data argumentation to reduce the impact. For tackling these difficulties, a contour detection neural network is considered to obtain accurate boundary determination. Furthermore, a semantic segmentation neural network via adaptive depth is proposed which can handle background and foreground semantic regions in different depth.

3. Contour-aware network for semantic segmentation via adaptive depth

In this section, the basic semantic segmentation model is first introduced, including pyramid CNN modeling, fully connected CRF construction and final decision for pyramid results. Subsequently, the semantic segmentation neural network via adaptive depth is described in detail. Furthermore, a contour-aware neural network is also introduced. Finally, the global inference procedure is formulated. The pipeline of the proposed method is illustrated in Fig. 2.

Semantic segmentation tends to assign semantic labels to each pixel and various semantic labels are defined in different dataset. Generally, the semantic labels can be categorized into two kinds. The first kind is the foreground object which have a specific shape prior, such as the car. This kind of object mostly occupies small area in image space and small number of training samples will be obtained. Another kind is the background region which is defined by repetitive pattern and no specific spatial extent, such as sky. This kind of region is in the majority and the number of training samples tend to be large compared with the first kind.

Two disadvantages are produced due to the facts mentioned above. The unbalance distribution of training samples is the first disadvantage and the learned model tends to ignore the small-sample classes and the test samples are more likely to be classified as large-sample classes. The other problem is the different perceptual complexities for foreground and background semantic objects. For example, textural features may help a lot for background classification while the shape information plays a key role for foreground objects determination. The feature maps from different depth play different roles for semantic segmentation and it is necessary to build a model which can adaptively select the proper level feature. To alleviate these problems, a semantic segmentation neural network via adaptive depth is proposed in this part and the implementation details are described in the following part.

3.1. Basic semantic segmentation modeling

Before introducing the adaptive depth neural network for semantic segmentation, the basic semantic segmentation model

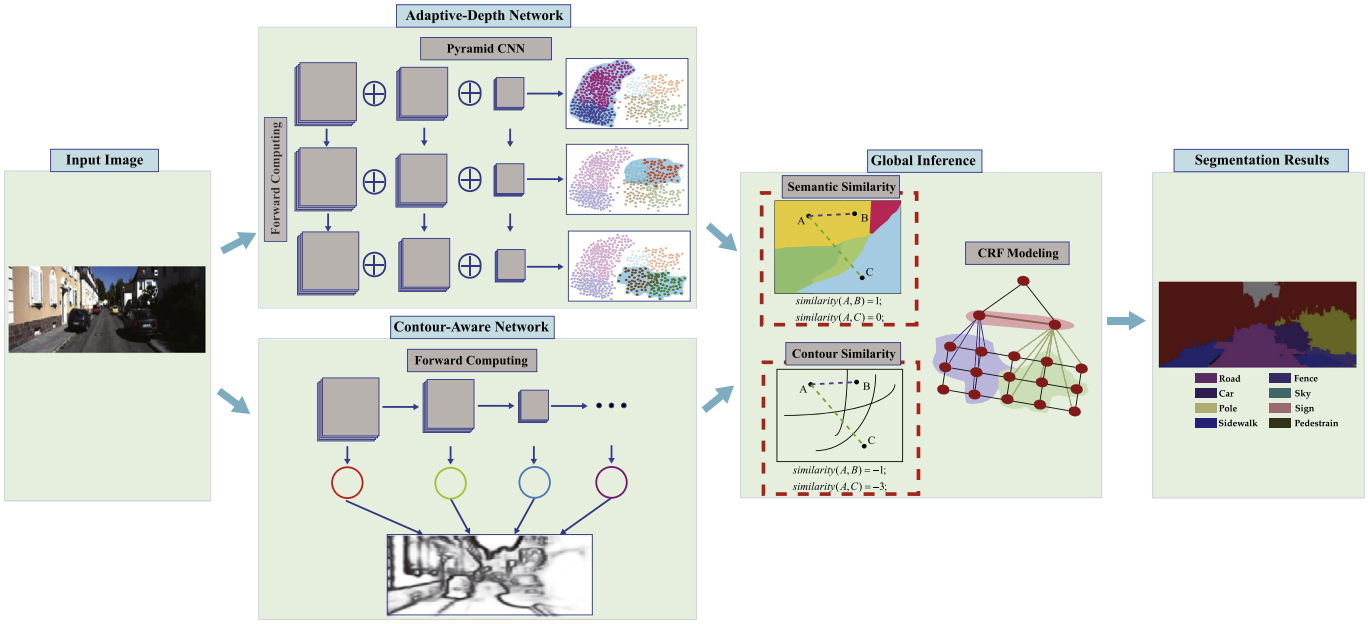


Fig. 2. Semantic segmentation pipeline. An adaptive-depth neural network is built to obtain the coarse semantic segmentation results. Simultaneously, the contour information is inferred through a contour-aware network. Furthermore, both the coarse semantic information and contour information are modeled in the same manner. Finally, the semantic labels are obtained through global inference based on CRF.

without adaptive depth is described. We adopt the common practice in semantic segmentation and formulate the semantic segmentation task as a discrete energy minimization problem. Specifically, three parts are included to accomplish the semantic segmentation task and the detailed information is introduced in the following section.

3.1.1. Pyramid CNN modeling

For a certain image, redundant information exists for the spatial similarities between nearby pixels and image segmentation is an efficient strategy to reduce the spatial redundancy. In this work, image pyramid is first built for adjusting the receptive field and only down sampling is considered. Specifically, each pyramid image is segmented into fix-sized pathes and small-resolution pyramid image is segmented into small-sized pathes. For each patch, it will forward a CNN model and the pre-trained VGG model [38] is utilized to initialize the CNN model and the last FC-layer is replaced by FC-k layer, where k is the number of semantic classes. Moreover, fine-tune procedure is adopted through different pyramid scale.

3.1.2. Fully connected CRF construction

It is widely believed that CNN model is good at feature learning and it is also necessary to model the correlations between patches and semantic labels. For semantic segmentation, each pixel should be corresponded to a certain semantic label and a certain semantic label can be assigned to any pixel. Based on this formulation, a fully connected CRF model is constructed for the capability of capturing contextual information between nearby pixels and the constructed fully connected CRF model [39] can be defined as

$$E(\boldsymbol{Y}, \boldsymbol{f}) = \sum_i \psi_u(y_i, \boldsymbol{f}) + \sum_{i<j} \psi_p(y_i, y_j, \boldsymbol{f}), \tag{1}$$

where i and j indicate the ith and jth samples. The unary energy $\psi_u(y_i, \boldsymbol{f})$ measures the cost of assigning sematic label y_i to the sample x_i given the features $\boldsymbol{f}$. The pairwise function is defined as in [40]

$$\psi_p(y_i, y_j, \boldsymbol{f}) = \mu(y_i, y_j) \sum_{m=1}^{M} \omega^{(m)} k^{(m)}(\boldsymbol{f_i}, \boldsymbol{f_j}), \tag{2}$$

where the number of Gaussian kernels is M. Each $k^{(m)}$ is a Gaussian kernel depending on pixel feature $\boldsymbol{f}$ and $\omega^{(m)}$ is weighted parameters. The parameter μ is defined as indicating value which is based on whether the semantic labels of y_i and y_j are the same. The object is to minimize the energy function defined in Eq. (1) and the optimal label assignment for all the samples will be determined. Truncated EM method [40] is utilized to solve the Eq. (1) for its good performance.

3.1.3. Final decision for pyramid results

After CRF inferring, how to make decision on the pyramid results is necessary. Intuitively, only small number of semantic labels play key roles for the final decision across pyramid results and a sparse learning model is employed to infer the final results. For pixel x_i, the pyramid feature after nearest interpolation can be written as $\boldsymbol{a}_i$ all the sample features is $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]^T$. The objective equation can be written as

$$\boldsymbol{w}^* = \arg\min \|\boldsymbol{Aw} - \boldsymbol{Y}\|_2^2, s.t. \|\boldsymbol{w}\|_1 \leq \varepsilon, \tag{3}$$

where $\boldsymbol{Y}$ is the semantic labels of all the samples and ε is the residual error and Eq. (3) can be solved by Lasso [41]. The final semantic labels can be obtained by

$$\boldsymbol{y}_i^* = \arg\min_{\boldsymbol{y}_i \in \mathcal{L}} \|\boldsymbol{a}_i \boldsymbol{w}^* - \boldsymbol{y}_i\|_2, i \in [1, \ldots, N]. \tag{4}$$

3.2. Semantic segmentation neural network via adaptive depth

After introducing the basic semantic segmentation model, the semantic segmentation model via adaptive depth is introduced in this part.

For CNN models, as the depth of neural network increases, the response of the feature map tends to be large amplitude for certain semantic regions. However, the resolution of feature map is decreased due to the convolutional and pooling operations and

the detailed information is also eliminated. Meanwhile, different semantic regions are sensitive to different level of feature maps, thus an adaptive depth learning CNN model is proposed. Specifically, the semantic regions corresponding to each semantic label can adaptively determine the depth of forward and feedback computing. The pipeline is illustrated in Fig. 1. The detailed architecture is explained as follows.

For the input image, the pyramid CNN model is firstly built as described in Section 3.1.1 without batch segmentation. Unlike the basic semantic segmentation model, the feature maps from different pyramid scale are up-sampled to the original image size through bilinear interpolation. Simultaneously, the feature maps are connected parallelly and a fully connected CRF as described in Section 3.1.2 is followed. Starting from the 2nd layer of the CNN model, the CRF is trained and the validation dataset is also utilized to test the performance and the top k semantic classes with high accuracy are determined. Subsequently, the 3rd layer of the CNN model and CRF are trained and the top k semantic classes are determined eliminating the k classes determined in the 2nd layer. More importantly, during the training procedure, the samples corresponding to the determined k semantic classes in the 2nd layer are taken into consideration with small weight. The weight value can be changed by tuning the definition of μ in Eq. (2) and the new definition can be written as

$$\mu(y_i, y_j) = \begin{cases} 1 - G(min(n_i, n_j)), & if \quad y_i \neq y_j, \\ 0, & if \quad y_i = y_j, \end{cases} \tag{5}$$

where $G(\cdot)$ is a Gaussian function and n_i is the training sample number of the ith semantic label. Repeating these procedures until the last semantic class are selected and the semantic segmentation via adaptive depth is ready for testing. As for testing procedure, each layer with CRF models can determine k semantic labels and repeat this step for next layer until the last semantic regions are determined.

Two advantages are claimed for the adaptive depth semantic segmentation. Firstly, for a certain semantic region detection, the forward step of the input image can be adaptive based on the semantic label and only small number of semantic region determination need to forward the whole model. This characteristic can decrease the computing complexity efficiently. The other advantage is alleviating the unbalanced problem of training samples. For each CRF model, only the samples with similar attributes are determined. For example, the shallow-layer models tend to determine the background regions and the deep-layer models are likely to analyze foreground objects.

3.3. Contour-aware neural network

For semantic segmentation, high-level contour information can efficiently alleviate the ambiguousness of semantic regions. In this section, the proposed contour-aware neural network is introduced for the final semantic segmentation. The proposed pixel-wise contour detection architecture is first introduced, and then the detailed procedures are also discussed.

Intuitively, the operation of convolution tends to respond to image's edge position in the shallow-layer of neural network. And semantic regions are more likely to correspond to large weight magnitude value in deep-layer of neural network. Based on these facts, the feature map of neural network can be directly adapted to contour detection which can efficiently distinguish semantic regions. Meanwhile, the detailed structural information, such as the shape of semantic objects, is eliminated in deep-layer network for convolution and pooling operations while shallow-layer is the opposite. Consequently, it is necessary to take both the shallow-layer and deep-layer network into consideration for contour detection and a linear combination of all the feature maps. Firstly, the multi-layer features of pixel p can be written as a column descriptor:

$$\boldsymbol{f}(p) = [f_1(p), f_2(p), \ldots, f_L(p)], \tag{6}$$

where $f_n(p), n \in \{1, \ldots, L\}$ is the feature map response of pixel p in the nth layer, L is the number of neural network layer. However, due to the convolution and pooling operations, the resolution of feature maps is decreased as n increases. Consequently, proper interpolation method is essential and bilinear interpolation of each feature map is adopted to adjust the response map to the original pixel resolution. For simplicity, $f_n(p)$ is the response map after interpolation. Subsequently, the linear combination of all the feature maps can be defined as

$$h_{\boldsymbol{W}}(p) = \boldsymbol{W} \times \boldsymbol{f}(p) = \sum_{n=1}^{L} w_n f_n(p), \tag{7}$$

where $\boldsymbol{W}$ is the weight of each layer and a sigmoid cross-entropy loss is utilized to determine the weight value which can be written as

$$J(\boldsymbol{W}) = -\frac{1}{m}\sum_{i=1}^{m} y(p^{(i)}) \log(h_{\boldsymbol{W}}(p^{(i)})) + (1 - y(p^{(i)})) \log(1 - h_{\boldsymbol{W}}(p^{(i)})), \tag{8}$$

where $y(p^{(i)})$ is the true label of pixel $p^(i)$, $y(p^{(i)}) = 1$ indicates the contour position and $y(p^{(i)}) = 0$ for other regions.

For the training procedure, the training samples are generated from semantic segmentation dataset by simply detecting the edges of semantic label image. Considering the balance of training samples, a sparse set of samples are efficiently generated from the original pixel space. Moreover, two advantages of this strategy are concluded. Firstly, nearby pixels in image space are highly correlated and the bilinear interpolation tends to make the nearby responses different while their labels are the same. This phenomenon is extremely obvious for the pixels close to contour positions and these samples will result in making the model confusion. Based on this fact, sampling strategy can efficiently reduce the influence. To ensure a diverse set of training samples, about 2000 pixels are sampled from a single image and the edge positions are sampled in high frequency for the balance of training samples. Secondly, smaller number of training samples of per image results in sampling more images per batch for a certain GPU memory. At each iteration of SGD training procedure, the gradient over the model parameters is computed over a relatively small number of samples from the training set. Consequently, if more images are considered in a SGD iteration, the sample diversity will be more increased and the convergence speed is also accelerated.

3.4. Global inference

In this part, the semantic segmentation results via adaptive depth and the contour results are simultaneously considered for the final global inference. It is necessary to transform these results into a unified viewpoint.

For the contour results as illustrated in Fig. 3, the contour line is not closed and it is impossible to transform it into closed regions. Intuitively, two pixels are similar if there is no contour lines between them. On the contrary, if the straight path between two pixels are crossed with a contour line and then these two pixels are likely to belong to different semantic regions. Moreover, the larger the crossed contour line number, the more dissimilar the two pixels tend to be. Based on this fact, the similarity of pixel i and pixel j can be encoded as follows:

$$w_{i,j}^{ct} = \exp\left(\frac{-L_{i,j}}{\sigma_{ct}}\right), \tag{9}$$

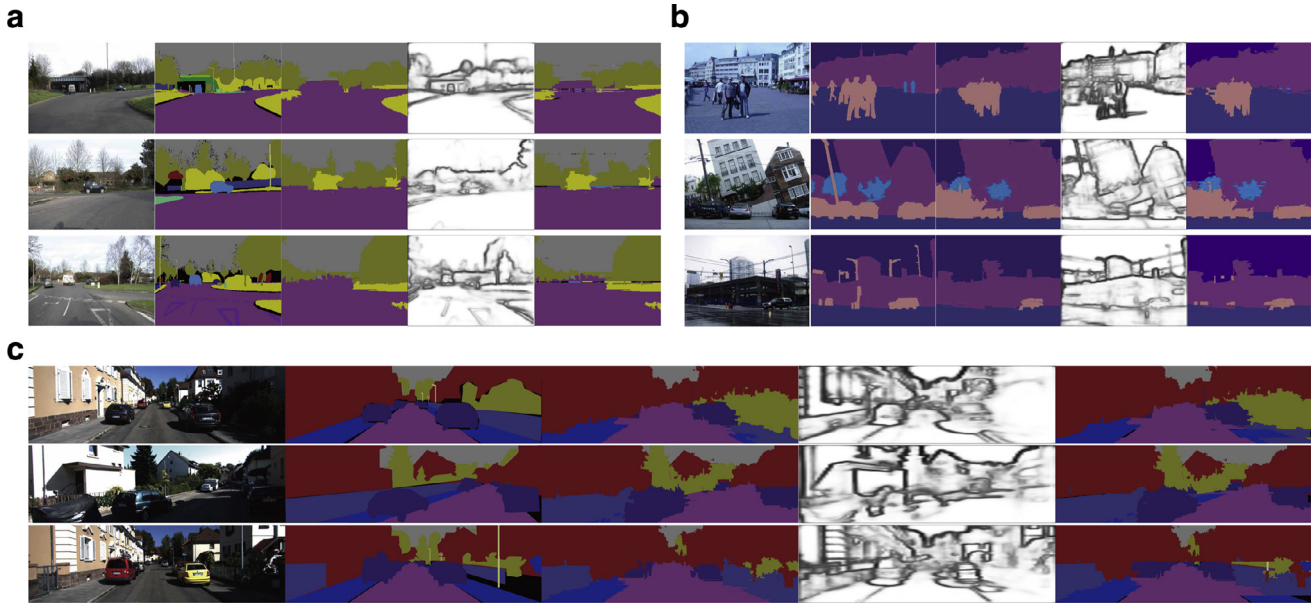


Fig. 3. Qualitative semantic segmentation results. CamVid results are shown in (a) and (b) is stanford-background results. KITTI results are demonstrated in (c). For each dataset results, the first column indicates the input images, the second column provides the groundtruth. The third column shows the semantic segmentation results based on the proposed basic model, the fourth column is the contour detection results and the last column shows the global inference results.

where $L_{i,j}$ is the sum of crossed contour values with the straight line between pixel i and pixel j. And σ_{cont} is the normalized parameter. Similarly, the semantic segmentation results can also been transformed into similarity pattern. Specifically, if two pixels belong to different semantic labels, the similarity $w_{i,j}^{sm} = 0$. Otherwise, $w_{i,j}^{sm} = 1$. The formulation can be written as

$$w_{i,j}^{sm} = \begin{cases} 1, & if \quad y_i = y_j, \\ 0, & if \quad y_i \neq y_j, \end{cases} \tag{10}$$

The final similarity can be defined as

$$w_{i,j} = \frac{k \times w_{i,j}^{ct} + w_{i,j}^{sm}}{\sigma}, \quad k \in (0,1), \tag{11}$$

where k is the hyper-parameter. According to Eq. (11), the importance of contour results is decreased, and this is because the semantic regions with the same semantic labels may be apart and $w_{i,j}^{ct}$ ignores this situation.

For global inference of the final semantic labels, CRF is considered for its global energy function definition and it can also preserve local consistency. Based on the Eqs. (1) and(2), the global inference can be formulates as:

$$\boldsymbol{Y}^* = \arg\min_{\boldsymbol{Y}} \sum_i d_i \left(y_i - \frac{\boldsymbol{f}_i}{d_i} \right)^2 + \sum_{i<j} w_{i,j} (y_i - y_j)^2, \tag{12}$$

where $\mathbf{f}_i$ is the semantic label obtained from the semantic segmentation via adaptive depth, d_i is the degree of pixel i. This equation can also be solved by truncated EM method [40].

4. Experiments

In this section, both quantitative and qualitative results for semantic segmentation on three public datasets are presented. Moreover, two evaluation metrics are also defined and some experimental details are explained subsequently. Finally, thorough analyses on the experimental results are conducted.

4.1. Datasets

Through the years, semantic segmentation has been mostly focused on outdoor images and a lot of semantic segmentation datasets are published. In this section, we will describe three semantic segmentation datasets which are utilized in this work to verify the performance of the proposed method. These three datasets are popular and have attracted much attention for their challenging properties.

CamVid dataset [42] is a road/driving scene understanding dataset which is originally captured as five video sequences with a 960×720 resolution camera. 701 frames are sampled from the sequences and they are manually annotated with 32 classes. Sturgess et al. [43] divided the dataset into 367 training images, 100 validation images and 233 testing images. Moreover, 11 semantic classes are selected from the original semantic classes, including building, tree, sky, car, sign, road, pedestrian, fence, pole, sidewalk, and bicyclist.

Stanford-background dataset [44] with outdoor images are imported from existing public datasets. The dataset contains 715 images with two separate label sets: semantic and geometric. We conduct our experiments for predicting the semantic label only. The semantic classes include seven background classes and a generic foreground class.

KITTI dataset [45] is a large publicly available road scene dataset and some images are extracted and manually annotated for scene parsing. For convenience of the comparison, the labeled images by Ros et al. [46] are utilized as experimental dataset which contains 142 images. Moreover, 11 semantic classes, such as buildings and road, are severely imbalanced distributed.

4.2. Evaluation criteria

For semantic segmentation, most evaluation criteria are focused on pixel accuracy and class accuracy. These two criteria can efficiently evaluate the performance of the proposed method in pixel level and semantic class level. For better understanding, some notations are firstly defined. The total number of semantic classes is

defined as k. n_{ij} is defined as the amounts of pixels which are predicted as class j while their true label is i.

Pixel accuracy indicates the percentage of pixels correctly labeled over all the test pixels without considering the semantic class. It can be written as

$$Pixel\ Accuracy = \frac{\sum_{i=1}^{k} n_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}}. \tag{13}$$

Class accuracy reflects the average percentage of pixel accuracy for every sematic class. Class accuracy focuses on the performance of the proposed method on the semantic class corresponding to small samples. The calculation formula is defined as follows:

$$Class\ Accuracy = \frac{1}{k}\sum_{i=1}^{k}\frac{n_{ii}}{\sum_{j=1}^{k} n_{ij}}. \tag{14}$$

4.3. Implementation details

For evaluating the contributions of the proposed method, two models are firstly defined in this part and some implementation details are also described in this part.

- *Basic model:* As described in Section 3.1, the basic semantic model consists of pyramid CNN modeling, fully connected CRF construction and final decision procedure. This model is served as a baseline and some details can be found in [56].
- *Adaptive-depth model:* As described in Section 3.2, this model can adaptively learn the feedback and forward flow of the neural network. Specifically, the last two layers (FC layer is not counted) of VGG model [38] is utilized to learn the adaptive procedure for this model due to the fact that the depth of VGG model is relatively small and deeper-layer tends to have better performance. Moreover, the parameter k is set as the half number of semantic classes.
- *Global inference model:* As described in Section 3.4, this model utilizes a unified framework which can simultaneously model the coarse semantic results and the contour information under global inference.

4.4. Performance analysis

Both qualitative and quantitative results are shown in this work. Typical scene parsing results on the three popular datasets are presented in Fig. 3. Intuitively, the contour information focuses on semantic boundaries and the global inference model achieves the best. For a more objective comparison, both pixel accuracy and class accuracy defined in Section 4.2 are considered. And the quantitative results are shown in Tab. 1. Although the basic model is a little weak on the class accuracy, it is clear that the global inference model achieves the highest scores. In the following part, a more detailed analyses on the three datasets are presented.

Basic model performance. From Table 1, the basic model has good performance when taken the pixel accuracy as the evaluation criterion. For example, Although recursive neural network model [52] and recurrent neural network model [53] can efficiently take the contextual constraints into account, the basic model performs better which takes adequate contextual information from both pyramid modeling and probabilistic graphical model construction into consideration. Specifically, the pyramid model takes hierarchical inferred labels into consideration and the final decision is based on sparse learning. Moreover, the probabilistic graphical model is built focusing on both the local and global contexts.

However, the basic model has shown its weakness on the aspect of class accuracy. The reason is that the basic model is based on sampling certain number of patches from Gaussian pyramid. And this strategy would ignore the small-sized semantic classes. For example, for the KITTI dataset, the number of pixels defines as pole [46] is very small and nearly zero number of pixels are correctly labeled based on the basic model. To alleviate this problem, an adaptive-depth model is proposed and the detailed analyses are described as follows.

Adaptive-depth model performance. The adaptive-depth model has great improvements on class accuracy. For example, nearly 5% improvements are obtained compared with the basic model on the KITTI dataset. The adaptive-depth strategy helps a lot for the good performance. This is because the samples corresponding to a certain semantic label can adaptively choose the proper feature layer. Besides, the samples are also parted due to adaptive-depth strategy and the small-sample problem is largely alleviated.

However, the performance of the adaptive-depth model is a little degraded for the pixel accuracy. For example. the adaptive-depth model decreased about 2% on the Stanford dataset. Conversely, the class accuracy is increased about 3%. This is because the samples with small-sized semantic labels are correctly labeled.

Global inference model performance. For simultaneously increasing the pixel accuracy and class accuracy, both the contour information and adaptive-depth semantic segmentation results are modeled in a global inference manner based on energy minimization. For better understanding this model, the performance is analyzed on the three datasets, respectively.

CamVid dataset. The images are sampled from two daytime and one dusk sequences. The first block of Table 1 shows the performance of the proposed method compared with state-of-the-arts. It is obvious that the performance of the global inference model is well considering both pixel accuracy and class accuracy. For example, the appearance model [47] and the local labeling method [50] perform worse in the dust sequences for their low-level feature representation. On the contrary, our work exploits the power of CNN model and Gaussian pyramid strategy, and adequate contextual information is utilized to improve the performance of the basic model. In addition, the CRF method [43] performs well when considering the class accuracy criteria. The proposed method takes advantage of the CRF model and takes different levels of the features into consideration which leads to higher pixel accuracy. On the other hand, the proposed model is better than the basic model considering both average accuracy and class accuracy. This phenomenon shows that the global inference can efficiently improve the accuracy of the small samples with a little loss on the normal semantic class.

Stanford-background dataset. The second block of Table 1 shows the superiority of the global inference method. For example, recursive neural network model [52] and recurrent neural network model [53] can efficiently take the contextual constraints into account on the structure of the models. Moreover, the global inference model shows priority in two aspects. Firstly, contour information can efficiently enhance the semantic segment coherence and accurately locate the boundary of semantic regions. Secondly, taking the advantage of adaptive-depth model, the global inference model can properly handle the small-sample learning problem.

KITTI dataset. This dataset is captured with wide viewing-angle and it is sampled from videos under a certain frequency. Moreover, the semantic label is imbalanced distributed and the long-tail phenomenon is obvious. Addressing these difficulties, temporal constraint is considered by Ros et al. [46] and high class accuracy verifies the effectiveness of the temporal information. On the contrary, temporal context information does not take into account in our method. The competitive results on the pixel criterion show the superiority of the proposed method. Compared with the basic model, the global inference model shows good performance on both evaluation criteria.

From pixel accuracy and class accuracy, the proposed method have achieved good performance. Moreover, the contributions of

Table 1
Quantitative semantic segmentation results, including pixel accuracy and class accuracy (%). The bold numbers represent the best-3 scores.

Dataset	Approach	Pixel accuracy	Class accuracy
CamVid	SFM+appearance [47]	69.1	53.0
	Boosting [43]	76.4	**59.8**
	Structured random forests [48]	72.5	51.4
	Local label descriptors [49]	73.6	36.3
	Boosting+pairwise CRF [43]	79.8	59.9
	Local labeling+MRF [50]	77.6	43.8
	Basic model (ous)	**81.1**	49.9
	Adaptive-depth model (ous)	**81.5**	**60.1**
	Global inference model (ours)	**81.7**	**60.2**
Stanford	Stacked labeling [51]	76.9	66.2
	Recursive neural networks [52]	78.1	N/A
	Recurrent neural networks [53]	80.2	69.9
	Hierarchical features [54]	**81.4**	**76.0**
	WAKNN+MRF [55]	74.1	62.2
	Basic model (ous)	**81.7**	70.6
	Adaptive-depth model (ous)	79.8	**73.4**
	Global inference model (ours)	**82.5**	**76.2**
KITTI	Temporal semantic segmentation [46]	51.2	**61.6**
	Semantic segmentation retrieva [46]	47.1	58.0
	Basic model (ous)	**79.8**	45.84
	Adaptive-depth model (ous)	**76.5**	**60.3**
	Global inference model (ours)	**79.8**	**62.3**

the proposed method are also verified. Firstly, the adaptive depth semantic segmentation can efficiently alleviating the unbalanced problem of training samples by adaptively deciding the feedback layer. Moreover, the forward step of the input image can also be adaptive based on the wanted semantic labels. Secondly, contour-aware neural network is proposed which can efficiently enhance semantic segment coherence and improve the localization of semantic regions. Thirdly, the global inference of the final semantic segmentation is beneficial to taking both coarse semantic labels and contour information into a unified framework. Besides, the contributions of coarse semantic labels and contour information are empirically decided. In conclusion, the adaptive semantic segmentation model achieves good performance on both pixel accuracy and class accuracy for its adaptive feedback and forward strategies. Moreover, the contour information can help a lot under the global inference framework.

5. Conclusion

In this work, a contour-aware network for semantic segmentation via adaptive depth is proposed. Firstly, a basic semantic segmentation model is introduced. Specifically, pyramid CNN is built for feature representation and fully connected CRF is constructed to model the relationships between nearby samples. The final decisions for pyramid results are made from sparse learning. Secondly, the semantic segmentation neural network via adaptive depth is described in detail and two advantages are claimed. One is alleviating the unbalanced problem in training and the other one is the model can adaptively determine the forward procedure in testing. Thirdly, a contour-aware neural network is proposed and it can efficiently alleviate the ambiguousness of semantic regions. Moreover, global inference of the final semantic label is proposed through global energy minimization.

Three contributions are claimed in this work. Firstly, a semantic segmentation neural network via adaptive depth is proposed. Secondly, contour-aware network is built for semantic segmentation. Thirdly, global inference for the final semantic segmentation is introduced. Experiments are conducted on three popular datasets and several state-of-the-art methods are served as competitors. The quantitative and qualitative results verified the superiority of the proposed method.

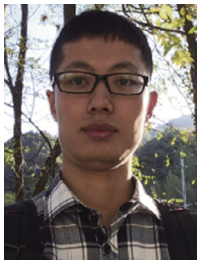
Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant 61773316 and 61379094, National Basic Research Program of China (Youth 973 Program) under Grant 2013CB336500, State Key Program of National Natural Science of China under Grant 60632018 and 61232010, and the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences.

References

[1] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, J.G. Rodríguez, A review on deep learning techniques applied to semantic segmentation, CoRR (2017). abs/1704.06857. arXiv: 1704.06857.
[2] S. Gould, X. He, Scene understanding by labeling pixels, Commun. ACM 57 (11) (2014) 68–77.
[3] J. Shotton, J. Winn, C. Rother, A. Criminisi, TextonBoost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context, Int. J. Comput. Vis. 81 (1) (2009) 2–23.
[4] N. Zeng, Z. Wang, H. Zhang, W. Liu, F.E. Alsaadi, Deep belief networks for quantitative analysis of a gold immunochromatographic strip, Cognit. Comput. 8 (4) (2016) 684–692.
[5] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A.M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 (2017) 643–649.
[6] Q. Wang, J. Gao, Y. Yuan, A joint convolutional neural networks and context transfer for street scenes labeling, IEEE Trans. Intell. Transp. Syst. PP (99) (2017) 1–14.
[7] Q. Wang, J. Gao, Y. Yuan, Embedding structured contour and location prior in siamesed fully convolutional networks for road detection, IEEE Trans. Intell. Transp. Syst. PP (99) (2017) 1–12.
[8] H. Cheng, X. Jiang, Y. Sun, J. Wang, Color image segmentation: advances and prospects, Pattern Recognit. 34 (12) (2001) 2259–2281.
[9] J. Kasson, W. Plouffe, An analysis of selected computer interchange color spaces, ACM Trans. Graph. 11 (4) (1992) 373–405.
[10] Z. Jiang, Q. Wang, Y. Yuan, Adaptive road detection towards multiscale-multilevel probabilistic analysis, in: Proceedings of the 2014 IEEE China Summit and International Conference on Signal and Information Processing, 2014, pp. 698–702.
[11] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[12] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[13] Y. Yuan, Z. Jiang, Q. Wang, Video-based road detection via online structural learning, Neurocomputing 168 (2015) 336–347.

[14] C. Chen, J. Luo, K. Parker, Image segmentation via adaptive k-mean clustering and knowledge-based morphological operations with biomedical applications, IEEE Trans. Image Process. 7 (12) (1998) 1673–1683.
[15] C. Zhang, L. Wang, R. Yang, Semantic segmentation of urban scenes using dense depth maps, in: Proceedings of the 2010 European Conference on Computer Vision, 2010, pp. 708–721.
[16] M. Atkins, B. Mackiewich, Fully automatic segmentation of the brain in MRI, IEEE Trans. Med. Imaging 17 (1) (1998) 98.
[17] K. Jiang, Q. Liao, S. Dai, A novel white blood cell segmentation scheme using scale-space filtering and watershed clustering, in: Proceedings of the 2003 International Conference on Machine Learning and Cybernetics, 2003, pp. 2820–2825.
[18] Q. Wang, S.Y. Li, Database of human segmented images and its application in boundary detection, IET Image Process. 6 (3) (2012) 222–229.
[19] Z. Liu, X. Li, P. Luo, L. Change, X. Tang, Deep learning Markovrandom field for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2016) 1.
[20] R. Vemulapalli, O. Tuzel, M. Liu, R. Chellappa, Gaussian conditional random field network for semantic segmentation, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3224–3233.
[21] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[22] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the 2016 IEEE International Conference on Computer Vision, 2016, pp. 1520–1528.
[23] T. Pohlen, A. Hermans, M. Mathias, B. Leibe, Full-resolution residual networks for semantic segmentation in street scenes, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[24] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for scene segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[25] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2016) 1.
[26] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the 2015 IEEE International Conference on Computer Vision, 2015, pp. 1529–1537.
[27] P. Pinheiro, T. Lin, R. Collobert, P. P. Dollr, Learning to refine object segments, in: Proceedings of the 2016 European Conference on Computer Vision, 2016, pp. 75–91.
[28] W. Liu, A. Rabinovich, A. Berg, ParseNet: looking wider to see better, in: Proceedings of the 2016 International Conference on Learning Representations, 2016.
[29] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[31] H. Li, Contour-aided accurate semantic segmentation using deep network, Bol. Téc. 55 (7) (2017) 105–111.
[32] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Proceedings of the 2004 Workshop on Statistical Learning in Computer Vision (ECCV), 44, 2004, pp. 1–22.
[33] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.
[34] J. Carreira, C. Sminchisescu, Constrained parametric min-cuts for automatic object segmentation, in: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3241–3248.
[35] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the 2015 International Conference on Learning Representations, 2015.
[37] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[38] K. Chatfield, K. Simonyan, A. Vedaldi, A. Zisserman, Return of the devil in the details: delving deep into convolutional nets, in: Proceedings of the 2014 British Machine Vision Conference, 2014.
[39] P. Krahenbuhl, V. Koltun, Efficient inference in fully connected CRFs with Gaussian edge potentials, Advances in Neural Information Processing Systems, 2011.
[40] J. Domke, Learning graphical model parameters with approximate marginal inference, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2454–2467.
[41] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. 58 (1) (1996) 267–288.
[42] G. Brostow, J. Fauqueur, R. Cipolla, Semantic object classes in video: a high-definition ground truth database, Pattern Recognit. Lett. 30 (2) (2009) 88–97.
[43] P. Sturgess, K. Alahari, L. Ladicky, P.H.S. Torr, Combining appearance and structure from motion features for road scene understanding, in: Proceedings of the 2009 British Machine Vision Conference, 2009.
[44] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: Proceedings of the 2009 IEEE International Conference on Computer Vision, 2009, pp. 1–8.
[45] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: the KITTI dataset, Int. J. Robot. Res. 32 (11) (2013) 1231–1237.
[46] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, A. Lopez, Vision-based offline-online perception paradigm for autonomous driving, in: Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 231–238.
[47] G. Brostow, J. Shotton, J. Fauqueur, R. Cipolla, Segmentation and recognition using structure from motion point clouds, in: Proceedings of the 2008 European Conference on Computer Vision, 2008, pp. 44–57.
[48] P. Kontschieder, S. Bulo, H. Bischof, M. Pelillo, Structured class-labels in random forests for semantic image labelling, in: Proceedings of the 2011 International Conference on Computer Vision, 2011, pp. 2190–2197.
[49] Y. Yang, Z. Li, L. Zhang, C. Murphy, J.V. Hoeve, H. Jiang, Local label descriptor for example based semantic image labeling, in: Proceedings of the 2012 European Conference on Computer Vision, 2012, pp. 361–375.
[50] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with superpixels, in: Proceedings of the 2010 European Conference on Computer Vision, 2010, pp. 352–365.
[51] D. Munoz, J. Bagnell, M. Hebert, Stacked hierarchical labeling, in: Proceedings of the 2010 European Conference on Computer Vision, 2010, pp. 57–70.
[52] R. Socher, C. Lin, A. Ng, C. Manning, Parsing natural scenes and natural language with recursive neural networks, in: Proceedings of the 2011 International Conference on Machine Learning, 2011, pp. 129–136.
[53] P. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene labeling., in: Proceedings of the 2014 International Conference on Machine Learning, 2014, pp. 82–90.
[54] C. Farabet, C. Couprie, L. Najman, Y. Lecun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
[55] G. Singh, J. Kosecka, Nonparametric scene parsing with adaptive feature relevance and semantic context, in: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3151–3157.
[56] Y. Yuan, Z. Jiang, Q. Wang, HDPA: hierarchical deep probability analysis for scene parsing, in: Proceedings of the 2017 IEEE International Conference on Multimedia and Expo, 2017, pp. 313–318.



Zhiyu Jiang is currently working toward the Ph.D. degree in the Center for Optical Imagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His current research interests include computer vision and scene understanding.

Yuan Yuan is currently a Full Professor with the Chinese Academy of Sciences, Beijing, China. She has authored or coauthored over 150 papers, including about 100 in reputable journals such as IEEE Transactions and Pattern Recognition, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Qi Wang received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, and Center for OPTical IMagery Analysis and Learning (OPTIMAL), and Unmanned System Research Institute (USRI), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.