IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 29, 2018, accepted November 27, 2018, date of publication December 3, 2018,
date of current version December 31, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2884502

Modeling With Prejudice: Small-Sample Learning via Adversary for Semantic Segmentation

ZHIYU JIANG, QI WANG, (Senior Member, IEEE), AND YUAN YUAN, (Senior Member, IEEE)
Center for Optical Imagery Analysis and Learning (OPTIMAL), School of the Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding author: Yuan Yuan (y.yuan1.ieee@gmail.com)

This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1002202, in part by the State Key Program of National Natural Science Foundation of China under Grant 61632018, in part by the National Natural Science Foundation of China under Grant 61773316, in part by the Natural Science Foundation of Shaanxi Province under Grant 2018KJXX-024, in part by the Fundamental Research Funds for the Central Universities under Grant 3102017AX010, and in part by the Open Research Fund of Key Laboratory of Spectral Imaging Technology, Chinese Academy of Sciences.

ABSTRACT Semantic segmentation has become one of the core tasks for scene understanding and many high-level works heavily rely on its performance. In the past decades, much progress has been achieved. However, some problems still need to be settled. One problem is about the challenging classification of various objects, which are with diverse viewpoints, illumination, appearance, and cluttered backgrounds, in a unified framework. The other one is focusing on the unbalanced distribution of semantic labels, where long-tail phenomenon exists and the trained model tends to be biased toward the majority classes when testing. And this problem can be regarded as the small-sample learning problem in semantic segmentation for the number of training samples upon the minority classes are small. For tackling these problems, a small-sample learning method via adversary is proposed and three contributions are claimed: 1) discriminatory modeling for semantic segmentation: two submodels are simultaneously built based on the attribute of semantic classs; 2) hierarchical contextual information consideration: both local and global contextual relationships are equally modeled under a hierarchical probabilistic graphical method and neighborhood relationship in label space are also considered; and 3) adversary learning for small-sample modeling: according to the structural relationships between small samples and the others, semantic classes are adversely modeled through computing the weighted costs. Experimental results on three benchmarks have verified the superiority of the proposed method compared with the state-of-the-arts.

INDEX TERMS Adversary learning, CRF, probabilistic graphical model, semantic segmentation, small-sample learning.

I. INTRODUCTION

People can rapidly understand images by recognizing objects and their spatial relationships through visual system. Similarly, semantic segmentation tends to assigning a label from a predefined label set to every region corresponding to a specific object. Semantic segmentation plays an important role in scene understanding and many other vision applications, which are naturally formulated as classification task with a proper label space, can benefit a lot from semantic segmentation [1]–[4], such as Advanced Driver Assistant System (ADAS), human-computer interaction, robot navigation and video surveillance [5]–[9].

Semantic segmentation has been historically studied and great progress has been achieved in recent decades. However, the task is still challenging for objects can vary a lot with diverse viewpoints, illumination, appearance, as well as various backgrounds. In addition, unbalanced distribution of the semantic labels from the training data, which is obtained from pixel-wise labeling, hinders the small-sample regions from being accurately classified. For solving these issues, lots of frameworks have been proposed. Most works focus on the utilization of contextual information because it can provide important cues between neighboring regions. For example, a bike is more likely to be neighbored with the road and it is not likely to be surrounded by the sky. Moreover, contextual information is more important when the object is ambiguously represented in feature space.

Probabilistic Graphical Model (PGM) is one of the most successfully models when considering contextual information and it can be explained as follows. Samples (pixels, superpixels, or segmented areas) are defined as the nodes, and all nodes have connections which are defined as edges.

Meanwhile, each edge is assigned a value which represents the similarity of the connected nodes. The objective is making the visually alike or spatially close regions share the same semantic label, whereas those disparate or remote regions tend to be labeled differently. Conditional Random Field (CRF), which is the most successful and widely utilized PGM model, has witnessed great success in semantic segmentation [10], [11]. Concretely, CRF model formulates the semantic segmentation as a conditional probability inference problem, which can simultaneously consider local appearance information and smoothness priors [10]. Although these models have worked well on semantic segmentation, they can not capture large input contexts which is very important for segmenting large regions, such as road [12].

Convolutional Neural Network (CNN) has revolutionized the area of computer vision recently due to its outstanding performance on object detection and recognition [13], [14]. The multi-layer representation can efficiently extract local significant cues and the local contextual information is aggregated together to represent higher contextual information. However, the original CNN can not directly adapt to semantic segmentation for the elimination of detailed information, such as local position and edge information. This phenomenon motivates experts to explore CNN and its varieties for scene parsing, such as Fully Convolutional Networks (FCN) [15] and Recurrent Neural Network (RNN) [16]. Although CNN models have strong ability to represent images, the smooth constraint of similar pixels sharing similar labels are underestimated [17].

For unbalanced samples, traditional methods tend to mining the true sample distribution or structural information through data dimension reduction, such as PCA, LDA. Through this strategy, the feature dimension is less than the amount of small samples and a more competitive model will be obtained. Another way to handle this problem is data augmentation. The original samples are transformed through adding noise, rotation, flipping, shifting, scaling, *etc.*, and by putting the transformed results back to the raw data, the unbalanced problem is eliminated. However, the original samples are all changed or ''polluted'' by the new samples through the mentioned methods and the entropy is also increased.

Starting from these limitations, a hierarchical model which focuses on contextual information is built. Specifically, the local and the global spatial contextual relationships are both considered under pyramid probabilistic graphical model. Meanwhile, the neighboring correlations in label space are also constrained as shown in Fig. 1. Moreover, an adversary strategy is proposed which fully utilizes the structural information between small samples and the others. Overall, the main contributions of this work are summarized and explained as follows:

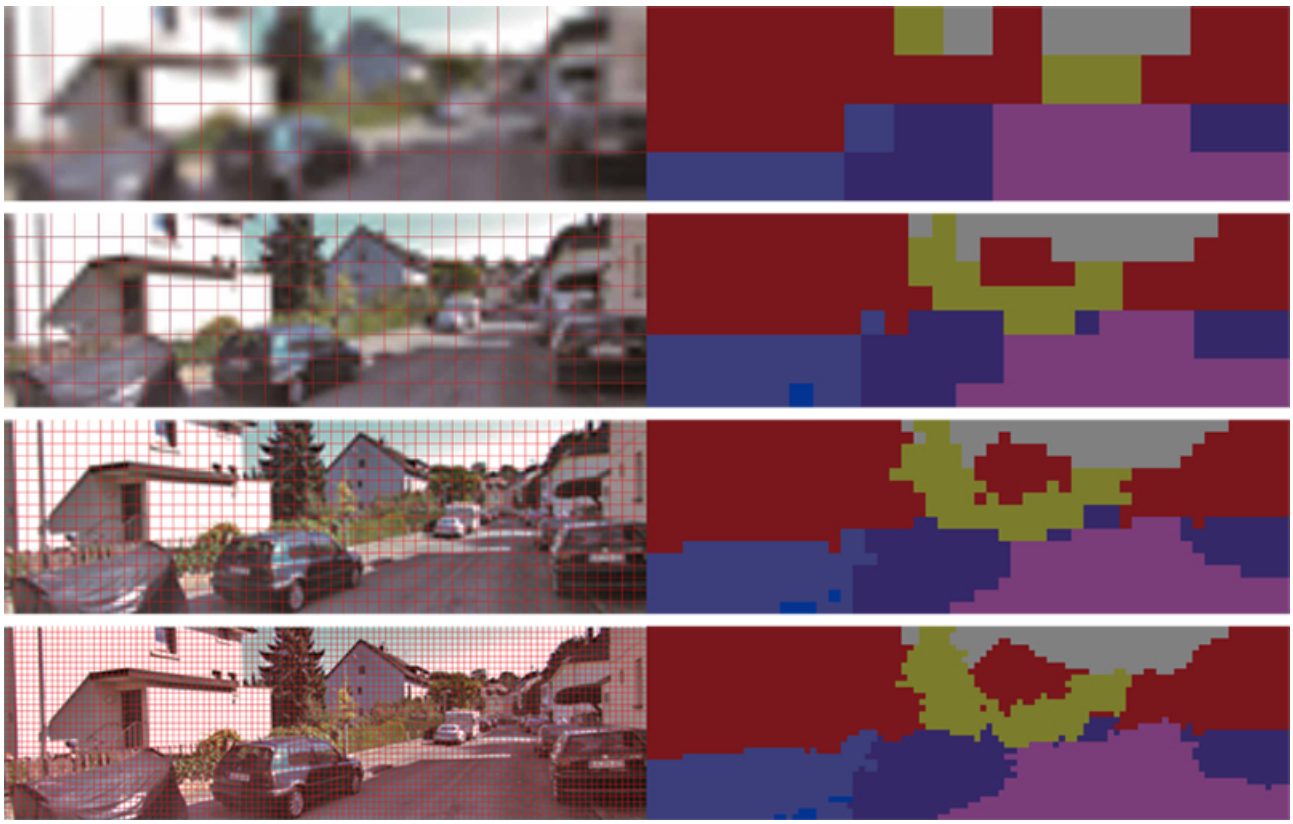


FIGURE 1. Typical pyramid strategy results. In this work, the image is directly segmented into patches and the inferred results in high resolution is still convincing. Furthermore, the correlations between multi-scale semantic label results can provide effective context relations.

1) *Discriminatory model for semantic segmentation.* Recent progress in semantic segmentation is more likely to segment objects in a unified framework without considering their attributes. Specifically, objects often vary a lot and necessary prior information can help a lot. For example, the size of semantic object is one of the most important priors for semantic segmentation and it is important to take region size into consideration. On the one hand, the region size can indicate the degree of unbalaced distribution of semantic samples, the scale information can also be inferred on the other hand. Consequently, a discriminatory model is introduced which takes region size as a judgement and two submodels are simultaneously built.

2) *Hierarchical contextual information consideration.* Taking wide range of contextual information into consideration is a challenging problem for semantic segmentation. Traditional methods tend to utilize probabilistic graphical models, such as CRFs, to solve this problem. Although significant progress has been made through CRF models, they can not efficiently capture the contexts of large regions. To solve this problem, a hierarchical contextual information consideration model is built and two advantages are concluded. Firstly, both local and global contextual relationships in spatial space are considered under pyramid probabilistic graphical model. Secondly, the neighboring correlations in label space are also served as statistical constraints in the decision procedure.

3) *Adversary strategy for small-sample learning.* For semantic segmentation, the phenomenon that the label space is distributed in an unbalanced manner is widely known and it can lead to bad capability of modeling semantic samples, especially for small-sample data. Many previous works focus on dimension reduction or increasing the number of small samples through data augmentation. Although this strategy can help decrease the influence of unbalanced data, it will increase the number of training samples and this can result in increased consumptions of memory and computation. Inspired from the idea that samples with different semantic labels should weighted contribute to the loss function, an adversary strategy for small-sample learning without transforming or increasing training samples is proposed in this work.

The remaining parts of this paper are organized as follows. The related works are revised in Section II. In Section III, the formulation of the proposed method is described in detail. Section IV demonstrates the experimental setup and results are also analyzed in this part. Finally, conclusions are drawn in Section V.

II. RELATED WORKS

Semantic segmentation has attracted much attention in recent decades. And in some literature, semantic segmentation is regarded as semantic annotation, image parsing, scene parsing, and Full Scene Labeling (FSL). Among various works, we review and analyze two slices of works which are most relevant to the proposed method.

A. CRF AND ITS VARIETIES

Graphical models have been widely utilized for semantic segmentation for their ability of taking contextual information into consideration. Markov Random Fields (MRF) [18]–[20] and CRF [10], [11], [21], [22] are most successful models. Multi-level strategy is always considered in CRF modeling. Lempitsky *et al.* [23] formulate a joint-CRF on multiple levels of an image segmentation hierarchy. He *et al.* [21] utilize CRF to capture different contextual relationships ranging from local to global under multi-scale consideration. A hierarchical CRF is introduced by Ladický *et al.* [22] to combine features extracted from pixels and segments. Furthermore, dense CRF is defined over pairwise pixels [24] and [25]–[27] introduce dense CRF for higher order labeling coherence. Higher order relations are also modeled by Kohli *et al.* [28] and Kontschieder *et al.* [29]. Inspired by the advances in CNN, Zheng *et al.* [17] firstly formulate the CRF as a part of neural network, and makes the inference procedure of CRF equal to training the same neural network until convergence.

B. CNN AND ITS VARIETIES

It is noticeable that deep learning models, particularly deep convolutional neural network [30], start to be popular in semantic segmentation. A number of early works focus on adapting the pre-trained models, such as ImageNet [30] learned from classification task, to detection or segmentation [27], [31]. Pre-processing steps, such as superpixel segmentation, object detection or classification tasks [31], [32], and post-processing steps, such as CRF and feature classifiers [31], [32], are all essential for these approaches. Recently, end-to-end and pixel-to-pixel methods, which are named as Fully Convolutional Network (FCN) [15] have achieved great progress in semantic segmentation. These approaches do not depend on any pre-trained CNN models or classifiers and directly learn a FCN model for semantic segmentation. As FCN methods do not have better performance on small objects, Mohan [12] introduces deep deconvolution networks to capture high-order image structure beyond edge primitives. Chen *et al.* [27] also simultaneously take a fully connected CRF and deep convolutional network into consideration to improve the localization accuracy.

Furthermore, recent literatures mainly focus on the following two aspects. The first aspect is focusing on the structure of the CNNs which tend to capture large-wide contextual information, such as convolutional layer and mapping strategy. For example, Wang *et al.* [33] implement a hybrid dilated convolution layer which can effectively enlarge the receptive fields to aggregate global information. The second aspect is about small-sample problem. Zhao *et al.* [34] simultaneously utilize local and global clues to make the final prediction more reliable under pyramid model.

Although significant progress has been made for semantic segmentation, some challenges still exist. Firstly, adequately consideration of contextual information can improve the performance a lot and how to take this information into

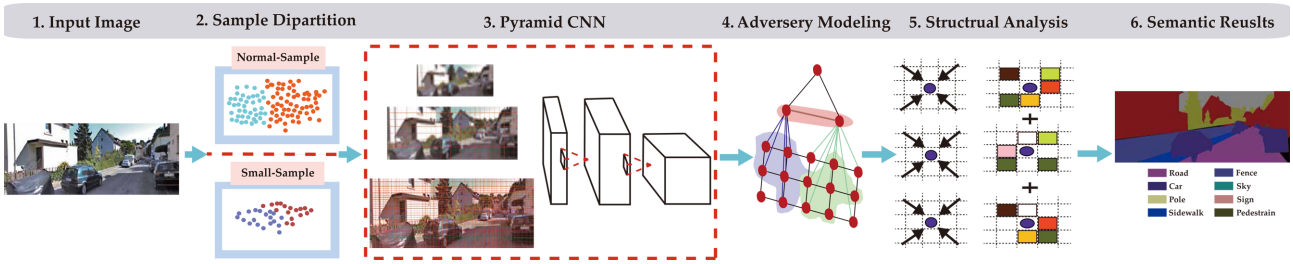


FIGURE 2. Semantic segmentation pipeline. The input image is firstly divided into normal samples and small samples. Subsequently, Gaussian pyramid is built and each image in a certain scale is segmented into patches. Pyramid CNN is constructed for patch representation and the adversary modeling is proposed in this part. Finally, the semantic labels are obtained through sparse structural analysis.

consideration is still challenging. Secondly, the trained model will be biased towards the majority classes without considering the small-sample problem. For tackling these difficulties, both local and global contextual relationship are equally modeled under a hierarchically probabilistic graphical method in this paper. Furthermore, the constraints of neighborhood relationship in label space are also considered. On the other hand, according to the structural relationships between small samples and the others, a small-sample learning method via adversary is proposed.

III. SMALL-SAMPLE LEARNING VIA ADVERSARY FOR SEMANTIC SEGMENTATION

In this section, semantic segmentation is elaborately formulated. Concretely, a basic semantic segmentation model is first formulated through discrete energy minimization. Subsequently, for adequate contextual information consideration, a hierarchical model is built upon CNN structure under image pyramid. Furthermore, an adversary strategy is proposed to handle small-sample learning problem. The pipeline of the proposed method is illustrated in Fig. 2.

A. BASIC FORMULATION

We adopt the common practice in semantic segmentation and formulate the semantic segmentation of an image as discrete energy minimization. Specifically, given an image $\{x_1, x_2, \ldots, x_n\} \in \mathcal{X}$ with pixel-wise labels $\{y_1, y_2, \ldots, y_n\} \in \mathcal{Y}^n$. Here n is the number of pixels and $\mathcal{Y}$ is the set of all possible labels, which are defined as semantic labels. The quality of semantic labeling is measured by s loss function $L(y', y)$, where y' is the predicted labels and y is the ground truth of x. The goal of semantic segmentation is to find a mapping $F : \mathcal{X} \to \mathcal{Y}$ such that the loss function L can be minimized. Instead of modeling F as a single parametric model, such as CRF, the proposed method tends to model the problem under pyramid.

Although it is widely believed that CNN is more capable of feature learning, it is significant to model the correlations between patches and labels. For semantic segmentation, each pixel is expected to correspond to only one semantic label, and a semantic label can be distributed to multiple pixels. Based on these assumptions, a fully connected CRF model $G = \{V, E\}$, which is an undirected graph, is built for semantic segmentation. Here, V represents all the nodes and E refers to the set of edges. Specifically, every pixel x_i and its unpredicted label $y_i \in \mathcal{L}$ are defined as nodes V and two nodes (x_i, x_j) are connected which is weighted for the similarity of the two nodes in feature space. The Gibbs energy of fully connected pairwise CRF model [24] can be written 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 $i, j \in \{1, \ldots, N\}$ and N is the amount of training pixels. The unary energy component $\psi_u(y_i, \boldsymbol{f})$ measures the cost when assigning sematic label y_i to node x_i given image features $\boldsymbol{f}$.

The pairwise energy components $\psi_p(y_i, y_j, \boldsymbol{f})$ measures the cost of assigning labels (y_i, y_j) to the connected nodes (x_i, x_j) simultaneously. In this work, the unary energy is obtained from CNN model, which focuses on inferring semantic labels and ignoring the smoothness constraint in label space. To alleviate this problem, the pairwise energy provides a data-dependent smoothing term that encourages neighbor pixels to be assigned the same semantic label. As was done in [35], the pairwise function is defined as

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

where M is the number of Gaussian kernels. 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 defined as

$$\mu(y_i, y_j) = \begin{cases} 1, & \text{if } y_i \neq y_j, \\ 0, & \text{if } y_i = y_j, \end{cases} \tag{3}$$

The objective is to find an optimal label assignment for all the samples. And this is equivalent to minimizing the CRF energy defined in (1):

$$y^* = \arg\min_y E(\boldsymbol{Y}, \boldsymbol{f}), \tag{4}$$

and for solving this equation, truncated EM in [35] is adopted for its good performance.

After solving (4), the parameters of the energy function in (1) are estimated. For a test image, the conditional probabilities of each pixel can be inferred and the semantic label is determined as the label corresponding to the maximum conditional probability.

B. HIERARCHICAL CONTEXTUAL INFORMATION CONSIDERATION

For CRF method mentioned above, the smooth constraints between neighboring semantic labels can be ensured in some degree. However, two issues still need to be considered. The first issue is about modeling context relations which refer to the semantic correlation between one object and its neighboring objects. The contextual information can provide crucial clues and adequate consideration is essential. To model the context relations, a hierarchical model is built under image pyramid. The second one is about the image representation. For machine learning problem, feature representation is absolutely important and CNN feature is considered in this work for its strong representation ability. Furthermore, the image pyramid model can decrease the affect of CNN feature's scale sensitivity.

1) PYRAMID IMAGE BUILDING

In this work, Gaussian pyramid is considered for its linear property and given an image X, the Gaussian pyramid can be represented as X^s, where $s \in \{1, \ldots, S\}$. In this work, only down sampling is considered and larger s would result in smaller image size. The reason can be explained as follows. On one hand, up sampling strategy, such as linear interpolation and nearest neighbor interpolation, can import redundant information. On the other hand, the size of receptive field can be adjusted through the neighborhood size. Based on these considerations, each pyramid image under a certain scale is segmented into fixed-size pathes. Besides, for the pyramid image with larger Gaussian kernel size, i.e., larger s, smaller number of pathes are segmented. This setup can be explained as follows. It is a tradeoff between the capability of providing detailed information and owing wide receptive field for a certain CNN model. Consequently, detailed texture information in a small neighborhood and global relations in a widely receptive field are synchronously obtained through Gaussian pyramid.

2) MULTI-SCALE CNN MODELING

Given an image, amounts of redundant information exists for the pixels' spatial similarity. Image segmentation is an efficient strategy to alleviate the effects of this situation. In this work, the pyramid images are directly segmented into patches for simplification. Although superpixel segmentation may be more reasonable, it is time-consuming over pyramid images and the amount of patches is set as a big number for high-resolution image which will result in each patch nearly belonging to one semantic label.

For each patch, it will be represent by CNN feature. The Deep Convolutional Neural Network is widely applied for its abundant information from low-level to high-level as the depth of CNN increases. Deeper CNN can achieve higher semantic information while less detailed information which is essential for small objects. Although Long *et al.* [15] proposed layer fusion strategy to solve this problem, the proposed method in this work solves this problem in a different angle. For a convolutional network with parameters θ, including all the hyperparameters and weights of each layer, the multi-scale networks θ^s can be obtained through instantiation.

The pre-trained VGG model [36] is utilized to initialize the CNN model and the last FC-1000 layer is replaced by FC-k layer, where k is the class number. After initialization, each network of the multi-scale networks is updated through fine-tuning separately. For the training data, the label of each segmented patch is defined as the pixel's label corresponding to the highest frequency. In the fine-tuning step, the CNN model is re-trained with small learning rate ultimately.

3) STRUCTURAL DECISION

After multi-scale CNN modeling, the conditional distributions $P(\boldsymbol{Y}^s|\boldsymbol{X}^s), \forall s \in \{1, \ldots, S\})$ are obtained and the inferred semantic labels of a certain scale s can be written as

$$\boldsymbol{Y}^s = \arg \max_{\boldsymbol{Y}^s \in \{1,\ldots,l\}^N} P(\boldsymbol{Y}^s|\boldsymbol{X}^s), \quad s \in \{1, \ldots, S\} \tag{5}$$

where N is the amount of patches in current scale s.

As illustrated in Fig. 1, how to take full advantage of the inferred results across scale is meaningful. Traditional methods tend to utilize voting strategy although significant contextual information is eliminated. For example, the CRF model trained by the patches in small scale tends to label the testing patch as large things or stuff, such as road and sky. Moreover, the inferred labels between different scales also show strong correlations which can be regarded as the contextual information between scales. For example, given a pixel which is labeled as road in small scale model and labeled as car in large scale model, and then we can strongly believe that the pixel belongs to car region. On the contrary, if a pixel is labeled as road region in small scale model and labeled as pedestrian class in large scale model, and then we will be confused about labeling the pixel for the pedestrian is impossible to be located in the sky region. One solution to this problem is to calculate the joint conditional probability$P(\boldsymbol{Y}|\boldsymbol{X}) = P(\boldsymbol{Y}|\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_S)$. However, this equation is difficult to calculate for the absence of essential information. In order to solve this problem, a sparse based model is built to calculate the final scene parsing results considering the prediction results of neighboring pixels. For each scale, the predicted labels $\boldsymbol{Y}^s$ are firstly resized to the original image size with the nearest interpolation method. Subsequently, for a certain pixel x_i, its label and neighboring labels inferred from the total S scales can be written as $\boldsymbol{a}_i \in \mathcal{L}^{k\times 1}$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N]^T$, where k is the total number of its neighbours. The mathematical equation can be

written as

$$\boldsymbol{w}^* = \arg\min \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{Y}\|_2^2, \quad s.t. \|\boldsymbol{w}\|_1 \leq \varepsilon \tag{6}$$

where ε indicates the residual error and this problem can be solved by Lasso [37]. The final labels can be obtained by

$$\boldsymbol{y}_i^* = \arg\min_{\boldsymbol{y}_i \in \mathcal{L}} \|\boldsymbol{a}_i \boldsymbol{w}^* - \boldsymbol{y}_i\|_2, \quad i \in [1, \ldots, N]. \tag{7}$$

The flowchart of this model is illustrate in Fig. 3(a).

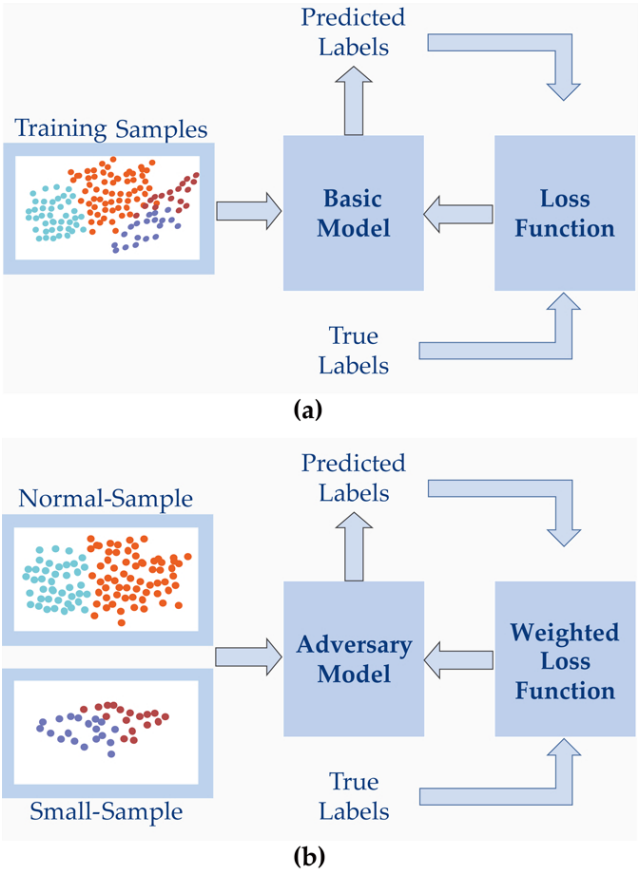


FIGURE 3. The basic model for semantic segmentation is illustrated in (a) and (b) demonstrates the adversary model proposed in this work.

C. ADVERSARY STRATEGY FOR SMALL-SAMPLE LEARNING

For semantic segmentation, unbalanced distribution of the semantic labels from the training data is a ubiquitous phenomenon. However, learning from unbalanced data is still a difficult task for its ambiguous structural relationships. It can be observed that models trained with unbalanced data tend to be biased towards the majority classes when testing. Many works have achieved good performance towards solving this problem, including sampling methods, data augmentation methods, and hybrid approaches. The proposed method handles this problem in another aspect. Before demonstrating the proposed method, small-sample should be defined. It can be explained in two aspects. Firstly, small-sample means the region size of the sample is small, such as the far-away cars and pedestrians. Secondly, small-sample can be regard as a kind of sample with smaller size compared with other sample with another different semantic label. For semantic segmentation, each pixel or segment is served as one sample and this fact makes the first aspect of explanation equals to the second aspect. As a consequence, small-sample learning is equal to unbalanced data learning in semantic segmentation. Inspired by Wang *et al.* [38], an adversary strategy is proposed to handle small-sample learning problem.

The goal of the proposed adversary model is to classify small-sample with higher accuracy. According to (2) and (3), it can be obvious that the weight of all samples are equal and each sample uniformly contributes to the final energy function within the image. This uniformly weighted strategy will result in the bias of the trained model which tends to infer a sample belonging to majority classes. In order to alleviate this issue, the contribution of each sample is weighted based on the distribution of the semantic labels. (3) can be rewritten as

$$\mu(y_i, y_j) = \begin{cases} adv_{ij}, & if\ y_i \neq y_j, \\ 0, & if\ y_i = y_j, \end{cases} \tag{8}$$

where adv_{ij} is the weighted parameters and it should be defined in the following principle.

For the training samples, it can be divided into two categories, normal sample and small sample. In this work, this procedure is processed by k-means for its convenience. Specifically, the number of each semantic label is served as feature and the number of cluster centers is set to 3 and the samples belonging to the minimum center are defined as small samples. The rest samples are defined as normal samples. Inferred from the loss function 1, the function is a tradeoff between the normal samples and small samples. For the purpose of high accuracy for small samples, small samples should contribute more to the overall loss function, while the normal samples contribute less to the overall loss function. The loss function defined on these two kinds of samples can be regarded as adversarial learning. If the small samples' loss is weighted high, the accuracy of small-samples will increase while the accuracy of normal sample decreases. This adversary phenomenon can help to build a small-sample learning model which can obtain high accuracy on samll samples. Concretely, the small sample penalty factor adv_{ij} should be proportional to the corresponding sample number. In this work, adv_{ij} is defined as

$$\begin{aligned} &adv_{ij} = 1 - G\left[\min(n_i, n_j)\right], \\ &\underset{i<j}{} \\ &s.t. \quad i, j \in \{1, \ldots, l\}, \end{aligned} \tag{9}$$

where $G(\cdot)$ is a Gaussian function estimating the distribution upon the amount of semantic labels. n_i is the number of the ith semantic label in the training set. l is defined as the number of semantic labels.

Based on (9), the CRF model can be re-trained and the obtained CRF model mainly focuses on small samples. Fig. 3(b) shows the adversary model proposed in this work.

IV. EXPERIMENTS

In order to evaluate the proposed method, the datasets are first introduced. Subsequently, the evaluation criteria and the parameter setting are explained in detail. Finally, thorough analyses on the experimental results are conducted.

A. DATASETS

Semantic segmentation has attracted many experts exploring efficient frameworks and a number of semantic segmentation datasets are developed. Three datasets [18], [39], [40] are utilized to verify the performance of the proposed method for their broad application and challenging properties.

- **CamVid dataset** is the first collection of videos with corresponding semantically labeled images at 1Hz. It mainly focuses on road scene understanding, including 468 training images and 233 testing images of day and dusk scenes [39]. The target of CamVid dataset is to segment 11 semantic classes such as road, buildings, cars, pedestrians, signs, side-walk, etc.
- **Stanford-Background dataset** [18] contains 715 images of outdoor scene 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** [40] is a large publicly available road scene dataset and some images were extracted and manually annotated for scene parsing. For convenience of the comparison, the labeled images by [41] are utilized as experimental data which contain 142 images. Moreover, 11 semantic classes, such as buildings and road, are severely imbalanced distributed.

B. ACCURACY

Many evaluation criteria have been proposed for semantic segmentation, and these metrics are usually variations on pixel accuracy and class accuracy. Hence pixel accuracy and class accuracy are considered in this work. For explanation, we remark the following notations: the total semantic class number is k (from class 1 to class k) and n_{ij} is the amounts of pixels which are predicted as class j whereas the true label is class 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

$$Pixel\ Accuracy = \frac{\sum_{i=1}^{k} n_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}}. \tag{10}$$

- **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 as follows

$$Class\ Accuracy = \frac{1}{k}\sum_{i=1}^{k}\frac{n_{ii}}{\sum_{j=1}^{k} n_{ij}}. \tag{11}$$

C. PARAMETER SETTINGS

For the patch defined upon pyramid images, the smallest scale image is segmented into $4 \times (4 \times w)$ patches, where w is empirically set as the ratio of image height to width. For the next level, the image is segmented into $8 \times (8 \times w)$ patches, and so on. Naturally the patch size varies over image size. The level of the pyramid is set to 5 for the trade-off between computation cost and accuracy. Experiments have shown that deeper depth could lead to better performance with less accuracy growth rate.

D. PERFORMANCE ANALYSIS

Before analyzing the experimental results, two models should be clearly described. As illustrated in Fig. 3, the basic model is mainly described in Sec. III-B, including Gaussian pyramid, CNN feature representation, CRF modeling, and strucutral decision based on sparse learning. The adversary model is improved based on the basic model and adversary strategy is considered in this model. In order to verify the performance of the proposed method, especially the adversary strategy, experiments on the basic model and the adversary model are simultaneously conducted. Experimental results are evaluated by qualitative and quantitative measures. Quantitative results are shown in Table 1 and typical scene parsing results of the three datasets are presented in Fig. 4.

For a more objective comparison, we calculate the two accuracies introduced in Section IV-B. The results are shown in Table 1. We can see clearly that the highest scores are mostly achieved by the proposed method. Only Hierarchical Features proposed by Farabet *et al.* [31] is a little better than ours on the class accuracy, but the differences are very small. Besides, the average accuracy of [31] is not comparable with ours. Therefore, it is reasonable to claim that the proposed method is more effective than the other competitors.

In the following, a more detailed analysis on the three datasets will be presented.

- **Basic Model Performance**
  Significant results have been achieved when taking the pixel accuracy as the evaluation criterion according to the Table 1. For example, Recursive Neural Network model [46] and Recurrent Neural Network model [16] can efficiently take the contextual constraints into account on the structure of the models. However, the basic model explores the power of contexts from two directions. The first one is focusing on the local and global contexts. And the probabilistic graphical model is built simultaneously. The other one is the strengths

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

Dataset	Approach	Pixel Accuracy	Class Accuracy
CamVid	SFM+Appearance [42]	69.1	53.0
	Boosting [43]	76.4	59.8
	Structured Random Forests [29]	72.5	51.4
	Local Label Descriptors [44]	73.6	36.3
	Boosting+pairwise CRF [43]	79.8	59.9
	Local Labeling+MRF [32]	77.6	43.8
	Basic Model (ous)	81.1	49.9
	Adversary Model (ours)	**81.6**	**62.4**
Stanford	Stacked Labeling [45]	76.9	66.2
	Recursive Neural Networks [46]	78.1	N/A
	Recurrent Neural Networks [16]	80.2	69.9
	Hierarchical Features [31]	81.4	**76.0**
	WAKNN+MRF [47]	74.1	62.2
	Basic Model (ous)	81.7	70.6
	Adversary Model (ours)	**81.9**	75.8
KITTI	Temporal Semantic Segmentation [42]	51.2	61.6
	Semantic Segmentation Retrieva [42]	47.1	58.0
	Basic Model (ous)	**79.8**	45.84
	Adversary Model (ours)	79.6	**64.7**

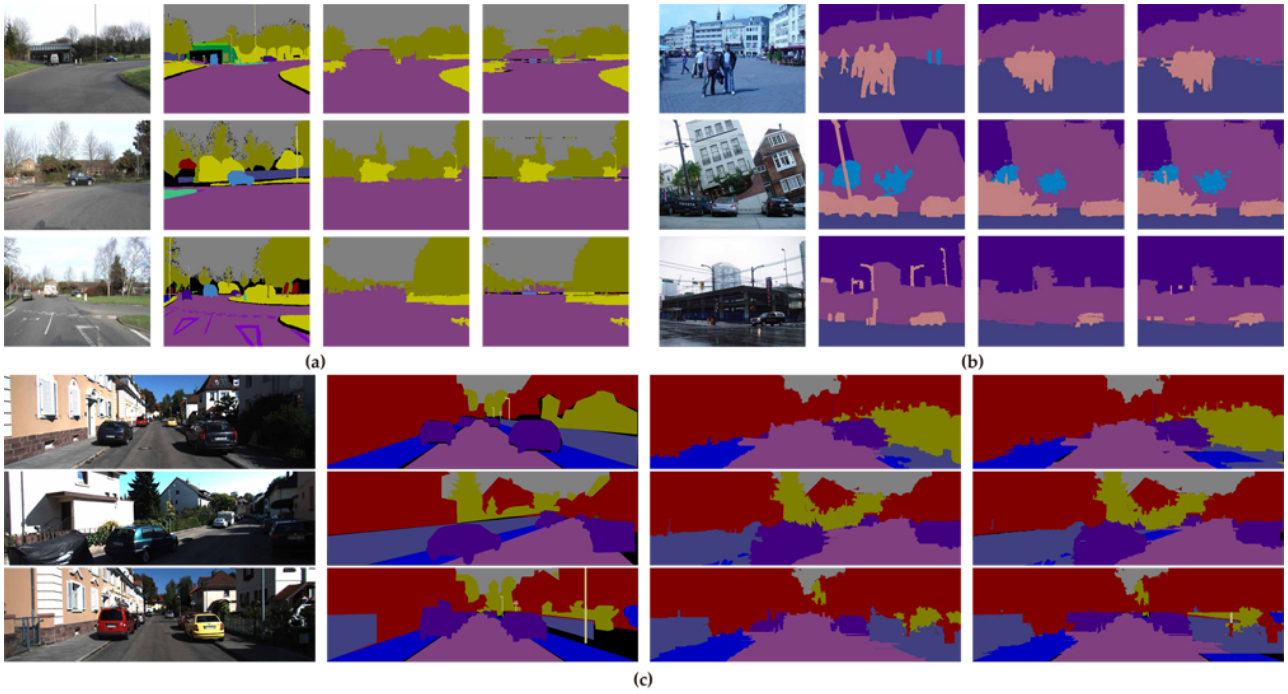


FIGURE 4. Qualitative semantic segmentation results. CamVid results are shown in (a) and (b) is Stanford-Background results. KITTI results are demonstrated in (c). For each dataset results, the first column indicates the input images, the second column provides the groundtruth. The third column shows the semantic segmentation results based on the proposed basic model and the last column shows the adversary results.

of pyramid model by taking hierarchical inferred labels into solving a sparse problem. Generally, a hierarchical CRF models based on Gaussian pyramid can take different levels of object parsing into consideration which leads to adequate contextual for semantic segmentation. However, the proposed basic model has shown its weakness on the aspect of class accuracy. The reasons can be explained as follows. The basic model is based on sampling certain number of patches from Gaussian pyramid and this strategy would ignore the small-sized semantic class. For example, for the KITTI dataset, the number of pixels defines as pole label [41] is very small and nearly zero number of pixels were correctly though the basic model. In order to alleviate this problem, the adversary model is proposed and the detailed analyses are described as follows.

- **Adversary Model Performance**
  CamVid dataset. The images in this dataset are sampled from two daytime and one dusk sequences and the first block of Table 1 shows the performance of the proposed method compared with state-of-the-arts. We can observe the positive impact of the proposed basic model and the adversary model in this work. For example, the appearance model [42] and the local labeling

method [32] 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, 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. Our method takes advantage of the CRF model and takes different levels of the scene into consideration which leads to higher pixel accuracy. On the other hand, the adversay model is better than the basic model considering both average accuracy and class accuracy. This phenomenon shows that the proposed adversary strategy can efficiently improve the accuracy of the small samples with small losses on the normal semantic class.

Stanford-Background dataset. Experiments on this dataset are conducted over 5-fold validation. Concretely, 572 images are served as training examples and the other 143 images are utilized to test the performance of the proposed basic method and the adversary model each time. The second block of Table 1 shows the superiority of our method. For example, Recursive Neural Network model [46] and Recurrent Neural Network model [16] can efficiently take the contextual constraints into account on the structure of the models. On this foundation, our work explores the power of contexts from two directions. The first one is focusing on the local contexts and the probabilistic graphical model is built. The other one exploits the strengths of pyramid model by taking hierarchical inferred labels into solving a sparse problem. Experiments on pixel accuracy verified the contributions of this work. Considering the class accuracy, the proposed adversary method is a little worse than the hierarchical features proposed by Farabet *et al.* [31]. However, compared with the basic model, the class accuracy is increased.

KITTI dataset. This dataset is captured by RGB camera with wide view angle and 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.* [41] and high class accuracy verified the effectiveness of the temporal information. On the contrary, temporal context information does not take into account in our method temporarily and competitive results on the pixel criterion also show the superiority of the proposed method. Compared with the basic model, the adversary model is better on the class accuracy whereas the average accuracy is a little worse. This results verified the ability of small-sample learning for the adversary strategy.

Significant results have been reached when take the pixel accuracy and class accuracy as the evaluation criterion. Meanwhile, the contributions of the proposed methods in this work are also verified. On one hand, the basic model proposed in this work takes hierarchical contextual information into consideration. Specifically, hierarchical CRF models based on Gaussian pyramid can take different level of object parsing into considerations and the structural decision based on sparse learning simultaneously lead to adequate context relations for semantic segmentation. On the other hand, discriminatory model is built upon small samples and normal samples. Meanwhile, the adversary strategy which focuses on small samples is strengthened through weighted loss definition. These characteristics can efficiently increase the power of small-sample learning the class accuracy is also increased. Generally, the basic model proposed in this work achieves good performance on average accuracy for its adequate contextual information consideration. Moreover, the adversary model has good performance on the class accuracy for the utilization of the adversary strategy.

V. CONCLUSION

In this work, a small-sample learning model via adversary for semantic segmentation is proposed. Firstly, a discriminatory model for semantic segmentation is introduced and two submodels based on pyramid convolutional neural network are simultaneously built. In this model, images are segmented into patches under Gaussian pyramid and each patch is represented by a CNN model. Secondly, hierarchical contextual information is taken into consideration to alleviate the wide-range perception problem through discrete energy minimization. CRF method is considered in this work and structural decision is made through sparse representation. Thirdly, an adversary strategy is proposed for small-sample learning by weighting the energy function of CRF. Experiments are conducted on three datasets and several state-of-the-arts are served as competitors. The quantitative and qualitative results verifies the superiority of the proposed work.

REFERENCES

[1] L. Zhang *et al.*, "Improving semantic image segmentation with a probabilistic superpixel-based dense conditional random field," *IEEE Access*, vol. 6, pp. 15297–15310, 2018.

[2] W. Yu, Z. Hou, P. Wang, X. Qin, L. Wang, and H. Li, "Weakly supervised foreground segmentation based on superpixel grouping," *IEEE Access*, vol. 6, pp. 12269–12279, 2018.

[3] L. Fan, H. Kong, W.-C. Wang, and J. Yan, "Semantic segmentation with global encoding and dilated decoder in street scenes," *IEEE Access*, vol. 6, pp. 50333–50343, 2018.

[4] S. Gould and X. He, "Scene understanding by labeling pixels," *Commun. ACM*, vol. 57, no. 11, pp. 68–77, 2014.

[5] M. Riveiro, M. Lebram, and M. Elmer, "Anomaly detection for road traffic: A visual analytics framework," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 2260–2270, Aug. 2017.

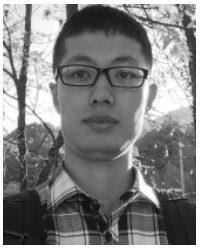
[6] Y. Yuan, Z. Xiong, and Q. Wang, "An incremental framework for video-based traffic sign detection, tracking, and recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 7, pp. 1918–1929, Jul. 2016.

[7] Z. Jiang, Q. Wang, and Y. Yuan, "Adaptive road detection towards multiscale-multilevel probabilistic analysis," in *Proc. IEEE China Summit Int. Conf. Signal Inf. Process.*, Jul. 2014, pp. 698–702.

[8] Y. Yuan, J. Fang, and Q. Wang, "Online anomaly detection in crowd scenes via structure analysis," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 548–561, Mar. 2015.

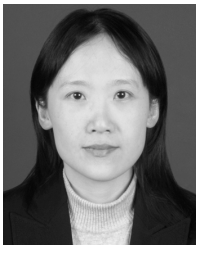
[9] J. Fang, Q. Wang, and Y. Yuan, "Part-based online tracking with geometry constraint and attention selection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 854–864, May 2014.

[10] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "TextonBoost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, Dec. 2007.
[11] M.-M. Cheng *et al.*, "ImageSpirit: Verbal guided image parsing," *ACM Trans. Graph*, vol. 34, no. 1, pp. 1–11, 2014.
[12] R. Mohan. (Nov. 2014). "Deep deconvolutional networks for scene parsing." [Online]. Available: https://arxiv.org/abs/1411.4101
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 346–361.
[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[16] P. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 82–90.
[17] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1529–1537.
[18] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2010, pp. 1–8.
[19] M. P. Kumar and D. Koller, "Efficiently selecting regions for scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3217–3224.
[20] J. Tighe and S. Lazebnik, "Superparsing," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 329–349, Jan. 2013.
[21] X. He, M. A. Carreira-Perpinan, and R. S. Zemel, "Multiscale conditional random fields for image labeling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun./Jul. 2004, pp. 695–703.
[22] L. Ladický, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical CRFs for object class image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2010, pp. 739–746.
[23] V. Lempitsky, A. Vedaldi, and A. Zisserman, "A pylon model for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1485–1493.
[24] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.
[25] A. Roy and S. Todorovic, "Scene labeling using beam search under mutex constraints," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, Jun. 2014, pp. 1178–1185.
[26] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 582–589.
[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[28] P. Kohli, L. Ladický, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 302–324, May 2009.
[29] P. Kontschieder, S. R. Bulò, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2190–2197.
[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[31] C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.
[32] J. Tighe and S. Lazebnik, "SuperParsing: Scalable nonparametric image parsing with superpixels," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 352–365.
[33] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.
[34] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[35] J. Domke, "Learning graphical model parameters with approximate marginal inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 10, pp. 2454–2467, Oct. 2013.
[36] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014, p. 1.
[37] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.
[38] X. Wang, A. Shrivastava, and A. Gupta, "A-fast-RCNN: Hard positive generation via adversary for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3039–3048.
[39] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 88–97, 2009.
[40] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[41] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, and A. Lopez, "Vision-based offline-online perception paradigm for autonomous driving," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 231–238.
[42] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 44–57.
[43] P. Sturgess, K. Alahari, L. Ladicky, and P. Torr, "Combining appearance and structure from motion features for road scene understanding," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–12.
[44] Y. Yang, Z. Li, L. Zhang, C. Murphy, J. V. Hoeve, and H. Jiang, "Local label descriptor for example based semantic image labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 361–375.
[45] D. Munoz, J. A. Bagnell, and M. Hebert, "Stacked hierarchical labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 57–70.
[46] R. Socher, C. C.-Y. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 129–136.
[47] G. Singh and J. Kosecka, "Nonparametric scene parsing with adaptive feature relevance and semantic context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3151–3157.



ZHIYU JIANG received the B.E. degree in intelligent science and technology from Xidian University, Xi'an, China, in 2013, and the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2018. He is currently a Post-Doctoral Researcher with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an. His current research interests include computer vision and scene understanding.

QI WANG (M'15–SM'15) 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 the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.



YUAN YUAN (M'05–SM'09) is currently a Full Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as the conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.