Contents lists available at ScienceDirect

Pattern Recognition

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

Scene parsing using inference Embedded Deep Networks

Shuhui Bu, Pengcheng Han, Zhenbao Liu*, Junwei Han*

Northwestern Polytechnical University, Xi'an 710072, China

ARTICLE INFO

Article history:
Received 5 August 2015
Received in revised form
10 December 2015
Accepted 22 January 2016

Keywords:
Convolutional Neural Networks (CNNs)
Conditional Random Fields (CRFs)
Inference Embedded Deep Networks (IEDNs)
Hybrid Features

ABSTRACT

Effective features and graphical model are two key points for realizing high performance scene parsing. Recently, Convolutional Neural Networks (CNNs) have shown great ability of learning features and attained remarkable performance. However, most researches use CNNs and graphical model separately, and do not exploit full advantages of both methods. In order to achieve better performance, this work aims to design a novel neural network architecture called Inference Embedded Deep Networks (IEDNs), which incorporates a novel designed inference layer based on graphical model. Through the IEDNs, the network can learn hybrid features, the advantages of which are that they not only provide a powerful representation capturing hierarchical information, but also encapsulate spatial relationship information among adjacent objects. We apply the proposed networks to scene labeling, and several experiments are conducted on SIFT Flow and PASCAL VOC Dataset. The results demonstrate that the proposed IEDNs can achieve better performance.

1. Introduction

As the intelligent time is coming, computer vision as an important technical field of artificial intelligence has achieved rapid development in recent years. Scene parsing, a complex high level vision task not only detecting and segmenting the different objects but also recognizing what classes the objects belong to, is primarily important for a wide scope of applications. The core technique for realizing the scene parsing is to label every pixel in images with accuracy as high as possible [1,2].

Compared to scene parsing, image classification task [3] recently has made a significant breakthrough. The image classification usually assumes that the object of interest is centered and at a fixed scale, as a consequence the object localization problem is not vital. However, real scenes are complex and volatile, rarely just containing a single object or object class, hence scene parsing belongs to a multi-label classification task caring the object location and recognizing every isolated object in a scene, which leads to some challenging problems. The latest researches show that there are two keys affecting the performance of scene labeling: one is how to extract good representations of images [4–6], and the other is how to infer and improve the object class based on their spatial relationship between different objects in images [7,8].

In the last decade, lots of researches have been conducted to find good features that represent the intrinsic properties of objects and many effective features are proposed, such as Gist [9], HoG [10], SIFT [11], SURF [12], and so on. Although these features have achieved great performance in some vision applications, they just depict one part information of object in images, which causes inconsistent performance in different tasks, thereby they have limited performance and application range. Moreover, these features are extracted with systems relying on carefully engineered design, which increases the difficulty for further improvement. Some researches [13,14] are explored through simulating human vision mechanism, and some achievements have been made. In order to overcome the shortcomings of these features, deep learning [15–22] based feature learning methods have been investigated and become the mainstream of vision researches. Over the several past years, Convolutional Neural Networks (CNNs) [23,24], one type of deep learning methods, have pushed the performance of computer vision systems to soar heights on a broad array of high-level techniques, including image classification [23,24], object detection [24,25], fine-grained categorization [26], and so on. The success is partially attributed to three aspects: First, CNNs, a powerful machine learning model, automatically learn feature from image databases and are independent of the target dataset, which removes the demand to compute over lots of hand-crafted features and the need for feature selection. Second, they are deep architectures having the capacity to learn more complex non-linear model than the traditionally shallow ones [4], such as Support Vector Machine (SVM) [27] or Neural Networks (NN) [28]. Third, CNNs are inherent invariance to local

* Corresponding authors.
E-mail addresses: bushuhui@nwpu.edu.cn (S. Bu), hanpc@mail.nwpu.edu.cn (P. Han), liuzhenbao@nwpu.edu.cn (Z. Liu), jhan@nwpu.edu.cn (J. Han).

http://dx.doi.org/10.1016/j.patcog.2016.01.027
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

image transformations, which underpins the ability to learn robust abstractions of data [29].

Despite the ability of learning powerful object representations, when CNNs are applied to scene labeling task there are still existing two obstacles including signal downsampling problem and relatively weak spatial description. The first problem comes from the repeated combination of max-pooling and downsampling operation carried out at each layer of standard CNNs, which reduces the chance of obtaining a fine segmentation output and then leads to non-sharp boundaries in semantic segmentation tasks. The second problem is related to the fact that obtaining high quality parsing requires strong invariance to global and local spatial transformations, which inherently limits the spatially representative accuracy of the CNNs model [26].

To remedy the issues from existing deep learning techniques, probabilistic graphical models have been adopted as effective methods to boost the accuracy of scene labeling tasks. Conditional Random Fields (CRFs) [30], one type of the most successful graphical models, have been widely applied to address the parsing problem due to its great performance on inferring the spatial dependencies between objects. It linearly combines class scores computed through multi-label classifiers with the low level features extracted from pixels, edges, or superpixels. The key idea of CRF inference for scene labeling is to formulate the label assignment problem as a probabilistic inference problem. CRF can be used to refine weak and coarse label predictions to generate sharp boundaries and fine-gained segmentations. Therefore, CRF can be applied to overcome the drawbacks of CNNs in scene labeling tasks. Farabet et al. [2] propose a method to use CRF to refine the classification results only using CNNs, and better results are achieved. Kae et al. [31] propose a method combining CRF and Restricted Boltzman Machine (RBM) to better accomplish face labeling.

In this paper we propose a novel neural network called Inference Embedded Deep Networks (IEDNs) for scene labeling which comprise both advantages of CNNs and CRFs. Different from the way of some previous work combining deep learning methods with CRFs [2,31], which regards the deep learning methods as feature generative model to extract high-level hierarchical features and treats the CRFs as a label discriminative model to give the final category of every pixel, we design a novel structure of networks considering CRFs model as one type of layer for IEDNs, which makes the networks have explicitly structural learning ability. In brief, IEDNs consist of following three major type layers:

(1) *Feature learning layer*: CNNs are adopted to generate a feature vector containing hierarchical information for every pixel in images. The networks can learn different scales of pyramid version representations of the input image, which means the representations can capture abundant shape and contextually hierarchical information with the properly trained networks. We call these features deep hierarchical features (DHF).
(2) *Structural learning layer*: In order to improve the structural representative performance of deep learning features, we formulate the CRF as one layer of the IEDNs to explicitly learn the spatial relationship of objects in images. The graphical model is trained with DHF, and gives optimal label of each pixel or superpixel according to the trained parameters. Then we encode the region information with spatial distance relationships to generate spatially inferred features (SIF).
(3) *Feature fusion layer*: Above-mentioned two types of features have their own advantages. To explore non-linear information and to generate stronger representative features, we use Deep Belief Networks (DBNs) [4,32] to fuse DHF and SIF.

Because the structural learning is integrated into the deep learning, consequently compared to previous works, there are three main contributions as follows:

- *Explicitly structural learning*: The CNN uses the pooling and downsampling to make upper layer covering larger region. Through this way the CNN implicitly learns structural relationship in small regions. However, a massive training samples including various combination of different objects location relationship are needed to fully train the networks. In this work, CRF is regarded as a layer of the network, therefore, the structural learning can be conducted explicitly, and consequently improve the performance of inference.
- *Spatially inferred features*: A novel feature encoding spatial relationship between objects in images is proposed. By this way, not only the feature of object itself is used to estimate its class, also its neighbor information is encoded. Thereby the spatial inference performance can be better.
- *Fusing multimodal features*: To further improve the performance, feature fusing is adopted to learn their intrinsic non-linear relationships. Through fusing multimodal features, comprehensive local and global information can be encapsulated to increase the accuracy of scene labeling.

In the proposed method, CRF is trained separately and back propagation optimization is not performed, therefore, the structural learning is not a strict layer of neural networks. However the structural learning in the proposed method can be regards as a processing layer of deep neural networks. The CRF model can explicitly learn the spatial relationship of objects in image, therefore, it can overcome drawback that traditional deep neural networks lack long-range spatial inference. Consequently, the scene parsing performance can be boosted. Furthermore, one benefit of training train these layers individually is that these three separated modules can be trained repeatedly at the same time, which makes the deep networks trained more easily and quickly, so that this way reduces the time consumption.

To evaluate the proposed IEDNs, several experiments are conducted on SIFT Flow and PASCAL VOC datasets. Comprehensive comparisons and analyses indicate that the IEDNs have better performance on scene parsing.

2. Related work

In recent years the scene parsing has been explored by a lot of researches, many of which depend on Markov random fields (MRFs) [7], CRFs [30], or other graphical models [33] to ensure the consistency of the labeling. Jamie et al. [1] propose a mixture model effectively incorporating context, layout, and texture information to extract features, and then use a conditional distribution over the class labeling given a image. The use of a CRF allows to fuse context, layout, texture, color, edge, and location cues in a single unified model, which is a benefit for automatic visual understanding and semantic understanding of photographs. Motivated by the observation that each quantization has its fair share of pros and cons and the existence of a common optimal quantization level suitable for all object categories is highly unlikely, Russell et al. [34] propose a hierarchical random field model allowing integration of features computed at different levels of quantization hierarchy, which can infer better segmentation and recognition results. Gould et al. [35] propose a region-based model that combines appearance and scene geometry to automatically decompose a scene into semantically meaningful regions, and present an effective inference algorithm to optimize the energy function of graphical model. Tighe et al. [36] present a simple and effective nonparametric approach to the problem of image parsing, the method works by superpixel-level matching with local features and efficient MRF optimization for incorporating neighborhood context information. This method does not require training, and furthermore it can easily scale to datasets with thousands of images and hundreds of labels, therefore it is fast used to realize vision task. In order to

Please cite this article as: S. Bu, et al., Scene parsing using inference Embedded Deep Networks, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.027

overcome limitations that traditionally methods use homogeneous superpixel based modeling, Wang et al. [37] propose a unified framework which aggregates multiple saliency cues in a global context. In addition, an energy minimization function is formulated to model neighborhood consistence constraint and appearance consistency jointly. Ordinary CRFs used in practice typically have edges only between adjacent image pixels, to represent object relationship statistics beyond adjacent pixels, a fully connected CRFs [38] that has an edge for each pair of pixels, which augments the pixel-wise CRFs to capture object spatial relationships and improves the performance of the graphical model.

Even though graphical models have achieved an ascendant performance on scene understanding, they still strictly rely on the representative ability of features. Deep learning based feature extraction methods recently boost the development of scene labeling task. Farabet et al. [2] propose an approach that uses the multi-scale convolutional network trained with raw pixels to learn dense feature vectors that encode regions of multiple sizes centered on each pixel. The method learns feature without engineered design, and generates a powerful representation capturing texture, shape, and contextual information. Then a classical CRF model is constructed on the superpixels as the strategy to complete the scene parsing task. Girshick et al. [39] propose region based CNNs, which first generate category-independent region proposals to get the set of candidate regions, then uses a large convolutional neural network to extract a fixed-length feature vectors from every region, finally adopts linear SVMs model to tag the labels for each feature vector. This method applies high-capacity CNNs to bottom-up region proposals, therefore it can localize and segment objects accurately. Socher et al. [40] notice that recursive structure is commonly found in the different modal inputs like natural language sentences and natural scene images, and discover this structure can help us to identify not only the units that an image or sentence contains but also how they interact to form a whole. Motivated with this observation, they introduce a max-margin structure prediction architecture based on recursive neural network (RNN), which can successfully recover such structure both in complex scene images as well as sentences. It is well known that directly applying the forward and backward propagation to pixelwise classification in patch-by-patch scanning manner is extremely inefficient due to the fact that surrounding patches of pixels have large overlaps. Li et al. [41] present highly efficient algorithms for performing forward and backward propagation of CNNs for pixelwise classification on images. The algorithms cut all the redundant computation in convolution and pooling operation through introducing novel d-regularly sparse kernels.

Recently, CRFs and deep learning methods have been fused to complete scene parsing. Zheng et al. [42] introduce a new form of Convolutional Neural Network that combines the merits of CNN and CRF. They formulate CRF as recurrent neural networks and call the networks CRF-RNN. Then CRF-RNN is plugged in as a part of CNN to obtain a deep network that has desirable properties of both CNN and CRF. In this way, their system fully integrates CRF modeling with CNN, making it possible to train the whole deep network end-to-end with the usual back-propagation algorithm, avoiding offline postprocessing methods for object delineation. Through combining the strengths of deep CNN to learn powerful feature representations, with CRFs which can capture contextual relation modeling, Lin et al. [43] show great performance on semantic segmentation task. Unlike previous works, their formulation of “deep CRFs” learns both unary and pairwise terms using multi-scale fully Convolutional Neural Networks (FCNNs) in an end-to-end fashion, which enables to model complex spatial relations between image regions. Moreover, they propose a novel method for efficient joint training of the deep structured model based on piecewise training, which avoids repeated inference, and so is computationally tractable. Cogswell et al. [44] present a two-module approach to semantic segmentation that incorporates CNN and Graphical Models. Graphical models are used to generate a small (5–30) set of diverse segmentations proposals, such that this set has high recall. A Novel CNN called SegNet is used to generate complex features. Importantly, SegNet is specifically trained to optimize the corpus-level PASCAL IOU loss function. This two-module approach achieves good performance on the segmentation challenge. Liu et al. [45] combine the CNN and CRF to complete the image segmentation task. They propose to exploit pre-trained large CNN to generate deep features for CRF learning. Then the CRF parameters are learned using a structured support vector machine (SSVM). To fully exploit context information in inference, they construct spatially related co-occurrence pairwise potentials and incorporate them into the energy function. Chen et al. [26] bring together methods of deep Convolutional Neural Networks (DCNNs) and probabilistic graphical models for addressing the task of pixel-level classification. They overcome the poor localization property of deep networks by combining the responses at the final layer of DCNN with fully connected CRF. Their “DeepLab” system is able to localize segment boundaries at level of accuracy which is beyond previous many methods. In addition to scene labeling task, convolutional network and graphical model based methods are also applied in human pose estimation. Tompson et al. [46] propose a hybrid architecture that consists of DCNNs and MRF. They use this hybrid architecture to exploit structural domain constraints such as geometric relationships between body joint locations, and improve the performance through jointly training these two model paradigms.

Different from above mentioned methods, we treat the graphical model as one type of layer in the deep neural network to improve the explicitly inference capability, and use the inference results to generate SIF which includes middle/large-range structural relationships. Thereby, both advantages of deep learning based feature extractor and graphical model can be retained at the same time. Moreover, to fully grasp the intrinsic of complex object, DBNs are adopted to fuse DHF and SIF, and non-linear relationships of them can be discovered.

3. Inference embedded deep networks

The proposed IEDNs consist of three units: feature learning layer, structural learning layer, and feature fusion layer. The feature learning layer mainly comprises the convolutional layer, pooling layer, and rectifier linear unit. The structural learning layer is used to learn spatial relationship of superpixels. The feature fusion layer uses DBNs to learn object features and spatial distribution features together. The flowchart of the proposed method is depicted in Fig. 1.

3.1. Feature learning layer

In computer vision task, strong representations are vital for good performance. Recent researches show that good intrinsic representations are hierarchical. The structure is like that pixels are assembled into edglets, edglets into motifs, motifs into parts, parts into objects, and objects into complex scenes, which means the step of extracting features is layer by layer. Deep leaning based feature extractors provide a powerful framework to learn such hierarchical features. Moreover, convolutional operation is close to mechanism of human eyes capturing features, which makes the networks highly invariant to translation, scaling, tilt, or other deformations. Here, we use CNNs to learn the DHF, and the operational principle is illustrated in Fig. 2.

3.1.1. Convolutional neural networks

Convolutional neural networks are trained with multiple stages. The input and output of each stage are sets of arrays called feature maps. In this research, color images are used as the input of neural networks, thus each feature map can be regarded as a two dimensional array containing color channels of the input images. After one stage, the output feature map is treated as further abstraction of input feature

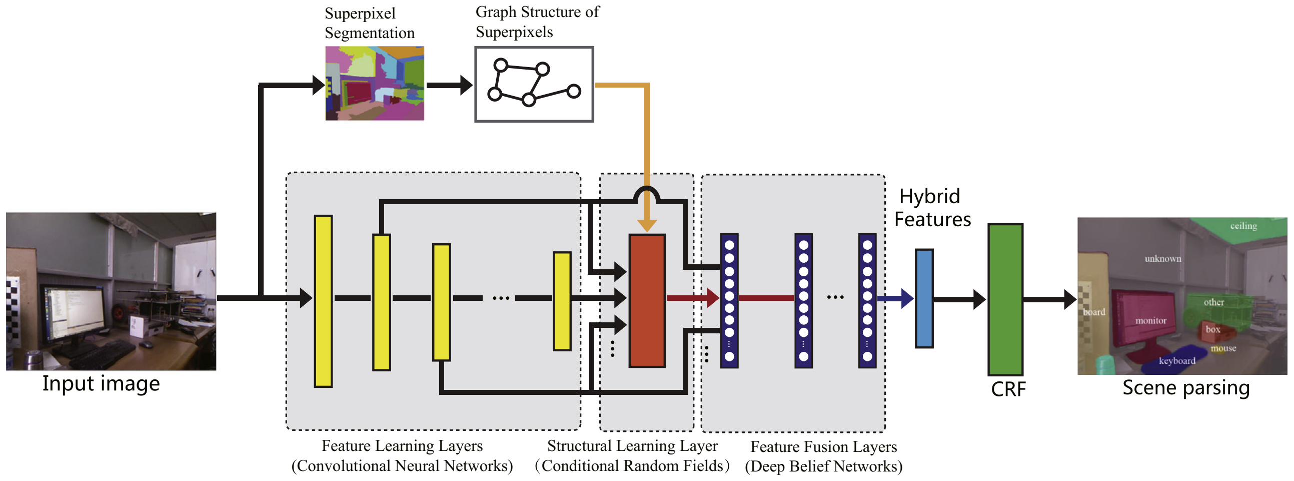
Superpixel Segmentation
Graph Structure of Superpixels
Hybrid Features
Input image
CRF
Scene parsing
ceiling
unknown
board
monitor
other
box
mouse
keyboard
Feature Learning Layers (Convolutional Neural Networks)
Structural Learning Layer (Conditional Random Fields)
Feature Fusion Layers (Deep Belief Networks)

Fig. 1. The flowchart of the proposed method. The IEDNs consist of three major types of layers: feature learning layer (Convolutional Neural Networks), structural learning layer (Conditional Random Fields), and feature fusion layer (Deep Belief Networks).

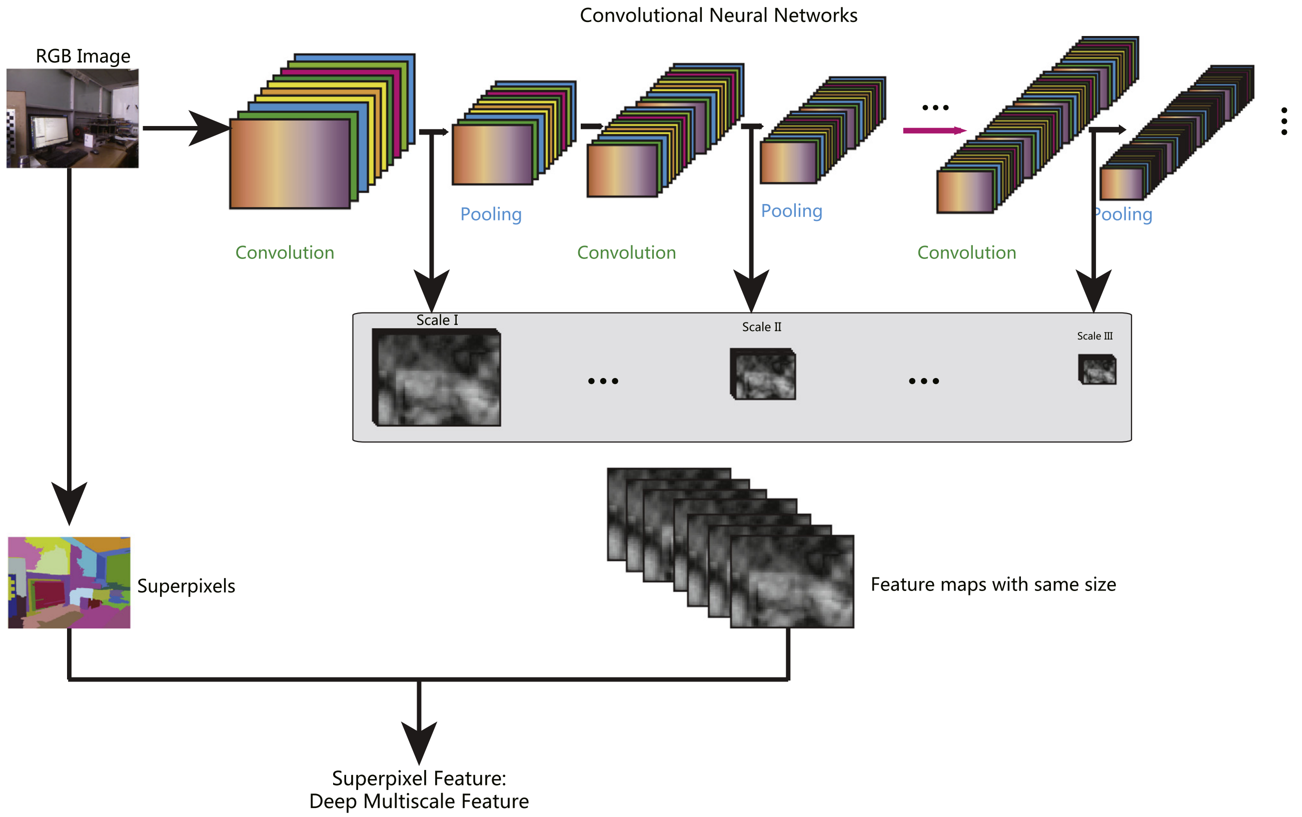

Fig. 2. The principle of feature learning.

map. Each stage is composed of three parts: convolutional operation, non-linearity transformation, and feature pooling. A typical CNN contains several such 3-part stages, and finally use softmax to perform the classification. The details of CNN are introduced as follows.

The CNNs with L layers can be described as a sequence of linear transformations ($*$ operator), interspersed with non-linear symmetric squashing units like sigmoid function or tan h function (non-linear operator), and pooling/subsampling operations (*pool* operator). The input image I of networks are seen as three dimensional arrays, consisting of the number of feature maps, the height of the maps, and the width. The output of the l-th stage is denoted with F_l. For each layer l, we have:

$$F_l = pool\,(\tanh(W_l * F_{l-1} + b_l)), \tag{1}$$

for $l \in 1, \ldots, L$, b_l is the bias parameter of the l-th layer, W_l is the convolutional kernel. The initial feature map is the input image $F_0 = I$. Therefore, each stage is stacked into the whole networks layer upon layer.

In our model, the convolutional kernels W_l and the biases b_l are the trainable parameters. The *tanh* function is the point-wise hyperbolic tangent function. The *pool* operation is a function considering a neighborhood of activations and generating one activation in every neighborhood. Max-pooling operator is regarded as the *pool* function, which gets the maximum activation in the neighborhood and brings the built-in invariance to translations.

Once output feature maps of all layers have been generated, we upsample them into the same size of input image and then concatenate them to produce a three dimensional arrays $F \in \mathbb{R}^{N \times H \times W}$,

Please cite this article as: S. Bu, et al., Scene parsing using inference Embedded Deep Networks, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.027

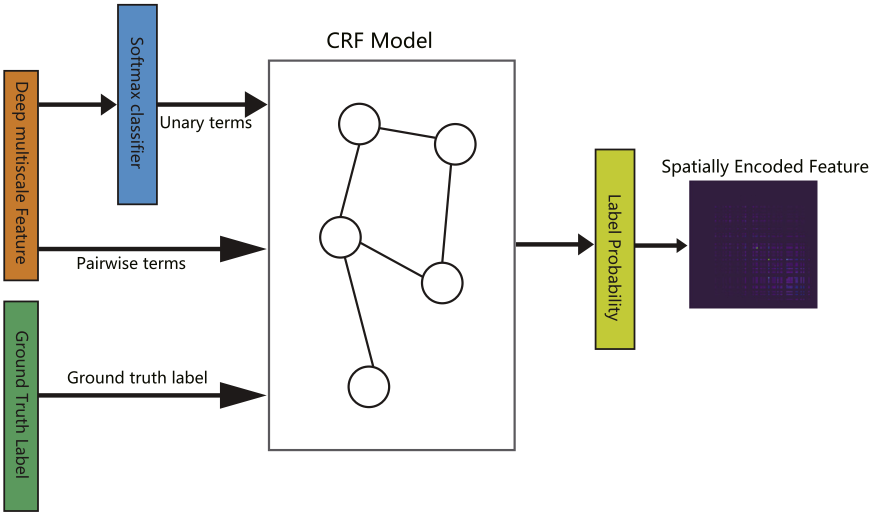

Fig. 3. The illustration of structural learning layer.

in which the three dimensions are the height of images H, the width of images W, and the number of feature maps N. The arrays F can be seen as hierarchical descriptors:

$$F = [up(F_1), up(F_2), \ldots, up(F_L)], \tag{2}$$

where up operator is an upsampling function, $up(F_l) \in \mathbb{R}^{N_l \times H \times W}$, N_l means the number of feature maps or the number of filter kernel in l-th layer. For a pixel in a image, its final descriptor can be denoted as $p \in \mathbb{R}^N$. In principle, making full use of the outputs of all layers can generate stronger features. However, in fact the information in some feature maps is redundant, which will reduce the computational effi-ciency. Therefore, in practice we just select the outputs of several layers to produce the descriptors F to improve the efficiency and ensure the quality of the features at the same time. The details of parameters will be explained in experiment section.

3.1.2. Superpixels

Predicting the class of each pixel independently from the neighbors will lead to error predictions resulted from various noise. A simple cleanup can be obtained through forcing local regions of the same color intensities to be assigned a single label. Therefore, we use simple linear iterative clustering (SLIC) [47] to generate superpixels from input images. There are three main advantages of using superpixels as the elementary units:

- To improve the ability of anti-noise.
- The number of pixel is far more than that of superpixel in a image, therefore the overall processing can be boosted.
- Since superpixels tend to preserve boundaries, a very accurate segmentation can be obtained with simply finding the superpixels in which part of the object, which boosts the performance in scene parsing task.

A complex scene image contains various type of boundaries, and SLIC can preserve the boundaries of adjacent objects well, which is more reasonable for scene parsing. In addition, it has high efficiency to generate superpixels. For each superpixel composed of lots of pixels, we compute the average feature value with features of pixels in the segmented region, where the feature is denoted as $S_p \in \mathbb{R}^N$.

3.2. Structural learning layer

Even that CNNs can learn great descriptors containing hierarchical information, it might still be prone to misclassification without neighborhood information. Moreover, the networks cannot explicitly learn features with middle or large range structures, what is to say, the hierarchical features learned with CNNs lack strong spatial relationship representations between objects. To remedy the drawbacks of CNNs, we introduce superpixels based CRF model to explicitly learn the spatial distribution of objects and generate the SIF. The illustration of the structural learning is shown in Fig. 3.

3.2.1. Conditional random fields

For the purpose above mentioned, we define a graph $G = (V, E)$, where vertices $v \in V$ and edges $e \in E \in \mathbb{R}^{V \times V}$. Each superpixel in the image is regarded as a vertex, and edges are the relationships between every neighboring nodes. An edge consisting of two vertices v_i and v_j is denoted with e_{ij}. The energy function of CRF is composed of a unary term enforcing the superpixels to take values close to the predicted label and a pairwise term enforcing regularity or local consistency. The CRF energy to minimize is given with:

$$E(l) = \sum_{i \in V} \psi(c_i, l_i) + w \sum_{e_{ij} \in E} \phi(l_i, l_j). \tag{3}$$

We define the unary term as:

$$\psi(c_i, l_i) = \exp(-\alpha_u c_i), \tag{4}$$

and the pairwise term is considered as:

$$\phi(l_i, l_j) = \begin{cases} 1 - exp(-\alpha_p \|S_{p_i} - S_{p_j}\|_2^2 / \sigma_\phi^2) & : \quad l_i = l_j \\ exp(-\alpha_p \|S_{p_i} - S_{p_j}\|_2^2 / \sigma_\phi^2) & : \quad l_i \neq l_j, \end{cases} \tag{5}$$

where c_i is a vector represents the probabilities of superpixel v_i belonging to all classes, which is generated from a trained softmax classifier. l is the resulting label. $\|S_{p_i} - S_{p_j}\|_2^2$ is the norm of superpixel features between v_i and v_j. The w is the parameter representing the tradeoff between spatial regularization and the confidence in the classification. This CRF energy is minimized using graph-cut algorithm [48,49]. Once the CRF graphical model has been learned properly, we infer the probability of each superpixel, which is denoted by $\theta_i \in \mathbb{R}^n$, where n is the number of category.

3.2.2. Spatially inferred feature

In most previous work, CRF is generally regarded as the last step of various vision tasks for refining the classified labels. In the proposed framework, the work is not done yet. We have got the inference label probability of each superpixel, but the results are generated with features learned with the CNNs, which lacks strong ability of learning spatial relationship. Although graphical model can remedy this shortcoming, we hope to further improve the performance. Thus, we propose SIF to represent not only the feature of superpixel by itself, but also spatial relationships. This way makes the whole networks have explicitly structural learning capability. For a superpixel μ and its local region generated connection graph $G_\mu = (V_\mu, E_\mu)$, the SIF is given by:

$$\odot(\mu) = \lambda \sum_{i \in V_\mu} \sum_{j \in V_\mu} \theta_i \theta_j^T \exp\left(-k_d \frac{d(v_i, v_j)}{\sigma_d}\right) \tag{6}$$

where λ is the normalized parameter, $d(v_i, v_j)$ is the distance of superpixel i and j, k_d is distance decay rate, and σ_d is the maximal distance of any vertices in the graph G_μ. The resulting representation $\odot$ is a $n \times n$ matrix, representing the frequency of appearance of nearby probability of vertices i and j. We call $\odot$the SIF.

3.3. Feature fusion layer

After the procedures of feature learning layer and structural learning layer, for each superpixel μ, it has two type of descriptors: DHF S_p and SIF $\odot$. We concatenate these two features into $[S_p, \odot] \in \mathbb{R}^{N+n\times n}$, and then use DBNs [4,32] to fuse the concatenated features and explore comprehensive non-linear relationships between every dimension of features. The flow of feature fusion is depicted in Fig. 4.

Recent works on DBNs [4,32] have shown that it is feasible to learn multiple layers of non-linear features that are useful for object classification without requiring labeled data. The features are trained layer by layer in a RBM [50,51] by means of contrastive divergence (CD) [51]. The feature activations learned by one layer RBM become the input data for training the next layer RBM. After a pre-training phase that learns layers of features which are good at modeling the statistical structure in a set of unlabeled data, supervised back-propagation can be used to fine-tune the features for classification. The last layer's output is a kind of highly representative feature which encodes the input data.

3.3.1. Restricted Boltzmann machines

In order to make the paper more self-contained, we succinctly discuss the concept of restricted Boltzmann machines. The RBM is a two layer, bipartite, undirected graphical model with a set of binary hidden unit $\mathbf{h}$, a set of (binary or real-valued) visible units $\mathbf{v}$, and symmetric connections between these two layers represented by a weighted matrix W. The joint distribution $p(\mathbf{v}, \mathbf{h}; \theta)$ over the visible units $\mathbf{v}$ and hidden units $\mathbf{h}$, given the model parameters $\theta = \{\mathbf{w}, \mathbf{a}, \mathbf{b}\}$, is defined in terms of an energy function $E(\mathbf{v}, \mathbf{h}; \theta)$ of

$$p(\mathbf{v}, \mathbf{h}; \theta) = \frac{\exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z}, \tag{7}$$

where $Z = \sum_v \sum_h exp(-E(\mathbf{v}, \mathbf{h}; \theta))$ is a normalization factor or partition function and the marginal probability that the model assigns to a visible vector $\mathbf{v}$ is

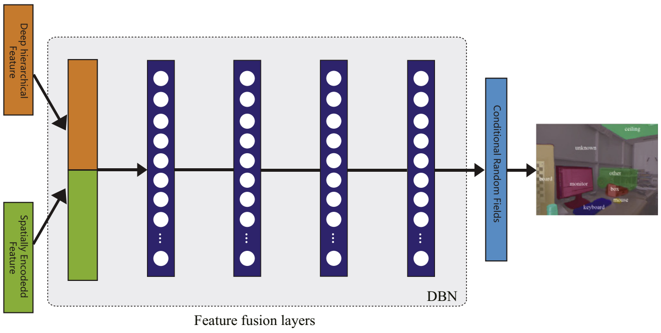
Deep hierarchical Feature
Spatially Encodedd Feature
DBN
Feature fusion layers
Conditional Random Fields
ceiling
unknown
board
monitor
other
box
mouse
keyboard

Fig. 4. The illustration of feature fusion.

$$p(\mathbf{v}; \theta) = \frac{\sum_h \exp(-E(\mathbf{v}, \mathbf{h}; \theta))}{Z}. \tag{8}$$

For a Bernoulli (visible)–Bernoulli (hidden) RBM, the energy is

$$E(\mathbf{v}, \mathbf{h}; \theta) = -\sum_{i=1}^{V} \sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j, \tag{9}$$

where w_{ij} represents the symmetric interaction between visible unit v_i and hidden unit h_j, b_i and a_j the biases, and V and H are the numbers of visible and hidden units. The conditional probabilities can be efficiently calculated as

$$p(h_j = 1 | \mathbf{v}; \theta) = \sigma\left(\sum_{i=1}^{V} w_{ij} v_i + a_j\right), \tag{10}$$

$$p(v_i = 1 | \mathbf{h}; \theta) = \sigma\left(\sum_{j=1}^{H} w_{ij} h_j + b_i\right). \tag{11}$$

where $\sigma(x) = 1/(1+\exp(-x))$ is a sigmoid activation function.

In principle, the RBM parameters can be optimized by performing stochastic gradient ascent on the log-likelihood of training data. Unfortunately, computing the exact gradient of the log-likelihood is intractable. Instead, the CD approximation [51] is typically used, which has been shown to work well in practice.

3.3.2. Deep belief networks

Stacking a number of the RBMs and learning layer by layer from bottom to top gives rise to a DBN. It has been shown that the layer-by-layer greedy learning strategy [32] is effective, and the greedy procedure achieves approximate maximum likelihood learning. In our work, the bottom layer RBM is trained with the input data of $[S_p, \odot]$, and the activation probabilities of hidden units are treated as the input data for training the upper-layer RBM, and so on. After this procedure, we can get the final feature called as “hybrid feature”. Then CRF is used to compute the label of each superpixel, which completes the scene labeling vision task.

3.4. Training procedure of deep neural networks

The proposed deep neural networks are composed of three types of layers including CNNs layer, CRF layer, and DBN layer. In this paper, we train the modules individually, which makes the deep neural networks trained more easily and quickly.

In the feature learning layer, we use MatConvNet [56] and public available pre-trained model to extract hierarchical feature. The adopted CNNs model is ‘imagenet-vgg-f’, which contains 21 layers. In the structural learning layer, in order to optimize the CRF energy function, we adopt the graph-cut algorithm [48,49] to get the optimal weights with the input CNNs features. In the DBN layer, Deep Belief Networks are decomposed into many RBMs trained by means of contrastive divergence (CD) [51]. The feature activations learned by one layer RBM become the input data for training the next layer RBM. After the pre-training phase, supervised back-propagation method is used to fine-tune the DBN weights.

4. Experiments

To evaluate the performance of the proposed IEDNs applied in scene parsing, we test and compare it with several state-of-the-art methods on two standard dataset: SIFT Flow [57] and PASCAL VOC [58]. The evaluation measures for the methods are per pixel accuracy which means the percentage of pixel correctly labeled

Please cite this article as: S. Bu, et al., Scene parsing using inference Embedded Deep Networks, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.027

and per class accuracy which denotes the average of semantic category accuracies. Due to the fact that our IEDNs are composed of layers belonging to three type of layers, for different datasets the parameters are not fixed. Related details can be found in following subsections.

4.1. *SIFT flow*

The SIFT Flow dataset, a challenging dataset composed of 2688 images with 33 semantic categories and background, are thoroughly labeled and split into 2488 training images and 200 test images. Our experimental results on this dataset are reported in Table 1. In the table, 'CNN' means the results just using CNN layers to extract hierarchical features, 'CNN + CRF' means the results using CNN and CRF, and 'IEDNs' means the results generated by the proposed method. From the results, we can see that the proposed IEDNs achieve good results both on the per-pixel and per-class accuracy in comparison with previous approaches. The intermediate results show that only using part of the information can not reach high accuracy. Fusing spatially information and hierarchical feature can boost the overall performance.

For the convolutional neural networks, we use MatConvNet [56] and public available pre-trained model to extract hierarchical feature. The adopted CNNs model is 'imagenet-vgg-f', which contains 21 layers. In order to accelerate the speed of our framework in practice, we just select the feature maps of three layers, which are 5-th, 13-th, and 16-th of the whole networks, and the feature map number of each selected layer is 64, 256, and 256, respectively. Through this way, the parsing accuracy and computation performance can be balanced. In the superpixel step, we conduct several experiments to choose a proper region size that keeps great experimental performance and high computation efficiency of whole processing at the same time. The optimal region size of each superpixel is finally set to 15. In the structural learning layer, we use a grid search method to estimate the optimal parameters, where $w=0.2$. The CRF model learns the spatial relationships between each superpixel effectively and rectifies the incorrect and unreasonable labels. Then in the process of generating SIF, the parameter of distance decay rate k_d is set to 0.1 in our experiments. In the feature fusion layer, we construct four layers for DBN including input and output layer. The number of nodes in each hidden layer is empirically set to 1000 and 800, and the node number of output layer is set to 400. The learning rate is set to be 0.1, and the momentum is 0.9, and maximum epoch is 5000.

The actual timings for different steps of the proposed IEDNs are listed in Table 2. The timings are measured on a computer with Xeon 3.2 GHz CPU and 16G memory. Calculating the feature map of CNNs and superpixel features are the most time consuming two steps. Therefore, the proposed IEDNs have competitive performance of computation.

Table 1
Comparison of parsing performance about per-pixel accuracy and per-class accuracy with different methods on SIFT Flow dataset. 'CNN' means the results just using CNN layers to extract hierarchical features, 'CNN + CRF' means the results using CNN and CRF, and 'IEDNs' means the results generated by the proposed method.

Method	Per-pixel accuracy (%)	Per-class accuracy (%)
Liu et al. [52]	74.75	–
Tighe et al. [36]	76.9	29.4
Eigen et al. [53]	77.1	32.5
Singh et al. [54]	79.2	33.8
Farabet et al. [2]	78.5	29.6
Pinheiro et al. [55]	77.7	29.8
CNN	69.5	27.2
CNN + CRF	72.8	28.7
IEDNs	80.4	35.8

Table 2
The statistic timing for each step of the proposed IEDNs on SIFT Flow dataset. The time unit is minute.

Procedure	Timing
Hierarchical feature extraction	13.80
Superpixel	4.01
Superpixel feature calculation	2.45
Structure learning	2.67
Spatially inferred feature	1.23
Feature fusion learning	8.31

To further analyze the results of our IEDNs experimented on this dataset, we visualize some good results in Fig. 5 and bad results in Fig. 6. From Fig. 5, there are two keys deserved attention: first, the proposed IEDNs can find rough area of objects except some very small ones. After all, this dataset has 34 categories including background, which is challenging for scene parsing. Second, if the outline of object is not complex, such as some simple geometries including straight line, triangle, square, and rectangle, the boundary will be closer to the ground truth, which results in more exact region of isolated objects. These successes are attributed to three fields: (1) we use CNNs to learn hierarchical features, which have very strong ability of capturing various information. The more useful information we get, the more powerful the representation ability of features will be. (2) We introduce inference layer to encode spatial features designed to rectify the label of each pixel, this feature has strong spatial constraints among objects, which boost performance of scene understanding, especially for complex scene. (3) DBNs can further exploit the non-linear relationships between various features. This step fuses two types of features to further abstract more representative features.

In addition to the good results analyzed above, we also show some bad results in Fig. 6. The most serious problem is that small size objects are difficult to be detected and recognized. We can find that person, tree, window, and branch are easily judged as other categories. In addition, if the boundaries of objects are more complex, the results will be worse. These two problems might be resulted from two points of our IEDNs. First, small size objects cannot be detected, which means that the details of objects are lost. In practice just three layers' feature maps are selected to conduct the experiments, some basic information are dropped unintentionally. What is more, in our IEDNs we use superpixel instead of pixel to complete the scene parsing task. Although this way accelerates the speed of whole network, it might also lose much detail of very small size objects.

4.2. *PASCAL voc*

We also test our IEDNs on the PASCAL VOC 2012 dataset, consisting of 20 foreground object classes and one background class. The original dataset contains 1464, 1449, and 1456 images for training, validation, and testing, respectively. To train the network completely, the dataset is augmented by the extra annotations provided by Hariharan et al. [59], resulting in 10 582 training images. The performance is measured by the mean intersection-over-union (IU) score [60]. Some parsing results are demonstrated in Fig. 7. We compare our network to the previous state-of-the-art methods. The related results can be found in Table 3. We can find that our network achieves the best results on mean IU. Through comparison experiments, we can conclude that the parsing accuracy using both CNNs and CRFs layers is better then that just using CNNs layer. Furthermore, after the feature fusion, the performance is shown boosted, which indicates that the proposed networks can learn spatial relationship better.

The parameters of our IEDNs on this VOC dataset are partly similar with those on SIFT Flow dataset. In feature learning layer, the feature maps generated from 5-th, 13-th, and 16-th of the whole networks are used to be as a feature representation of input raw images. In fact, we

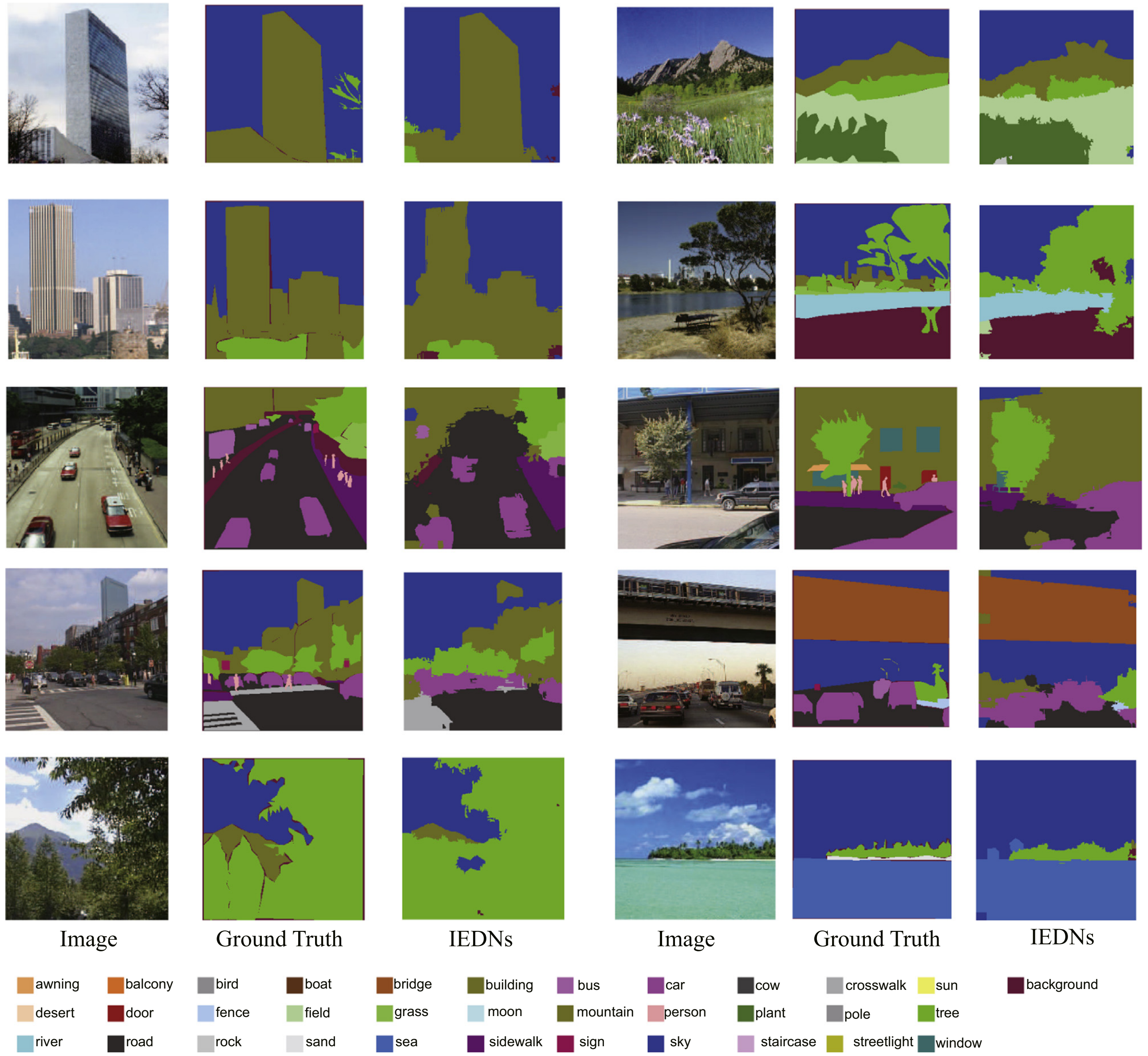

Fig. 5. Some scene parsing results from the SIFT Flow dataset.

select these three layers through many experiments to ensure high quality of scene parsing and fast computational efficiency. Moreover, CNNs are hierarchical feature extractors, which implies that different layers capture various scale information. In the superpixel step, we also set the region size of every superpixel as 15. For structural learning layer, the parameter w of CRF model is learned to be 0.11. The decay rate k_d is also set to 0.1 on this dataset. In the feature fusion layer, four layers are constructed for DBN. The number of nodes in each hidden layer is set to 800 and 600, and the node number of output layer is set to 400. The learning rate is 0.1, the momentum is 0.5, and maximum epoch is 5000.

From the comparison results, we can find that the proposed method achieves good result on VOC 2012 dataset. The reason might be found from the following aspects: (1) in the proposed method, the inference layer is utilized to encode spatial features designed to rectify the label of each pixel, and this feature has strong spatial constraints among objects, which boost performance of scene understanding, especially for complex scene; (2) DBNs can further exploit the non-linear relationships between various features, this operation can further abstract more representative features.

5. Conclusion

In this paper, we propose a novel deep learning model for structural learning. The proposed inference embedded deep networks mainly contain three types of layers: feature learning layer, structural learning layer, and feature fusion layer. Different from other methods, in the proposed model the structural learning is integrated into the network, which performs the structure inference to provide more robust spatial information. In addition, we also introduce unsupervised deep learning into the network to explore the non-linear relationships between image features and spatial information at the same time. Therefore, it can achieve better structural learning performance.

The experiments demonstrate that the proposed IEDNs achieve the state-of-the-art performance on standard scene understanding datasets, therefore it can be validated that the novel deep learning architecture is promising to perform scene parsing. The whole network does not depend on hand-crafted features, furthermore the statistic information and spatially information are integrated high efficiently. In fact, when we apply our IEDNs to scene parsing, just several layers of CNNs are used to generate feature of

Please cite this article as: S. Bu, et al., Scene parsing using inference Embedded Deep Networks, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.01.027

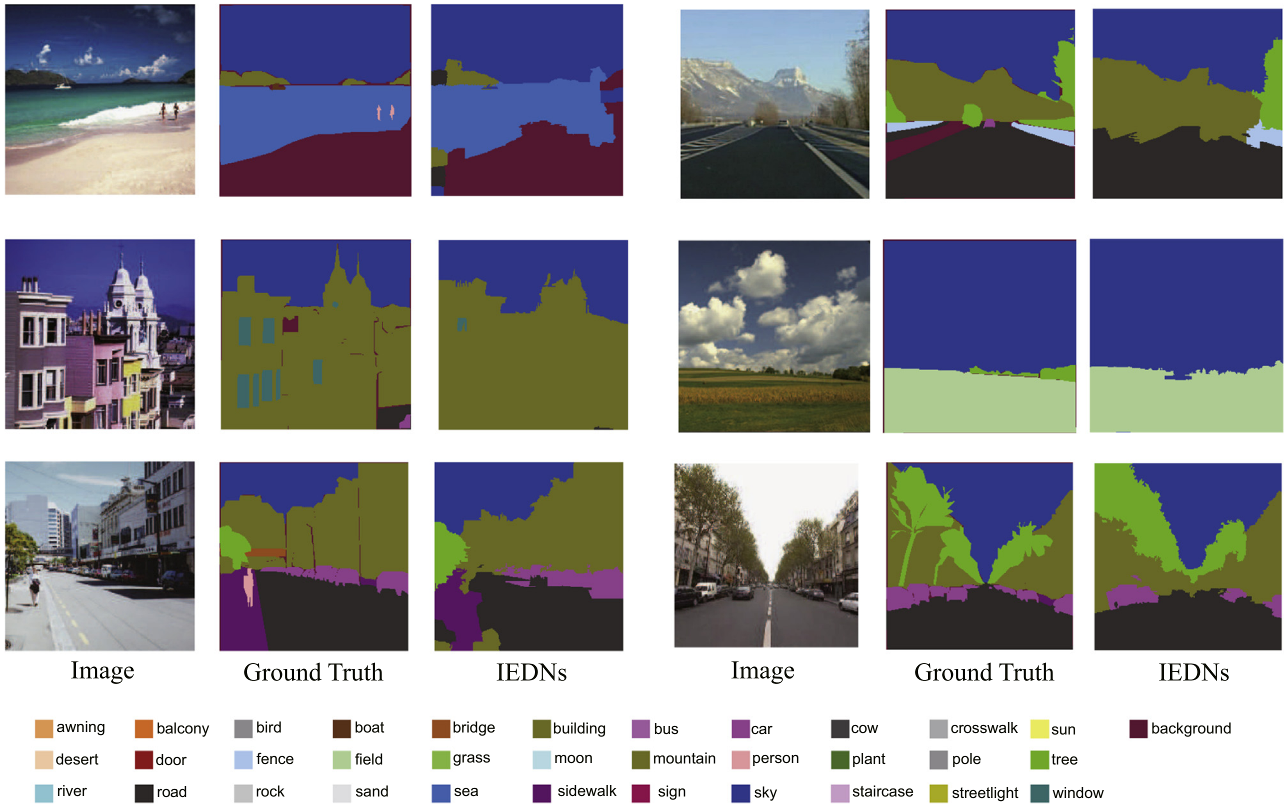

Fig. 6. Some bad scene parsing results from the SIFT Flow dataset.

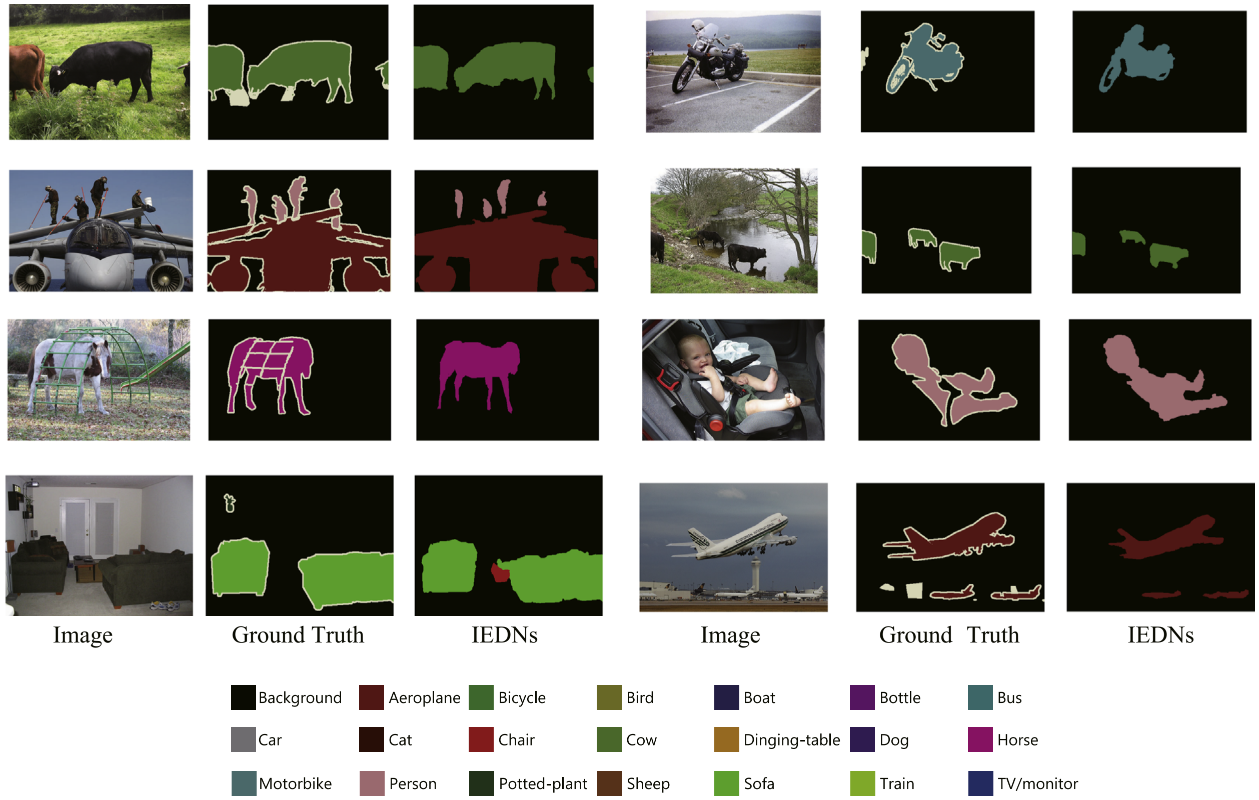

Fig. 7. Some scene parsing results from the VOC 2012 dataset.

superpixel. If we use all layers' feature maps, which will confidently improve the performance. In addition, the graphical model is treated as a layer of network, therefore other processing layers can be easily integrated into the IEDNs. This makes the proposed framework have better flexibility and further improves the performance.

Table 3
Comparison of parsing performance about Mean IU Accuracy with different methods on PASCAL VOC 2012 datset. 'CNN' means the results just using CNN layers to extract hierarchical features, 'CNN + CRF' means the results using CNN and CRF, and 'IEDNs' means the results generated by the proposed method.

Method	Mean IU accuracy (%)
Mostajabi et al. [61]	64.4
Lin et al. [43]	70.7
Papandreou et al. [62]	72.7
Zheng et al. [42]	74.7
CNN	67.5
CNN + CRF	69.1
IEDNs	75.4

Referring to the proposed IEDNs, we divide the network into three separated modules and train them individually. Although this way makes the deep networks trained more efficiently, it does not exploit full advantages of that CRF layer is regarded as one type of layer for deep neural networks. In the future work, we intend to research a novel model to integrate different layers more effectively, which is appropriate for adopting back propagation to jointly optimize all parameters. This way will fully extract benefits from these three types of layers, and further enhance the performance about scene parsing.

Although the proposed method achieves better performance, it inherits the drawback of CRF that small object might not be correctly estimated. As a consequence, high quality scene parsing of very complex scene is difficult to be achieved. To obtain better scene parsing, more elegant structural learning models for deep networks need to be investigated. In addition, currently the proposed IEDNs only adopt image information to do the scene parsing, but actually human brain performs scene parsing with 3D information. In the following research, to obtain better performance and robustness, 3D feature extraction and multimodal feature fusion need to be explored.

Conflict of interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments

This work is partly supported by grants from National Natural Science Foundation of China (61202185, 61473231, 61573284), the Fundamental Research Funds for the Central Universities (310201401-(JCQ01009,JCQ01012)), Open Projects Program of National Laboratory of Pattern Recognition (NLPR).

References

[1] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context, Int. J. Comput. Vis. 81 (1) (2009) 2–23.

[2] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.

[3] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE, Miami, FL, USA, 2009, pp. 1794–1801.

[4] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.

[5] J. Han, S. He, X. Qian, D. Wang, L. Guo, T. Liu, An object-oriented visual saliency detection framework based on sparse coding representations, IEEE Trans. Circuits Syst. Video Technol. 23 (12) (2013) 2009–2021.

[6] S. Bu, P. Han, Z. Liu, K. Li, J. Han, Shift-invariant ring feature for 3d shape, Vis. Comput. 30 (6–8) (2014) 867–876.

[7] S. Nowozin, C.H. Lampert, Structured learning and prediction in computer vision, Found. Trends® Comput. Graph. Vis. 6 (3–4) (2011) 185–365.

[8] L. Lin, X. Wang, W. Yang, J.-H. Lai, Discriminatively trained and-or graph models for object shape detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (5) (2015) 959–972.

[9] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.

[10] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005, vol. 1, IEEE, San Diego, California, USA, 2005, pp. 886–893.

[11] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[12] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Comput. Vis. Image Understand. 110 (3) (2008) 346–359.

[13] J. Han, C. Chen, L. Shao, X. Hu, T. Liu, Learning computational models of video memorability from fmri brain imaging, IEEE Trans. Cybern. 45 (8) (2015) 1692–1703.

[14] J. Han, D. Zhang, S. Wen, L. Gao, T. Liu, X. Li, Two-stage learning to predict human eye fixations via sdaes, IEEE Trans. Cybern. 46 (2) (2016) 487–498. http://dx.doi.org/10.1109/TCYB.2015.2404432.

[15] Q.V. Le, Building high-level features using large scale unsupervised learning, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Vancouver, Canada, 2013, pp. 8595–8598.

[16] Q.V. Le, W.Y. Zou, S.Y. Yeung, A.Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Colorado Springs, USA, 2011, pp. 3361–3368.

[17] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, Montreal, Canada, 2009, pp. 609–616.

[18] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: a deep convolutional activation feature for generic visual recognition, arXiv preprint arXiv:1310.1531.

[19] S. Bu, Z. Liu, J. Han, J. Wu, R. Ji, Learning high-level feature by deep belief networks for 3-d model retrieval and recognition, IEEE Trans. Multimed. 16 (8) (2014) 2154–2167.

[20] S. Bu, P. Han, Z. Liu, J. Han, H. Lin, Local deep feature learning framework for 3d shape, Comput. Graph. 46 (2015) 117–129.

[21] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, F. Wu, Background prior-based salient object detection via deep reconstruction residual, IEEE Trans. Circuits Syst. Video Technol. 25 (8) (2015) 1309–1321.

[22] J. Han, D. Zhang, G. Cheng, L. Guo, J. Ren, Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning, IEEE Trans. Geosci. Remote Sens. 53 (6) (2015) 3325–3337.

[23] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[24] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. LeCun, Overfeat: Integrated recognition, localization and detection using convolutional networks, in: International Conference on Learning Representations (ICLR 2014), CBLS, 2014.

[25] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognit. 48 (10) (2015) 2993–3003.

[26] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, in: International Conference on Learning Representations, 2015.

[27] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[28] S. Haykin, N. Network, A comprehensive foundation, Neural Netw. 2 (2004).

[29] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014, Springer, Zurich, Switzerland, 2014, pp. 818–833.

[30] J. Lafferty, A. McCallum, F.C. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, 2001.

[31] A. Kae, K. Sohn, H. Lee, E. Learned-Miller, Augmenting crfs with Boltzmann machine shape priors for image labeling, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2013, pp. 2019–2026.

[32] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[33] P. Luo, L. Lin, X. Liu, Learning compositional shape models of multiple distance metrics by information projection, IEEE Trans. Neural Netw. Learn. Syst. http://dx.doi.org/10.1109/TNNLS.2015.2440430.

[34] C. Russell, P. Kohli, P.H. Torr, et al., Associative hierarchical crfs for object class image segmentation, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, Miami, FL, USA, 2009, pp. 739–746.

[35] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: 2009 IEEE 12th International Conference on Computer Vision, IEEE, Kyoto, Japan, 2009, pp. 1–8.

[36] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with superpixels, in: Computer Vision–ECCV 2010, Springer, Crete, Greece, 2010, pp. 352–365.

[37] K. Wang, L. Lin, J. Lu, C. Li, K. Shi, Pisa: pixelwise image saliency by aggregating complementary appearance contrast measures with edge-preserving coherence, IEEE Trans. Image Process. 24 (10) (2015) 3019–3033.

[38] P. Krähenbühl, V. Koltun, Efficient inference in fully connected crfs with gaussian edge potentials, in: Advances in Neural Information Processing Systems, 2011, pp. 109–117.

[39] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, Ohio, 2014, pp. 580–587.

[40] R. Socher, C.C. Lin, C. Manning, A.Y. Ng, Parsing natural scenes and natural language with recursive neural networks, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 129–136.

[41] H. Li, R. Zhao, X. Wang, Highly efficient forward and backward propagation of convolutional neural networks for pixelwise classification, arXiv preprint arXiv:1412.4526.

[42] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. Torr, Conditional random fields as recurrent neural networks, arXiv preprint arXiv:1502.03240.

[43] G. Lin, C. Shen, I. Reid, et al., Efficient piecewise training of deep structured models for semantic segmentation, arXiv preprint arXiv:1504.01013.

[44] M. Cogswell, X. Lin, S. Purushwalkam, D. Batra, Combining the best of graphical models and convnets for semantic segmentation, arXiv preprint arXiv:1412.4313.

[45] F. Liu, G. Lin, C. Shen, Crf learning with cnn features for image segmentation, Pattern Recognit. 48 (10) (2015) 2983-2992. http://dx.doi.org/10.1016/j.patcog.2015.04.019.

[46] J.J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: Advances in Neural Information Processing Systems, 2014, pp. 1799–1807.

[47] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.

[48] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Trans. Pattern Anal. Mach. Intell. 23 (11) (2001) 1222–1239.

[49] Y. Boykov, V. Kolmogorov, An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision, IEEE Trans. Pattern Anal. Mach. Intell. 26 (9) (2004) 1124–1137.

[50] Y. Freund, D. Haussler, Computer Research Laboratory, Unsupervised learning of distributions of binary vectors using two layer networks, University of California, Santa Cruz, 1994.

[51] G.E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Comput. 14 (8) (2002) 1771–1800.

[52] C. Liu, J. Yuen, A. Torralba, Nonparametric scene parsing: label transfer via dense scene alignment, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE, Miami, FL, USA, 2009, pp. 1972–1979.

[53] D. Eigen, R. Fergus, Nonparametric image parsing using adaptive neighbor sets, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 2799–2806.

[54] G. Singh, J. Kosecka, Nonparametric scene parsing with adaptive feature relevance and semantic context, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Columbus, Ohio, USA, 2013, pp. 3151–3157.

[55] P.H.O. Pinheiro, R. Collobert, P.H.O. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene labeling, in: in International Conference on Machine Learning (ICML), 2014.

[56] A. Vedaldi, K. Lenc, Matconvnet – convolutional neural networks for matlab, CoRR abs/1412.4564.

[57] C. Liu, J. Yuen, A. Torralba, Nonparametric scene parsing via label transfer, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2368–2382.

[58] M. Everingham, S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.

[59] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: 2011 IEEE International Conference on Computer Vision (ICCV), IEEE, Barcelona, Spain, 2011, pp. 991–998.

[60] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[61] M. Mostajabi, P. Yadollahpour, G. Shakhnarovich, Feedforward semantic segmentation with zoom-out features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015. CVPR 2015. IEEE, Boston, USA, 2015, pp. 3376-3385.

[62] G. Papandreou, L.-C. Chen, K. Murphy, A.L. Yuille, Weakly- and semi-supervised learning of a dcnn for semantic image segmentation, arXiv preprint arXiv:1502.02734.

Shuhui Bu received the Ph.D. degree in College of Systems and Information Engineering from University of Tsukuba, Japan, in 2009. Currently, he is an Associate Professor at Northwestern Polytechnical University, China. Prior to joining Northwestern Polytechnical University, he was an Assistant Professor at Kyoto University, Japan. His research includes 3D shape analysis, image processing, and computer vision.

Pengcheng Han received the B.S. degree from Northwestern Polytechnical University of China, in 2013. Currently, he is working toward the M.S. degree at Northwestern Polytechnical University, China. His research interests include artificial intelligence, 3D shape analysis, image processing, and computer vision.

Zhenbao Liu received the Ph.D. degree in computer science from the College of Systems and Information Engineering, University of Tsukuba, Japan, in 2009. He is an Associate Professor at Northwestern Polytechnical University, Xi'an. He was a Visiting Scholar in the GrUVi Lab of Simon Fraser University, in 2012. His research interests include 3D shape analysis, matching, retrieval and segmentation.

Junwei Han is currently a professor with Northwestern Polytechnical University, Xi’an. He received his Ph.D. degree in pattern recognition and intelligent systems from the School of Automation, Northwestern Polytechnical University, in 2003. His research interests include computer vision and multi-media processing.