A Joint Convolutional Neural Networks and Context Transfer for Street Scenes Labeling

Qi Wang, *Senior Member, IEEE*, Junyu Gao, and Yuan Yuan, *Senior Member, IEEE*

***Abstract*— Street scene understanding is an essential task for autonomous driving. One important step toward this direction is scene labeling, which annotates each pixel in the images with a correct class label. Although many approaches have been developed, there are still some weak points. First, many methods are based on the hand-crafted features whose image representation ability is limited. Second, they cannot label foreground objects accurately due to the data set bias. Third, in the refinement stage, the traditional Markov random filed inference is prone to over smoothness. For improving the above problems, this paper proposes a joint method of priori convolutional neural networks at superpixel level (called as "priori s-CNNs") and soft restricted context transfer. Our contributions are threefold: 1) *a priori* s-CNNs model that learns priori location information at superpixel level is proposed to describe various objects discriminatingly; 2) a hierarchical data augmentation method is presented to alleviate data set bias in the priori s-CNNs training stage, which improves foreground objects labeling significantly; and 3) a soft restricted MRF energy function is defined to improve the priori s-CNNs model's labeling performance and reduce the over smoothness at the same time. The proposed approach is verified on CamVid data set (11 classes) and SIFT Flow Street data set (16 classes) and achieves a competitive performance.**

***Index Terms*— Scene labeling, convolutional neural networks, deep learning, label transfer, street scenes, data augmentation.**

I. Introduction

IN RECENT years, intelligent driving has been a hot topic for the research communities and industrial companies. It can promote the understanding towards fundamental computer vision and machine learning problems and enhance the actual experience of intelligent transportation. For this purpose, a critical challenge is how to understand the street scenes and react to the outside conditions efficiently. At present, researchers tackle this problem by an integration of several mature technologies, such as pedestrian detection [1], anomaly detection [2], vehicle detection [3], road surface detection [4], lane detection [5] and so on. However, these technologies are on the initial stage of scene understanding and far away from real requirement.

Fig. 1. Street scenes labeling examples. The images in the first row are street scenes and the second row illustrates the per-pixel labeling results.

In order to get a better knowledge of the street scene, a new computer vision task is proposed, semantic scene labeling. It combines segmentation, object detection and multi-object labeling into one single framework and can be regarded as a per-pixel labeling task. This is because for intelligent driving in street scenes, it is necessary to not only recognize the individual participant and event, but also have a thorough perception of the whole view. For instance, if the driver knows where the side buildings are, or what the traffic status is, he will drive more safely. Examples of street scene labeling are presented in Fig. 1.

However, because scene labeling is a unified framework and involves many fundamental computer vision tasks, it is still challenging since Wright [6] firstly put forward this concept in 1989. There are two questions to be solved in this topic: how to get distinctive internal representations of object appearance and how to improve labeling accuracies of foreground objects in the street scenes. Firstly, scene labeling is not like traditional single-object problem that needs to extract features between positive and negative samples. As a multi-object task, how to extract rich and discriminative features to describe different objects is essential to labeling, which is obviously more difficult than single-object task. For this purpose, many approaches (e.g. [7]–[12]) aim to exploit multiple features to characterize objects. The first five exemplar ones compute RGB based features to describe image by combination and fusion of them. The last two exploit 3D features (dense depth maps or 3D point clouds) to reconstruct 3D street scenes. Generally, the more features are extracted, the more

Manuscript received November 24, 2016; revised March 24, 2017 and June 17, 2017; accepted July 2, 2017. Date of publication August 17, 2017; date of current version May 2, 2018. This work was supported in part by the National Natural Science Foundation of China under Grant 61379094, 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. The Associate Editor for this paper was D. Fernandez-Llorca. *(Corresponding author: Qi Wang.)*

Q. Wang is with the School of Computer Science, with the Unmanned System Research Institute, and also with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com).

J. Gao and Y. Yuan are with the Center for OPTical IMagery Analysis and Learning, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: gjy3035@gmail.com; y.yuan1.ieee@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2017.2726546

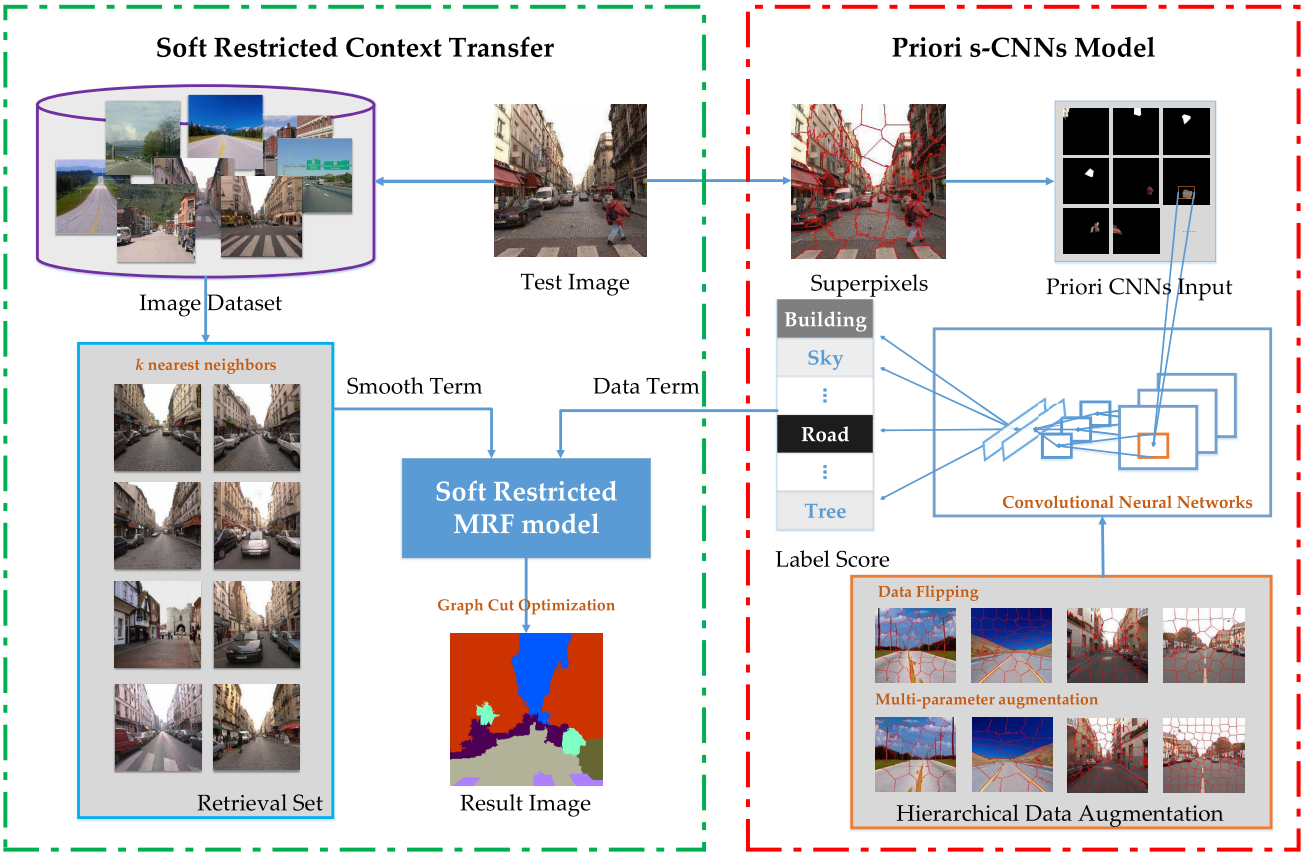


Fig. 2. The flowchart of our proposed joint method of priori s-CNNs and soft restricted context transfer. Firstly, given an input image, this paper generates a certain amount of superpixels. For learning priori location information, each superpixel is extracted from original image as a single input of priori s-CNNs. Then, the CNNs model outputs probability vectors (called as "label score") corresponding to each label. Secondly, the k nearest neighbors image retrieval searches for similar scenes to test image from the training set by global deep features. After obtaining retrieval set, this work computes conditional probabilities between adjacent superpixels. Finally, a soft restricted MRF model is constructed which integrates label score with priori probability between adjacent superpixels. Through optimizing MRF energy function to refine initial CNNs model's results.

information is exploited. However, the feature weight and fusion strategy are manually determined, and it is not easy to obtain some features, especially 3D features. Thus, how to automatically learn rich and discriminative features is an important issue that needs further research.

At the same time, labeling foreground objects is another intractable issue in the scene labeling. This is because the data distribution in the training set is unbalanced (called as "long-tail effect"). A few background objects (sky, road and building) account for the majority of the training data, while the foreground objects only take a small part. This phenomenon makes the training process or feature extraction are less adequate and the final result is: background objects labeling accuracy is far higher than foreground objects. However, because foreground objects may be more oriented to intelligent driving than background objects, such as traffic signs, pedestrian, surrounding preceding vehicle and so on, we think the foreground detection is more important to intelligent driving than background detection. Nevertheless, there is no approaches to solve it well because the long-tailed effect is a natural phenomenon and almost exists in every image. If the bias of the dataset can be reduced, the foreground objects detection will be promoted.

Therefore, our model focuses on how to learn more rich features and reduce the bias of dataset for labeling scene more accurately.

A. Overview of Our Approach

In this paper, we propose a joint method of superpixel-CNNs model and soft restricted context transfer to tackle the street scene labeling problem. The superpixel-CNNs model focuses on learning rich and discriminative features of image superpixels and exploiting priori location information effectively, which is called as "priori s-CNNs". The soft restricted context transfer aims to reduce the noises caused by the priori s-CNNs labeling results. The entire framework is illustrated in Fig. 2.

1) Training Priori s-CNNs With Anti-Bias Data Augmentation: In this stage, a priori s-CNNs is trained to label superpixels. At first, the training images are over-segmented to a certain amount of superpixels. Then all of them are input to the CNNs to train the model parameters by a supervised method. For a more efficient feature learning, the superpixel location prior is particularly considered in this procedure. Thus our CNNs feature does not only contain appearance but also reflect location information. At the same

time, in order to reduce the dataset bias, this work uses a hierarchical data augmentation to enlarge the original training set. It takes different numbers of training object classes into account and augments them separately. Thus the CNNs model can learn more rich features to describe superpixels.

2) Labeling Images With Context Smoothing: Given a test image, two processes are applied, initial label assignment and context transfer smoothing. The former aims to label each superpixel in the test image according to the previously learned model. But the obtained result is noisy and far from perfect. Therefore, the latter accordingly focuses on reducing the initial labeling noise by transferring contextual clue from the training set to the test image. To this end, we search for the k most similar images in the training set and transfer their structured labels to the examined test image, combined with an MRF post-optimization.

B. Contributions

In this work, we focus on learning more rich features to describe each superpixel and improving the problem of dataset bias. The main contributions of this work are threefold:

1) Learn rich feature (4096 dimensions) by finetuning the powerful CNNs (AlexNet, an image classifier network) to tackle our task - scene labeling. In order to get a coherent labeling result, we utilize a superpixel with location priors as an input unit instead of a traditional pixel based image. Our treatment keeps the structural relationship between the examined superpixel and the whole image and implicitly embeds the location prior in the CNNs processing. This is critical because the street scene understanding is highly dependent on the class spatial structure. For example, the sky is prone to be in the upper image and the road tends to be in the bottom. Thus, priori s-CNNs can extract rich feature for each superpixel to label scenes.
2) Propose a hierarchical data augmentation method to reduce overfitting and dataset bias. Traditional data augmentation expands the training data randomly and equally, which can not balance the number of different training classes. In order to tackle this problem, we propose to enlarge the training set in a more balanced manner. The classes with more training samples will be less augmented, and vice versa. Based on a well adjusted training set, the performance of the foreground objects labeling improve significantly.
3) Present a soft restricted MRF model to adapt to priori s-CNNs model's outputs and reduce over smoothness. Traditional approaches treat contributions of adjacent superpixels equally, which causes some foreground objects are smoothed improperly by its majority of adjacent background objects for consistence. In order to weaken this problem, adaptive weights are added to the smoothness term according to the similarity between the pairs. With such a soft restriction, our model can alleviate noises in the initial results and dose not result in serious over smoothness.

The rest of the paper is organized as follows. Section II reviews the related work briefly. Section III and Section IV describe the priori s-CNNs training process and soft restricted context transfer respectively. Section V shows the experimental results on two street scenes datasets and compares its performance with other competitors. In addition, some further discussion and analysis about some import modules in our methods are presented in this section. Finally, we summarize the work in section VI.

II. Related Work

In recent years, a large amount of approaches for scene labeling have been proposed. According to their pipelines, the algorithms usually consist of two components: extracting image feature and introducing contextual smoothness.

There are many methods for feature extraction. Liu *et al.* [13] use SIFT Flow feature to align the input image. Shotton *et al.* [8] define a texton and extract its texture, layout, and location information. Tighe and Lazebnik [9] compute around 20 features (five types: shape, location, texture, color and appearance) to describe superpixels. In addition to the above RGB features, some research [11], [14], and [15] exploit 3D features, such as 3D point clouds and depth maps. Brostow *et al.* [11] propose a method based on 3D point clouds derived from ego-motion. They design five cues (camera height, closest distance to camera, surface orientation, track density and back projection residual) to model patterns of motion and 3D structure. Xiao and Quan [14] propose a multi-view parsing method for image sequences. Zhang *et al.* [15] compute the scene depth information from video sequence through stereo reconstruction of dense depth maps. Peng *et al.* [16] propose an unsupervised subspace learning method which can automatically determine the optimal dimension of feature space.

In addition to the above hand-crafted features, deep feature is recently adopted to describe image. Compared to hand-crafted features, it can learn high-level features and fill in representation gap in some way. In an inchoate work, Grangier [17] propose a supervised greedy leaning scheme based on deep convolutional networks. The networks architecture can extract texture, shape, and contextual information. Farabet *et al.* [18] propose a method of learning hierarchical features based on multi-scale convolutional networks. However, because of the lack of training set, this method does not acquire a good results. Until 2014, Girshick *et al.* [19] solve this problem by representation transfer. They propose a region CNN (R-CNN), which use a high-capacity CNNs (AlexNet [20]) to process region proposal for localizing and segmenting object. Because the AlexNet's parameters are trained on ImageNet dataset, training model based on AlexNet can acquire its robust feature representation. After that, many similar methods exploit this strategy. Hariharan *et al.* [21] aim to detect all instances of a category in an image, and their algorithm is based on region proposals' features by extracting from both the region bounding box and the region foreground with a jointly trained R-CNN and box CNN. Shelhamer *et al.* [22] propose a fully convolutional networks based on AlexNet [20], VGG net [23] and

GoogLeNet [24], which only consists of convolutional layers without original fully connected layers.

As for the contextual information, Markov Random Field (MRF) and Conditional Random Field (CRF) models are very popular solutions. Early methods (e.g. [8]) exploit local feature information and smoothness prior adjacent pixels by defining second-order potential. Tighe and Lazebnik [9] define a prior conditional probability of adjacent superpixels as contextual information. For exploiting more wide contextual information, Ladicky *et al.* [25] introduce object detector terms into CRF function. Myeong *et al.* [26], [27] are based on [9] and model contextual relationships. Through learning the relationship of superpixels, the scheme transfer the object relationship from retrieved images to test images. Yang *et al.* [28] incorporate both local and global semantic context information via a feedback based mechanism to refine retrieval set and superpixels matching.

Besides the above two probabilistic graphical models, context and structure model is also a novel method. Generally speaking, the contextual information propagates in the trees, forests or networks. Sharma *et al.* [29] propose recursive neural network architecture (contains four networks) for the propagation of contextual information from a superpixel to other one through binary tree. Kontschieder *et al.* [30] exploit contextual and structural information in random forests by integrating the structured output predictions into a concise, global, semantic labeling. Shelhamer *et al.* [22] integrate appearance representation with semantic information from a shallow and a deep layer respectively. Peng *et al.* [31] propose a deep subspace clustering methods, which incorporates the structured global prior in representation learning.

In addition to the above two modules (feature extraction and contextual smoothness), it is important to mention the related works of data augmentation. In the real world, the objects' proportions are imbalanced because of the "long tail effect". In the field of knowledge discovery, imbalanced learning is a hot topic, which can affect the performance of learning algorithms in the presence of underrepresented data [32]. Data augmentation is one of imbalanced learning methods in the deep learning applications. For training AlexNet [20], Krizhevsky *et al.* apply image translation and horizontal reflections. They also alter the intensities of the RGB channels in the training images. Howard [33] extends image crops into extra pixels to capture translation and refection invariance, and adds randomly generated lighting which tries to capture invariance to the lighting and minor color variation. Wu *et al.* [34] adopt some color casting, vignetting and lens distortion to augment dataset, which can improve the CNNs' sensitivity to colors that are caused by the illuminants of the scenes.

III. Priori s-CNNs Based Feature Learning

This section mainly explains the training process. Based on a typical CNNs model, we transfer a robust representation to our specific application - scene labeling. For exploiting prior information, we propose a priori based s-CNNs. And for reducing dataset bias and getting a more balanced model, we propose a hierarchical data augmentation strategy. Thus, our CNNs model can learn rich and discriminative representations to describe images.

A. Priori s-CNNs

Scene labeling is a task that needs to annotate per pixel, but it is time-consuming to extract features for each pixel and construct a large graph to optimize the MRF energy function. We note that superpixel is a set of pixels that almost belong to the same class and have similar appearance and texture. Once a superpixel is classified as a label, the pixels in the superpixel are assigned as the same label. If we can regard each superpixel as a basic labeling unit, the time cost will decrease significantly. Based on this consideration, we propose a novel superpixel based CNNs, emphasizing the priori effect in the scene labeling application.

1) Convolutional Neural Networks: In this paper, in order to label street scenes datasets, we finetune AlexNet [20] that is pre-trained on ImageNet Large Scale Visual Recognition Challenge dataset (ILSVRC2012, 1.3 million images, 1000 object categories) by Caffe.[1] AlexNet consists of 5 convolutional layers and 3 fully connected layers (the last is soft-max layer). For finetuning it, a new soft-max layer replaces the original soft-max layer to predict street scene labeling classes (including the "void" class that is representative of the unannotated regions in the datasets).

2) Superpixel Generation: In this work, superpixel is a basic unit of labeling and each superpixel is produced by a typical and efficient method: simple linear iterative clustering (named as "SLIC") [35]. It adopts a k-means clustering algorithm to generate superpixels, which considers the color information in CIE-LAB space and the position of each pixel. SLIC has following advantages: 1) the boundaries of the generated superpixels are accurately; 2) the generation speed is fast. In a superpixel, nearly all pixels' labels are uniform and belong to the same class. Thus, it is reasonable to regard one superpixel as a processing unit.

3) Priori Superpixel Based Processing: After generating superpixels, we don't resize the them to the same dimension as the input. Alternatively, each superpixel is remained in the original image and the other outside areas are set as black color. Afterwards, these superpixel images enter into the CNNs model and the supervised parameter update is conducted during the training stage. The reason for this operation is explained as follows. In street scenes, we easily find that road region is usually located at the image bottom and the sky on the top. Taking full advantage of this location priori is essential to rule out the false labeling. Therefore, we propose this processing method, which can make CNNs learn location prior of superpixels and more discriminative feature.

Furthermore, we discuss the effect of location priori in the CNNs. In convolutional layers, because of the parameter sharing and small sizes of convolution kernels (11×11, 5×5 or 3×3 in AlexNet), the kernel's parameters are not changed by this strategy and they are only sensitive to object class. For example, a feature map O is a 3-D tensor with size of $H \times W \times H$, which is output by a convolutional layer. Here,

[1] http://caffe.berkeleyvision.org/

H is height, W is width and C is the number of channel in the feature map. The C-D vector at the i-th, j-th position in first two dimension represent the appearance information of the corresponding respective field in the input image. In addition, the entire feature map O is viewed as a permutation of the $H \times W$ C-D vectors. Such the permutation potentially contains the location information. Then, fully connected layers can integrate the last convolutional layer's feature map into a 4096-dimensional feature vector by inner product operation. Some neurons in these layers are sensitive to the data on the specific channel (the data are output by the specific kernel of the last convolutional layer) of the input. Thus, the fully connected layers model a relationship between appearance features and location priors. In other words, the neurons in fully connected layers can response to specific classes that often appear in specific regions while ignoring other classes. In summary, the fully connected layer can learn location priori for each superpixel.

B. Hierarchical Data Augmentation

In deep learning, overfitting caused by insufficient training data is a common phenomenon. One general method to alleviate overfitting is data augmentation that artificially expands the training set. Traditional strategies include image flipping, rotation, translation, rescaling, shearing, and so on. Unfortunately, these strategies share the same characteristic that all training data are expanded randomly and equally. Thus they cannot reduce dataset bias. Besides, some operations, especially rotation and translation, may change the location priori for the vehicle captured video. As a result, designing a self-tailored augmentation method is necessary.

We notice that common street scenes are roughly symmetric in the horizontal direction. Consequently, only horizontal flipping is adopted to avoid location prior changes when enlarging the training data. But this can not solve the dataset bias. In order to get a more balanced training set, different object classes should be augmented distinctively. Based on this consideration, a hierarchical data augmentation mechanism is presented to purposefully enlarge each class of training set.

To be specific, the objects in each training image are divided into four categories.

1) Majority objects: some background objects, such as sky, buildings and roads that count for the most part of an image;
2) Common objects: objects with label proportion more than 10% except "majority objects";
3) Unusual objects: objects with label proportion more than 3% and less than 10%;
4) Scarce objects: objects with label proportion less than 3%.

The label proportion is defined as $N_i / \sum_j N_j$, where N_i is number of pixels labeled as class i in the training image, and $\sum_j N_j$ is the number of image pixels. All foreground objects labels exist in the last three categories. As for the above four levels, in order to enlarge them to different extents, we present a multi-parameter data augmentation method to generate training data and the concrete implementation is described below.

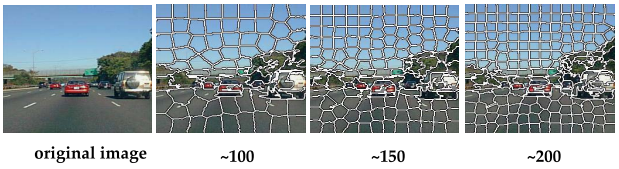


Fig. 3. The exemplar display of multi-parameter data augmentation. The number of superpixels is under each image.

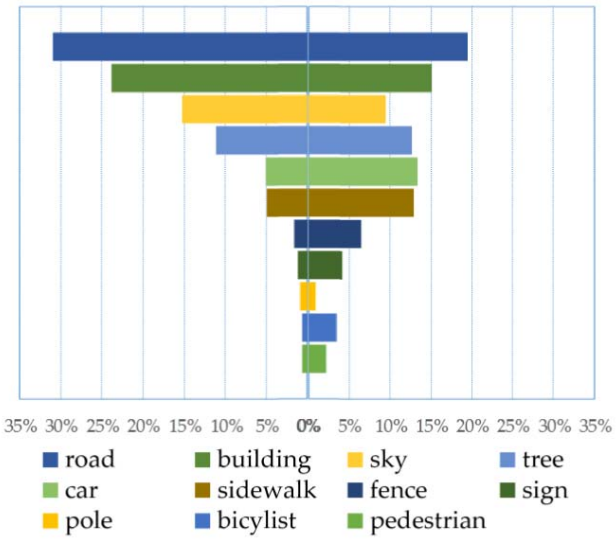


Fig. 4. The label distribution histogram of CamVid dataset [11]. The left is the label distribution of the original dataset before data augmentation, which shows a "Long Tail Effect". The right demonstrates a distribution after hierarchical data augmentation, which is more balanced among the different classes.

In our work, the superpixel is obtained by SLIC [35]. Under various parameters, each training image is over-segmented to different number of superpixels. However, not all of them are added to the training set, otherwise the dataset bias cannot be reduced. For complementary augmentation, object with less label proportion will acquire more augmentation with different parameter segmentations. Thus the majority objects get the least segmented superpixels as the training samples, while the scarce objects get the most. Eventually, a more balanced training set is achieved.

Actually, this multi-parameter augmentation also has the multi-scale effect. For example, in Fig. 3 the same image is segmented by different parameters of SLIC and some foreground objects with distinctive appearances and sizes are all involved in the training set. This makes the model learn more diverse and rich features.

In order to illustrate the effect of this self-tailored augmentation, we take CamVid [11] dataset as an example. The label distribution histograms before and after data augmentation are shown respectively in Fig. 4. The left part demonstrates clearly the "long tail effect" before augmentation. As can be seen from the bars, some background classes (sky, building and road) account for more than 70% of all training data, but the

important foreground objects (person, pole, fence, sign, etc.) only cover a few proportion. According to the proposed augmentation strategy, we divide these objects into four categories and expand their numbers through horizontal flipping and multi-parameter segmentation. The resulting proportion histograms alleviate the "long tail effect" greatly, which can be seen in the right part of Fig. 4. According to statistics, the number of superpixels in the training set increases from around 60,000 to more than 130,000 and the proportions of common, unusual and scarce objects raise greatly compared to the original training set. Therefore, we can say a more balanced training set is obtained after data augmentation.

C. Local Superpixel Labeling

With the above treatment, we can train the priori s-CNNs effectively. Then given a test image, the soft-max layer outputs a label score vector s for each superpixel, which represents the probability of being labeled as each class. Selecting the label with largest score as the superpixel's label and combining all the labels of superpixels in one test image forms the initial labeling result. However, only exploiting local feature is not enough, because the results may be noisy. Thus the initial results should be refined further.

IV. Soft Restricted Context Transfer

In Section III-C, the initial labeling results are obtained. Nonetheless, since the superpixels are individually examined, the spatial coherence needs to be improved further. SuperParsing [9] propose an effective method to utilize contextual information. It comprises two steps, nearest image retrieval and MRF optimization. However, the nearest image retrieval adopts hand-crafted features, which can not represent high-level image information. What's more, the traditional MRF model can result in over-smoothness. For solving the questions, we exploit deep features to search neighbors and propose a soft restricted MRF model, which can utilize internal difference in adjacent superpixels to alleviate the over-smoothness.

A. The k Nearest Image Retrieval

In order to transfer more accurate contextual information to the test image, the scenes that contain similar content structure should be considered. Thus we retrieve the k nearest images in the training set for the examined test image and exploit their contextual influence.

This system computes a deep global image features to search neighbors, which is 4096-D vector from fc7 layer of AlexNet. Here, the AlexNet is trained on Places Database (scene-centric databases, more than 7 million images, including 205 scenes categories), named as Places-CNN [36]. The model can effectively extract global feature of scenes. For each image in the training set, it will be ranked according to the increasing order of Euclidean distance to the test image on the computed 4096-D deep feature. Then the nearest k neighbors of the test image are chosen to transfer the contextual information in the next step (soft restricted MRF model inference). Compared to some traditional methods such as SuperParsing [9], deep features describe appearance and high-level semantic information better than hand-crafted features (More discussions about the advantages of deep features are presented in Section V-H). Thus, the more accurate contextual information is computed and transferred by Soft Restricted MRF Model in the next section.

B. Soft Restricted MRF Model Inference

For transferring contextual information from the retrieval set, the MRF model is a popular method. Given a superpixel, traditional methods employ its adjacent superpixels to smooth it equally. However, this is not a reasonable strategy. We think a pair of similar superpixels should smooth each other more than dissimilar pairs. Thus, we propose a soft restricted MRF model, which weights each adjacent superpixel to measure its contribution of spatial coherence.

We formulate an MRF energy function over the field of superpixel labels $\mathbf{l} = \{l_i\}$ as:

$$E(\mathbf{l}) = \sum_{s_i \in SP} E_d\left(s_i, l_i\right) + \lambda \sum_{\left(s_i, s_j\right) \in \varepsilon_w} E_s\left(l_i, l_j\right), \tag{1}$$

where SP is the set of superpixels in the test image, superpixel s_j is adjacent to superpixel s_i, ε_w is the set of edges of adjacent superpixels, and λ is a smoothing constant. The data term $E_d\left(s_i, l_i\right)$ denotes the cost of assigning superpixel s_i with label l_i and the definition is:

$$E_d\left(s_i, l_i\right) = \left(A^s_{s_i} - A^r_{s_i}\left(l_i\right)\right)^2, \tag{2}$$

where $A^s_{s_i}$ is the output label score vector for superpixel s_i from the priori s-CNNs, $A^r_{s_i}\left(l_i\right)$ is the observation value, an indicator vector whose l_i-th item is set as 1 and others 0. Suppose that the p-th item of label score vector have largest probability, if $l_i = p$, the $E_d\left(s_i, l_i\right)$ will be smallest and vice versa. The smoothness term $E_s(l_i, l_j)$ stands for the cost of a suerpixel smoothed by adjacent superpixels. If a pair of adjacent superpixels appear in the retrieval set frequently, the smoothness term should be small. This term is defined based on probabilities of label co-occurrence statistics:

$$E_s\left(l_i, l_j\right) = -w_{ij} \times \log\left[\frac{P(l_i|l_l) + P\left(l_j|l_i\right)}{2}\right] \times \delta\left[l_i \neq l_j\right], \tag{3}$$

where $P(l_i|l_j)$ is the conditional probability of assigning label l_i to the superpixel given its neighbor has label l_j, which is estimated from its corresponding retrieval set. $w_{i,j}$ is a soft restriction on the smoothness term, which represents the contribution of each pair of adjacent superpixels. It is defined as the squared Euclidean distance between label scores of two adjacent superpixels:

$$w_{ij} = \left(A^s_{s_i} - A^s_{s_j}\right)^2, \tag{4}$$

where $A^s_{s_i}$ and $A^s_{s_j}$ denote the label score of superpixel s_i and s_j. The more alike the adjacent superpixels are, the smaller w_{ij} is, and vice versa. Thus, w_{ij} enhances the smoothness between similar superpixels and reduces the smoothness between distinguished superpixels. The last

factor $\delta\left[l_i \neq l_j\right]$ can be considered as a Potts penalty, which is necessary to ensure that this term is semi-metric [37]. It is defined as below:

$$\delta\left[l_i \neq l_j\right] = \begin{cases} 0 & if\ l_i = l_j \\ 1 & if\ l_i \neq l_j \end{cases} \tag{5}$$

If the assigned labels for the two adjacent superpixels are the same, the smoothness term is not necessary and should be set as 0.

Exploiting the prior conditional probability aims to reduce labeling errors. For example, if a superpixel is a part of a person, it may be assigned with a label "pedestrian" or "bicyclist". But if its adjacent superpixels are likely to be "sidewalk", it is more probable to label it with "pedestrian" according to the learned prior conditional probability from the retrieval set. We perform MRF inference using an efficient graph cut optimization[2] [37]–[39].

V. Experiment

In this section, we report experimental details and results on the two challenging datasets: CamVid [11] and SIFT Flow Street dataset. Section V-A shows the two evaluation criteria in scene labeling. Section V-B presents some details and characteristic of the two datasets. Section V-C gives parameter setup and implementation details in the experiments. Then, the results and discussions are presented in Section V-D and V-E. Finally, we discuss the effects of the proposed priori location, the advantages of the soft restricted MRF model, the comparison between CNN and hand-crafted features in image retrieval and convergence issues of the stepwise models in last four sections (V-F, V-G, V-H and V-I).

A. Evaluation Criteria

In the scene labeling field, there are two metrics to evaluate each algorithm's performance: per-pixel accuracy and mean-class accuracy. The former is defined as:

$$r_p = \frac{\sum_i n_{ii}}{\sum_i \sum_j n_{ij}}, \tag{6}$$

where n_{ij} is the number of pixels assigning label i as label j, $\sum_i \sum_j n_{ij}$ and $\sum_i n_{ii}$ stand for the total number of pixels and total number of pixels that are assigned correct label, respectively. However, because the label distribution suffers from unbalanced problem in practice, only adopting the per-pixel accuracy is not precise. Moreover, in the street scenes, the foreground objects are essential to safe driving, but their contribution to per-pixel accuracy is limited. Therefore, a more reasonable criterion should be introduced. Specifically, the mean-class accuracy is defined as below:

$$r_c = \frac{1}{N} \sum_i \frac{n_{ii}}{\sum_j n_{ij}}, \tag{7}$$

where N denotes the number of the label classes. It is an average of per-pixel accuracy of each class, so it can evaluate the overall performance at the class level.

[2]The C++ code and MATLAB wrapper are developed by O. Veksler and A. Delong and available at http://vision.csd.uwo.ca/code/gco-v3.0.zip

TABLE I

The Detail Information of CamVid Dataset is Shown as Below, Including Sequence Name, the Number of Frames, Data Type and Scene Category

Video sequence	Frame no.	Type	Scene
0001TP-1	62	train	dusk
0016E5	305	train	daytime
0006R0	101	train	daytime
0001TP-2	62	test	dusk
Seq05VD	171	test	daytime

B. Dataset

1) CamVid Dataset: The Cambridge-driving Labeled Video Database (CamVid)[3] is a challenging road driving scenes dataset, which includes 4 video sequences (one video is divided into 2 parts) with image size of 960×720 pixels. Similar to [9], [30], and [40],we merge the 32 object classes of the original dataset into 11 classes. They are road, building, sky, car, sign-symbol, tree, pedestrian, fence, column-pole, sidewalk and bicyclist. Table I shows the detailed information of CamVid dataset.

2) SIFT Flow Street Dataset: The original SIFT Flow dataset[4] consists of 2,688 images of 33-class outdoor scenes, which is selected from LabeleME [41] and annotated by LabelME's users. These outdoor scenes include coast, forest, highway, inside city, street scenes and so on, with a resolution of 256×256. For doing the experiments in the specific street context, we only choose a part of them as our dataset which is called "SIFT Flow Street Dataset".

To be specific, we select the highway and street scenes from SIFT Flow dataset and remove those images that are not from the perspective of vehicles. The new dataset consists of 529 images (491 training images and 38 testing images are selected from original training and testing sets respectively). At the same time, the original label classes are updated by removing the unrelated labels. Eventually, there are 16 classes (road, sky, sidewalk, building, tree, car, field, fence, person, crosswalk, sign, streetlight, bus, bridge, window, and mountain) in the SIFT Flow Street dataset.

C. Implementation Details & Settings

1) Settings of the Priors s-CNNs: As for each image (training or testing), it is resized to 256×256px to adapt to the CNNs model and is oversegmented to ~ 150 superpixels (we treat "~ 150" as "the main parameter"). In the training priori s-CNNs stage, the learning rate is initialized at 10^{-4} and reduced ten times every ten thousand iterations. Our models are only sensitive to the learning rate: the smaller value selection results in the slower convergence speed and the higher loss. On the contrary, setting the more larger learning rate does not makes the model converge.

2) Settings of the Hierarchical Data Augmentation: As for the data augmentation, the majority objects are not enlarged; the other training samples are horizontally flipped but they are

[3]http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid/

[4]1http://people.csail.mit.edu/celiu/LabelTransfer/LabelTransfer.rar

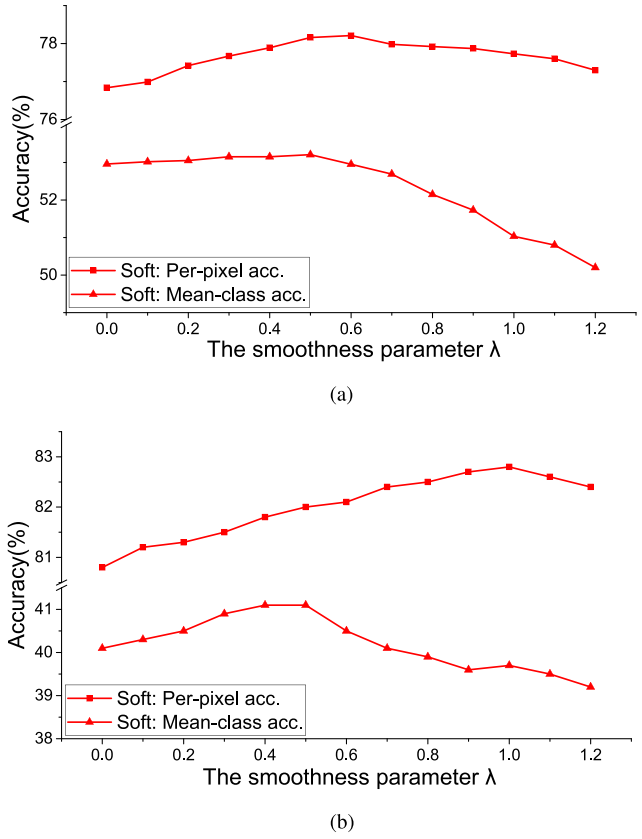


Fig. 5. The red solid lines demonstrate the effects of our proposed soft restricted MRF model under the different smoothness parameter λ. The (a) is on CamVid dataset, and the (b) is on SIFT Flow Street dataset.

segmented by different parameters. Specifically, the common objects are expanded under the main parameter; the unusual objects are augmented under the parameters of 100, 125 and 200. In addition to the above parameters, the scarce objects are expanded under more parameters, including 175, 130 and 170. With this strategy, a more balanced training set is obtained.

3) Settings of the Context Transfer: In the k nearest image retrieval, k is set as a default value 50 [9], which can achieve the best mean-class accuracy. Another important parameter is λ in the soft restricted MRF model. Fig. 5 demonstrates the performance under different λ choices on the two datasets. As can be seen from the red lines, the mean-class accuracy is almost stable at the beginning and decreases with the increase of λ. The per-pixel accuracy increases firstly and then decreases. This is because with the increase of λ, the labels with small area (e.g., foreground objects) are over-smoothed. Therefore, a moderate λ might be appropriate. Since it is more important to obtain a high mean-class accuracy than per-pixel accuracy for preserving the foreground objects, λ is set to 0.5 in this work.

After setting the above parameters, the entire model will perform automatically and without manual operation.

4) Settings of the Compared Algorithms: For showing the superiority of our method, the five mainstream algorithms are added to the comparison. They are SuperParing [9], LLD [40], LOR [26], SLiRF [30], THSRT [27] and FCN [22].

TABLE II
COMPARISON OF DIFFERENT APPROACHES ON CAMVID DATASET

Methods	Per-pixel	Mean-class
SuperPasing [9](Still Image)	78.6%	43.8%
LLD [40]	73.7%	36.6%
LOR [26]	72.5%	35.7%
THSRT [27]	73.1%	35.7%
SLiRF [30]	72.5%	51.4%
FCN-32s (AlexNet) [22]	80.1%	44.7%
FCN-8s (AlexNet) [22]	**80.8%**	47.4%
Ours methods:		
Baseline	77.1%	45.6%
Data Flipping	77.2%	47.9%
Hierarchical Augmentation	76.8%	53.0%
Full Model	78.1%	**53.2%**

The first five traditional approaches all exploit hand-crafted features: SuperParsing, LOR and THSRT use 20 features to represent superpixels; LLD designs a local label descriptor by concatenating label histogram; and SLiRF exploits low-level image features. The last two, FCN-32s and FCN-8s [22] that are finetuned on AlexNet, exploit the fully convolutional network to labeling scenes end-to-end. Because of no source code, we do not test LLD [40] and SLiRF [30] on SIFT Flow Street dataset.

D. Performance on CamVid Dataset

Table II shows the two metrics of different comparative methods. At first, the baseline only uses the original data to train CNNs model. Then the traditional data flipping and our hierarchical data augmentation are added to the training process respectively. The last one is the soft restricted MRF inference based on the CNNs model with the hierarchical data augmentation (called as "full model").

It can be seen that our per-pixel accuracy of 78.1% is not the best, but for mean-class accuracy, our full model achieves the best performance. Considering the above criteria collectively, our results is the best in all of the methods. On the one hand, compared with the traditional strategies, our priori s-CNNs can learn more discriminative features to describe various objects. Thus, more foreground objects are labeled accurately. On the other hand, compared with the FCN-xs [22], our model labels the foreground objects more accurately.

Next, we discuss the effects of hierarchical data augmentation. Our proposed augmentation method can improve the mean-class accuracy (from 45.6% to 53.2%, increasing by 16.7%) more significantly than traditional data flipping (from 45.6% to 47.9%, increasing by 5.0%). But for the per-pixel accuracy, the improvement is not obvious. The reason is that the labeling performance of foreground objects increases but the background objects' drops simultaneously. More discussions will be presented in the next paragraph.

In order to analyze the labeling performance further, the results of each class are shown in Table III. According to the distribution of bold statistics, we find that the best performance of almost all foreground classes are in the bottom

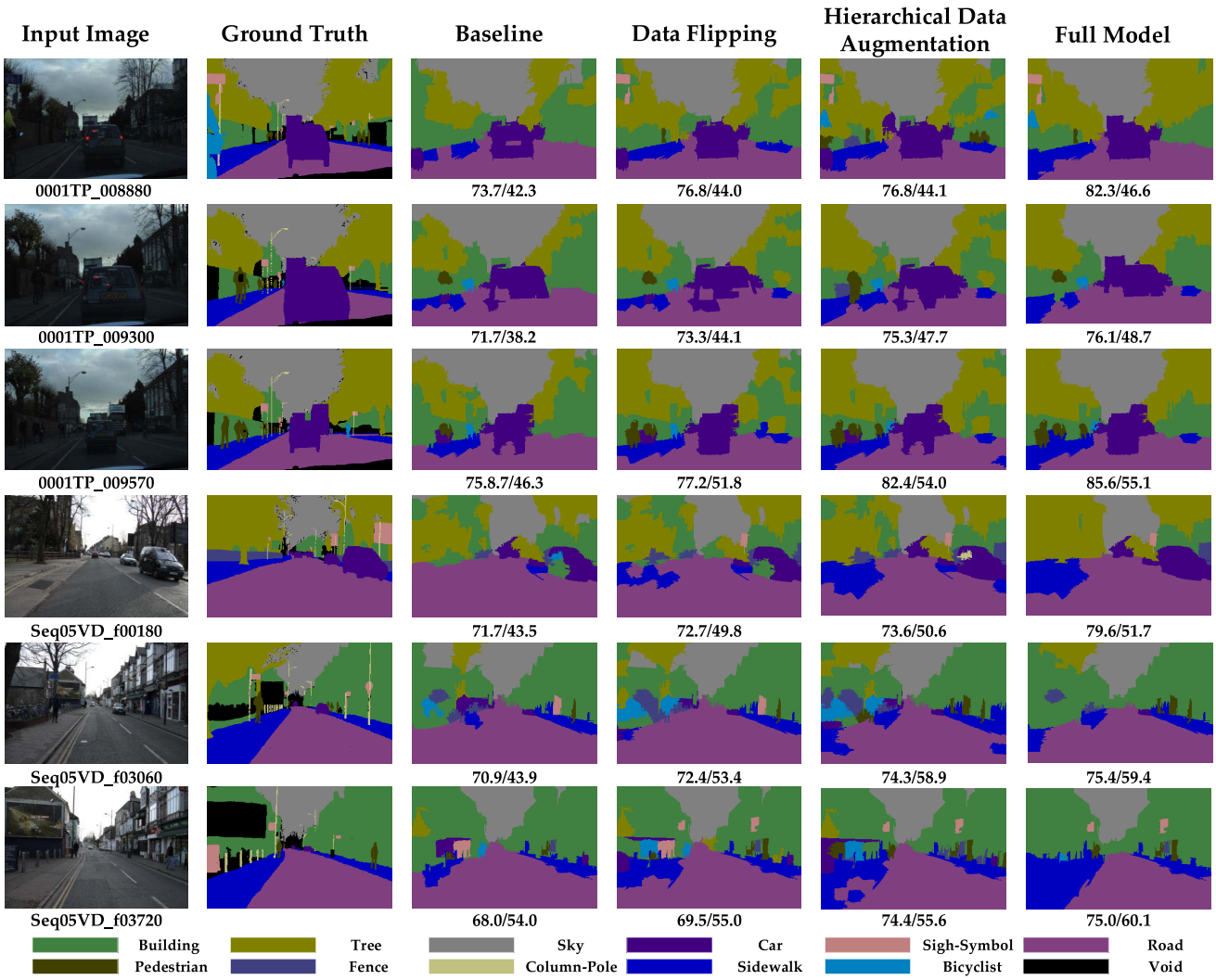


Fig. 6. Exemplar results on CamVid dataset. We report the four comparative results, namely the baseline, baseline+data flipping, baseline+hierarchical data augmentation and full model (baseline+hierarchical data augmentation+soft restricted MRF inference), respectively. The values under each image are the per-pixel/mean-class accuracies. The test images of the first three rows are from dusk sequence "0001TP" and the others are selected from the daytime sequence "Seq05VD".

two rows which utilize the hierarchical data augmentation. However, there is an exception, namely, "column-pole", whose performance is not good (2.2% versus the best 4.1% [40]). The reason is that the "column-pole" training data can not be segmented perfectly by SLIC and is not enlarged effectively (as shown in Fig. 4). We also notice that the accuracy of the majority objects (sky, building and road) decreases slightly after data augmentation. This is because with the increase of foreground labeling, the boundary pixels that previously belong to the background change their labels due to the unprecise superpixel segmentation.

For reporting the advantages of the our algorithm, Fig. 6 shows six typical exemplar labeling results. At first, we show the impacts of the hierarchical data augmentation on the labeling results. Without the data augmentation, the "tree" and "building" are prone to be mixed. After the hierarchical data augmentation, they are distinguished more clearly. In addition, more other foreground objects (sidewalk, pedestrian, sign and so on) are also labeled. For example, in the first input image, by our data augmentation the two signs are labeled correctly; in the second, third and last input images, several persons are labeled as "pedestrian" after data augmentation. Secondly, we present the effects of soft restricted context transfer. In the forth input image, the parts of the car are mislabeled as building, bicyclist and column-pole without the soft restricted MRF model inference. After considering contextual information by the MRF model, the car can be labeled entirely and accurately. In the fifth image, the left building is recognized as fence, bicyclist, car, pedestrian and so on. After smoothing this result, the error is mitigated considerably.

E. Performance on SIFT Flow Street Dataset

The results of SuperParsing [9], LOR [26], THSRT [27], FCN [22] and our models are listed in Table IV. From the table, we can see our full model achieves an excellent result (82.0% per-pixel accuracy and 41.1% mean-class accuracy). Obviously, our mean-class accuracy of 41.2% outperforms the SuperParsing (32.8%) [9], LOR (34.2%) [26],

TABLE III
COMPARISON OF PER-CLASS ACCURACY WITH SUPERPASING [9], LLD [40], LOR [26] AND FCN [22] ON CAMVID DATASET

	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Sidewalk	Bicyclist
SuperPasing [9](Still Image)	84.8	65.1	94.7	47.5	24.6	96.2	8.3	9.1	3.4	43.7	3.9
LLD [40]	80.7	61.5	88.9	16.4	-	**98.0**	1.1	0.01	**4.1**	12.5	0.01
LOR [26]	84.3	29.4	93.1	45.6	1.0	94.0	1.3	0.5	1.3	39.5	2.6
THSRT [27]	**87.2**	27.7	91.9	43.2	0.4	93.9	1.4	0.03	0.4	43.4	3.1
FCN-32s (AlexNet) [22]	85.5	63.6	90.3	63.4	10.4	94.1	5.0	10.7	0.0	69.0	0.3
FCN-8s (AlexNet) [22]	82.3	67.8	92.2	66.0	15.3	94.2	7.1	22.0	0.1	**71.8**	2.6
Our methods:											
Baseline	84.9	60.8	**95.3**	63.7	22.0	96.2	24.0	15.0	1.0	23.3	15.0
Data Flipping	79.1	70.0	94.2	67.4	26.6	95.2	28.3	16.9	2.1	30.5	17.0
Hierarchical Augmentation	70.4	73.9	93.6	68.8	**31.0**	92.4	**38.9**	**32.7**	2.3	50.3	28.4
Full Model	74.4	**74.9**	93.8	**69.8**	29.6	92.6	38.5	29.8	2.2	52.0	**29.0**

TABLE IV
COMPARISON OF DIFFERENT APPROACHES ON SIFT FLOW STREET DATASET

Methods	Per-pixel	Mean-class
SuperParsing [9]	79.9%	32.8%
LOR [26]	84.3%	34.2%
THSRT [27]	83.7%	33.2%
FCN-32s (AlexNet) [22]	84.0%	36.8%
FCN-8s (AlexNet) [22]	**84.7%**	37.2%
Ours methods:		
Baseline	80.9%	32.0%
Data Flipping	80.8%	36.3%
Hierarchical Augmentation	80.7%	40.1%
Full Model	82.0%	**41.1%**

THSRT (33.2%) [27] and FCN (36.8% and 37.2%) [22]. Compared with these mainstream methods, our model is trained on a more balanced dataset, which can learn rich and discriminative features to describe various objects. Thus, our model achieves the best mean-class accuracy. However, our per-pixel accuracy is not the best but it is close to the best (best 84.7% [22]). As a whole, our performance is competitive on the two criteria compared to other popular methods.

In addition to the above comparison with other approaches, we discuss the effects of different steps for the proposed method. Obviously, the mean-class accuracy of the original baseline is not very high. But after the traditional data flipping which augment the training set, the performance increases from 32% to 36.3%. The incensement becomes larger with the hierarchical data augmentation for a more balanced data augmentation (40.1%) and with a soft restricted MRF optimization for a smoother labeling (41.1%). These statistics give a hint that the proposed method is more effective than the competitors.

Fig. 7 illustrates the per-pixel accuracy, overall per-pixel accuracy and mean-class accuracy. The data statistics are similar to that of CamVid dataset: the performance of background objects decrease a little, and many foreground objects are promoted dramatically. The significant improvement of foreground objects labeling benefits from our priori s-CNNs trained on more foreground data after the hierarchical data augmentation. However, some foreground objects are not labeled correctly, such as "streetlight", "bus" and "window". The "streetlight" is similar to the "pole" in CamVid dataset, so it can not be segmented effectively and augmented. And the "bus" can not be trained enough because the training data are so rare that the effort of data augmentation is limited. The "window" is misclassified as "building" in the labeling stage.

Four typical results are shown in Fig. 8 to explain the effects of the hierarchical data augmentation and the soft restricted MRF inference. From (b) and (c), the "sidewalk" can be labeled more accurately with the hierarchical data augmentation than the baseline and traditional data flipping. In (c), the car in the right road is mislabeled as "road" by the first two methods, but our proposed augmentation method can label it correctly. In addition to the above intuitive displays, the statistics under the labeling images also illustrate the advantages of our augmentation strategy: the mean-class accuracy of each of the above images is promoted significantly by our data augmentation. As for the soft restricted MRF inference, in (b), the region mislabeled as "road" sidewalk shrink significantly because of its adjacent superpixels' smoothness. Similarly, some noises in the left (a) and (c) are reduced by contextual smoothing.

F. Effect of Prior s-CNNs

In the Section III-A, the learned location priors are explained theoretically. In order to show the effect of priori s-CNNs intuitively, the verified experiments are added. For comparison, the s-CNNs without location priors are trained on the two datasets. To be specific, during the training and testing stages, each superpixel is shifted by a random value at the x and y axes respectively in the original image, which removes the location priors from the superpixels input. In practice, the input superpixel image, the shift

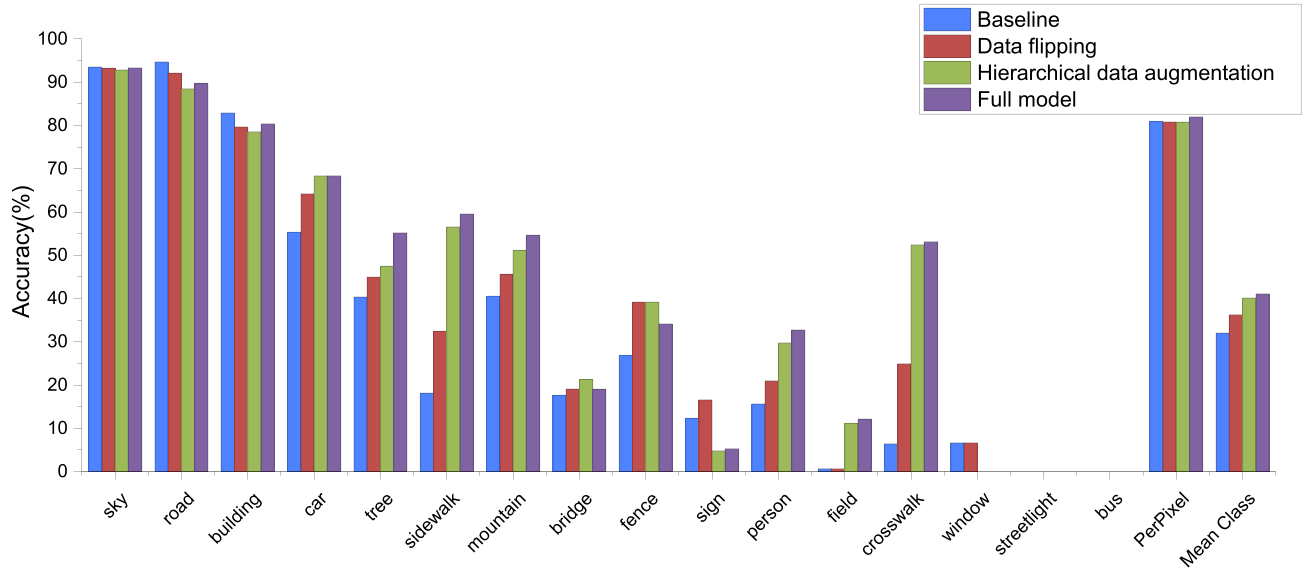


Fig. 7. The performance of each class and two metrics in the four stages.

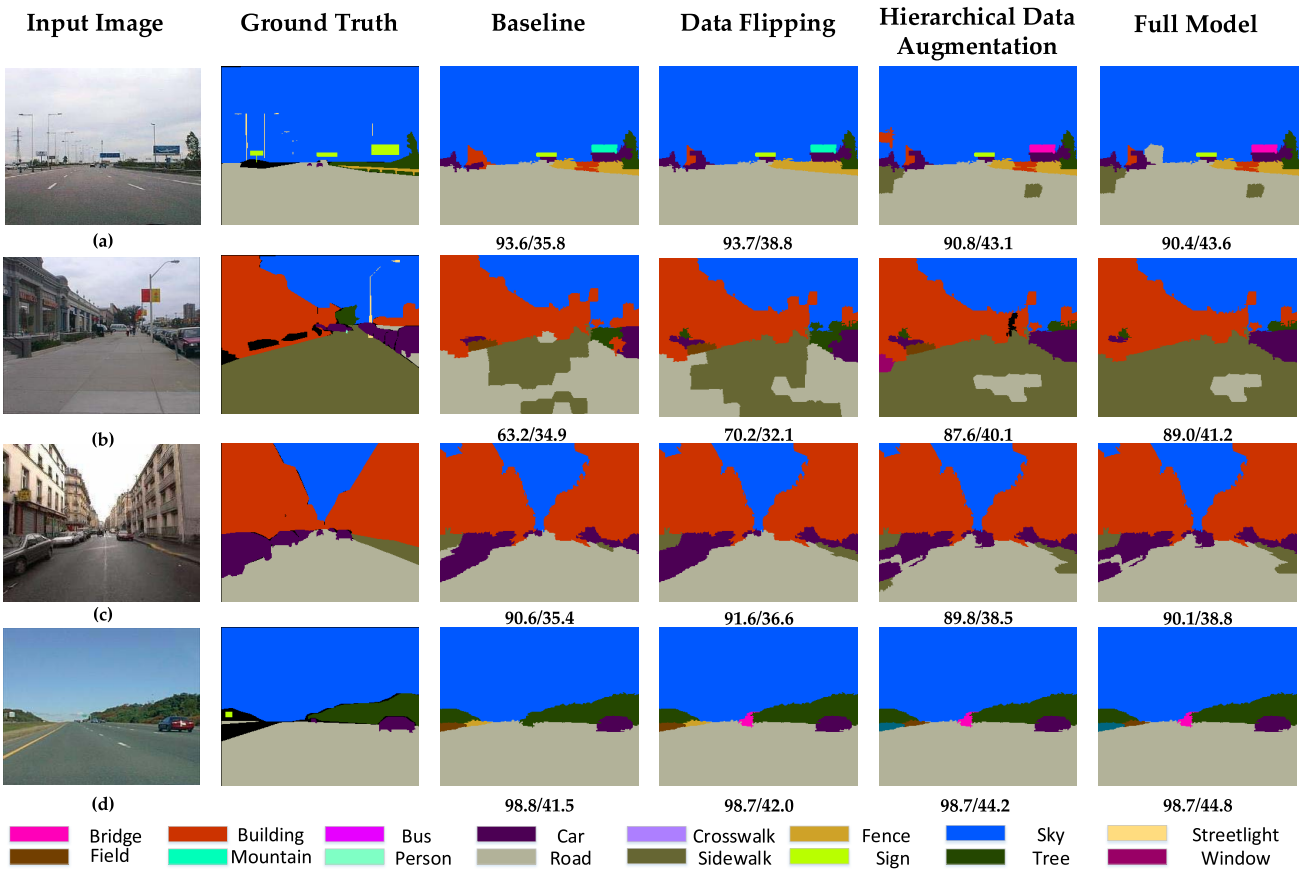


Fig. 8. Exemplar results from SIFT Flow Street dataset. The value under the each image labeled is the percentage of per-pixel and mean-class accuracy respectively. (a) highway_gre661. (b) street_bost136. (c) street_par177. (d) highway_bost164.

values $\Delta x, \Delta y \in [-255, 255]$ (image size is 256×256) are generated randomly on the X and Y-axis, respectively. If the superpixel is moved to the outside of the image, the Δx and Δy are regenerated until the new location of the superpixel is still in the image. That way, in an input image, all superpixels are move to a different random location, which eliminates the location priors in the original image. And the s-CNNs focuses on learning the features from the appearance information. The quantitative results are shown as in Table V. Specifically, the performance of the proposed priori s-CNNs is

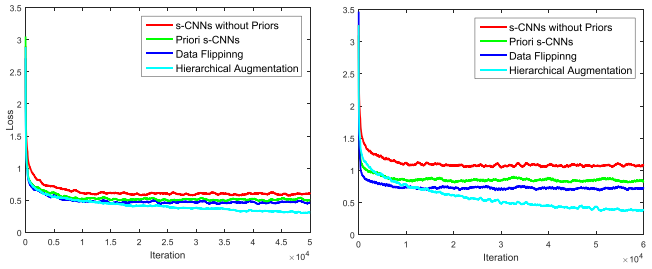


Fig. 9. Convergence curves of the stepwise models: s-CNNs without priors, priori s-CNNs, data flipping CNNs and hierarchical augmentation CNNs. The left and right present the results on CamVid and SIFT Flow Street dataset, respectively. (a) CamVid dataset. (b) SIFT Flow Street dataset.

TABLE V
COMPARISON OF S-CNNS WITHOUT PRIORS AND PRIORI S-CNNS ON THE TWO DATASETS

Methods	Per-pixel	Mean-class
CamVid dataset		
s-CNNs without priors	70.6%	35.2%
priori s-CNNs (Baseline)	**77.1%**	**45.6%**
SIFT Flow Street dataset		
s-CNNs without priors	79.8%	25.8%
priori s-CNNs (Baseline)	**80.9%**	**32.0%**

superior to that of s-CNNs without priors on the two datasets, which verifies the effectiveness of the former. In addition, the convergence curves of both models during training stage are shown in Fig. 9. Obviously, the priori s-CNNs converges a lower loss value than s-CNNs without priors.

G. Effect of Soft Restricted MRF

For explaining the advantage of the proposed soft restricted MRF intuitively, the comparisons with traditional hard MRF [9] on the two datasets are illustrated in Fig. 10. As can be seen from the bar chart, our method obtains higher mean-class accuracy than the traditional hard MRF, while the traditional method achieves a better per-pixel accuracy. But for the intelligent driving application, traditional hard MRF is not a good strategy, which sacrifices the performance of foreground labeling to get more overall per-pixel accuracy, because the foreground objects are more essential to safe driving than backgrounds. Therefore, the mean-class accuracy is more important than per-pixel accuracy. From this perspective, our model is much superior to the traditional model.

H. Comparison of CNNs v.s. Hand-crafted Features for Image Retrieval

In the k nearest image retrieval, the deep features are exploited instead of the hand-crafted global features, such as spatial pyramids, GIST and RGB-color histograms in Super-Parsing [9]. Because of classification capability of AlexNet, the 4,096-D feature in fc7 layer can represent the appearance and semantic information better than traditional features. Thus, the more accurate contextual information will be transferred to test images.

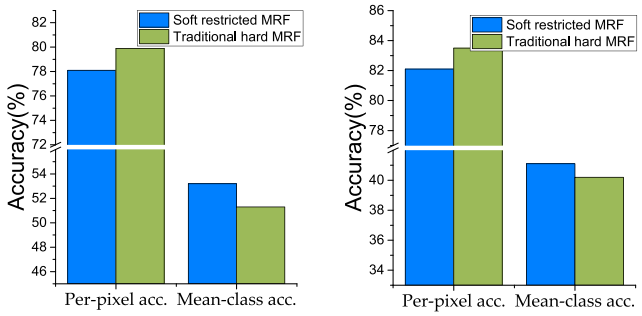


Fig. 10. Comparison of our proposed soft restricted MRF and traditional hard MRF at the optimal *lamda* 0.5. (a) CamVid dataset. (b) SIFT Flow Street dataset.

TABLE VI
COMPARISON OF CNNS V.S. HAND-CRAFTED FEATURES FOR IMAGE RETRIEVAL

Methods	Per-pixel	Mean-class
CamVid dataset		
Hand-crafted features	77.7%	53.0%
CNNs features	**78.1%**	**53.2%**
SIFT Flow Street dataset		
Hand-crafted features	81.4%	40.2%
CNNs features	**82.0%**	**41.1%**

In order to show advantages of deep features, we exploit two types of features to find similar images and transfer contextual information. Table VI shows the results of the two types of features in the full model. From the results, the improvement is not significant on CamVid dataset. To be specific, the retrieval results by two types of features are not so different. The main reason is that CamVid dataset includes continuous and similar image sequence and the retrieval set can be easily and accurately searched by hand-crafted features. On the SIFT Flow Street dataset, the improvement is obvious because of various scenes in the dataset so that the CNNs features demonstrate significant superiority.

For showing the difference of the two types of features intuitively, we select two typical test images from the two datasets and display respectively the retrieval results in Fig. 11. The left larger images are the query samples, and the right small images are the retrieval sets. Because of the limited space, the top-25 retrieved images are only displayed. About the "0001TP_008910" image in CamVid dataset, although the results of the two methods are similar as a whole, there are some subtle difference. The hand-crafted features are so sensitive to the color information that they ignore the images from other scenes. However, the CNNs features' results include some images that have the same content with the test image from other scenes. As for the test image "highway_1836030" in SIFT Flow Street dataset, the KNN that exploits CNNs features finds more similar images than the KNN that adopts three hand-crafted features. This is because the CNNs features describe more higher-level image representation including appearance, contextual and structural information than traditional hand-crafted features.

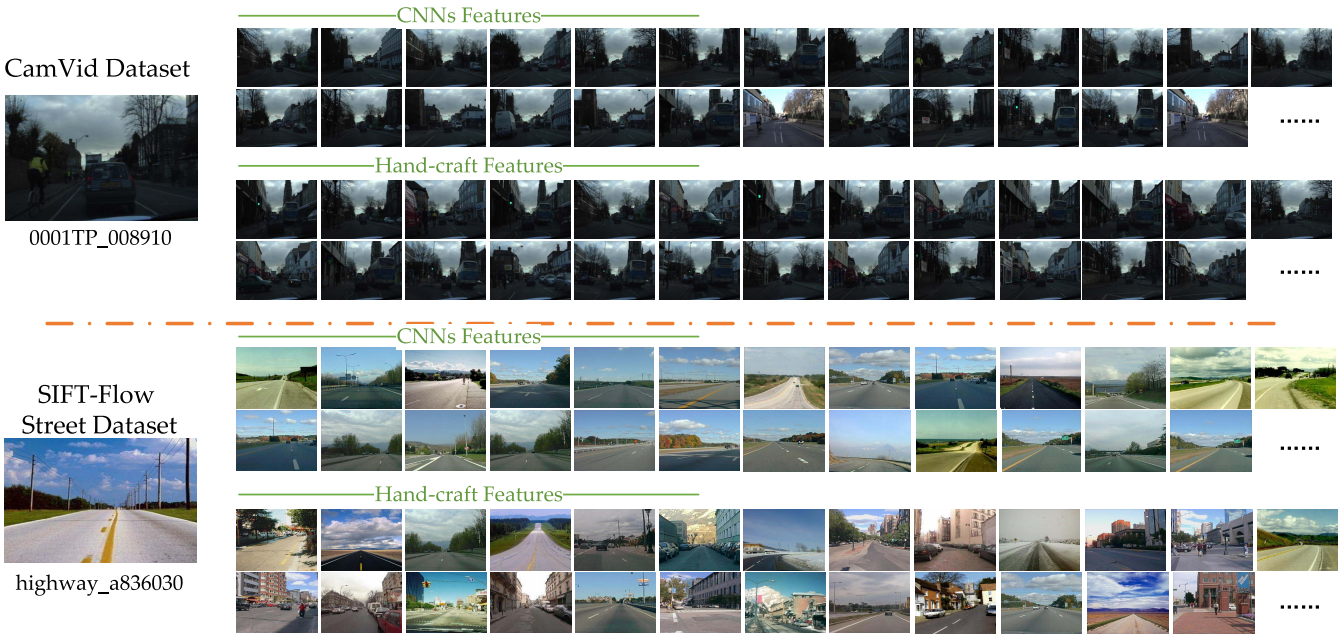


Fig. 11. The exemplar display of the retrieval sets generated by different global image features (CNNs versus hand-crafted features).

I. Convergence Analysis of the Stepwise Models

In this Section, we show the convergence curves of each stepwise models during the training stage, namely s-CNNs without priors, priori s-CNNs, data flipping CNNs and hierarchical augmentation CNNs in Fig. 9. As can be seen from the loss curves, the first three models converge after around 20, 000 iterations. The last model on the two datasets, however, converge at about 50, 000 and 60, 000 iterations, respectively. The main reasons are aggravating imprecise segmentation noises and increasing training samples caused by the hierarchical data augmentation. But eventually, the hierarchical data augmentation CNNs converges a lower training loss value and achieves a higher classification performance on the test set than the former three.

VI. Conclusion and Future Work

This paper proposes a joint framework of priori s-CNNs and soft restricted context transfer for street scenes labeling. The priori s-CNNs can fully exploit priori information through preserving superpixels' location in the image. Besides, it learns rich and discriminative features by the proposed hierarchical data augmentation. Compared with the traditional equal and random data augmentation, the proposed strategy can not only improve foreground objects labeling and mean-class accuracy significantly but also maintain the background objects labeling performance at the high level. In the context transfer, our proposed soft restriction on the smooth term of the MRF energy function can effectively reduce over smoothness, which makes the foreground objects not be improperly smoothed by the adjacent background objects. Extensive experiments have verified the effectiveness of the proposed method on the street scene datasets. Not limited to these street scenes, the proposed method also applies to other scenes (such as indoor and clothing parsing scenes) because no specific scene constraints are supposed in our approach.

With the proposed framework, the labeling accuracy of the foreground objects increases significantly. Nevertheless, the missing and false labeling phenomena are common in our results. Thus, we will focus on integrating objects detector into our model to enhance the labeling accuracy in the future.

References

[1] S. Zhang, C. Bauckhage, and A. B. Cremers, "Efficient pedestrian detection via rectangular features based on a statistical shape model," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 763–775, Apr. 2015.

[2] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, May 2017.

[3] T. Chen, R. Wang, B. Dai, D. Liu, and J. Song, "Likelihood-field-model-based dynamic vehicle detection and tracking for self-driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 11, pp. 3142–3158, Nov. 2016.

[4] Q. Wang, J. Fang, and Y. Yuan, "Adaptive road detection via context-aware label transfer," *Neurocomputing*, vol. 158, pp. 174–183, Jun. 2015.

[5] Y. Na, Y. Guo, Q. Fu, and Y. Yan, "Cross array and rank-1 MUSIC algorithm for acoustic highway lane detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 9, pp. 2502–2514, Sep. 2016.

[6] W. A. Wright, "Image labelling with a neural network," in *Proc. Alvey Vis. Conf.*, 1989, pp. 1–6.

[7] F. Nie, J. Li, and X. Li, "Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification," in *Proc. IJCAI*, 2016, pp. 1881–1887.

[8] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "*TextonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 1–15.

[9] J. Tighe and S. Lazebnik, "Superparsing—Scalable nonparametric image parsing with superpixels," *Int. J. Comput. Vis.*, vol. 101, no. 2, pp. 329–349, 2013.

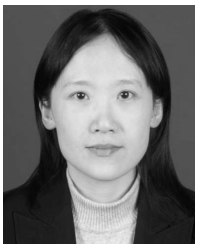
[10] X. Li, B. Zhao, and X. Lu, "A general framework for edited video and raw video summarization," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3652–3664, Aug. 2017.

[11] 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.

[12] S. Sengupta, E. Greveson, A. Shahrokni, and P. H. S. Torr, "Urban 3D semantic modelling using stereo vision," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2013, pp. 580–585.
[13] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1972–1979.
[14] J. Xiao and L. Quan, "Multiple view semantic segmentation for street view images," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 686–693.
[15] C. Zhang, L. Wang, and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 708–721.
[16] X. Peng, J. Lu, Z. Yi, and R. Yan. (2014). "Automatic subspace learning via principal coefficients embedding." [Online]. Available: https://arxiv.org/abs/1411.4419
[17] D. Grangier, L. Bottou, and R. Collobert, "Deep convolutional networks for scene parsing," in *Proc. Int. Conf. Mach. Learn. Deep Learn. Workshop*, vol. 3. 2009, pp. 1–8.
[18] 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.
[19] 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.
[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
[21] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 297–312.
[22] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[23] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[24] C. Szegedy *et al.* (2014). "Going deeper with convolutions." [Online]. Available: https://arxiv.org/abs/1409.4842
[25] 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. 2009, pp. 739–746.
[26] H. Myeong, J. Y. Chang, and K. M. Lee, "Learning object relationships via graph-based context model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2727–2734.
[27] H. Myeong and K. M. Lee, "Tensor-based high-order semantic relation transfer for semantic scene segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3073–3080.
[28] J. Yang, B. Price, S. Cohen, and M.-H. Yang, "Context driven scene parsing with attention to rare classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3294–3301.
[29] A. Sharma, O. Tuzel, and M.-Y. Liu, "Recursive context propagation network for semantic scene labeling," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2447–2455.
[30] P. Kontschieder, S. R. Bulo, M. Pelillo, and H. Bischof, "Structured labels in random forests for semantic labelling and object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 2104–2116, Oct. 2014.
[31] X. Peng, S. Xiao, J. Feng, W.-Y. Yau, and Z. Yi, "Deep subspace clustering with sparsity prior," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 1925–1931.
[32] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[33] A. G. Howard. (2013). "Some improvements on deep convolutional neural network based image classification." [Online]. Available: https://arxiv.org/abs/1312.5402
[34] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun. (2015). "Deep image: Scaling up image recognition." [Online]. Available: https://arxiv.org/abs/1501.02876
[35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[36] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.
[37] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[38] V. Kolmogorov and R. Zabin, "What energy functions can be minimized via graph cuts?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 2, pp. 147–159, Feb. 2004.
[39] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.
[40] 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.
[41] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and Web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.

Qi Wang (M'15–SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent system 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, with the Unmanned System Research Institute, and with the Center for OPTical IMagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.

Junyu Gao received the B.E. degree in computer science and technology from the Northwestern Polytechnical University, Xi'an, China, in 2015, where he is currently pursuing the master's degree from the Center for Optical Imagery Analysis and Learning. His research interests include computer vision and pattern recognition.



Yuan Yuan (M'05–SM'09) is currently a Full Professor with the Center for OPTical IMagery Analysis and Learning, School of Computer Science, 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 conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content.