Neurocomputing 304 (2018) 82–103

Contents lists available at ScienceDirect

Neurocomputing

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

Methods and datasets on semantic segmentation: A review

Hongshan Yu[a,b,*], Zhengeng Yang[a], Lei Tan[a,c], Yaonan Wang[a], Wei Sun[a], Mingui Sun[d], Yandong Tang[e]

[a] National Engineering Laboratory for Robot Visual Perception and Control Technology, College of Electrical and Information Engineering, Hunan University, Changsha, China
[b] Shenzhen Research Institute of Hunan University, Shenzhen, Guangdong 518057, China
[c] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[d] Laboratory for Computational Neuroscience, University of Pittsburgh, Pittsburgh, USA
[e] Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China

ARTICLE INFO

Article history:
Received 22 June 2017
Revised 31 January 2018
Accepted 19 March 2018
Available online 1 May 2018

Communicated by XIANG Xiang Bai

Keywords:
Semantic segmentation
Convolutional neural network
Markov random fields
Weakly supervised method
3D point clouds labeling

ABSTRACT

Semantic segmentation, also called scene labeling, refers to the process of assigning a semantic label (e.g. car, people, and road) to each pixel of an image. It is an essential data processing step for robots and other unmanned systems to understand the surrounding scene. Despite decades of efforts, semantic segmentation is still a very challenging task due to large variations in natural scenes. In this paper, we provide a systematic review of recent advances in this field. In particular, three categories of methods are reviewed and compared, including those based on hand-engineered features, learned features and weakly supervised learning. In addition, we describe a number of popular datasets aiming for facilitating the development of new segmentation algorithms. In order to demonstrate the advantages and disadvantages of different semantic segmentation models, we conduct a series of comparisons between them. Deep discussions about the comparisons are also provided. Finally, this review is concluded by discussing future directions and challenges in this important field of research.

1. Introduction

With the ever-increasing range of intelligent applications (e.g. mobile robots), there is an urgent need for accurate scene understanding. As an essential step towards this goal, semantic segmentation thus has received significant attention in recent years. It refers to a process of assigning a semantic label (e.g. car, people) to each pixel of an image. One main challenge of this task is that there are a large amount of classes in natural scenes and some of them show high degree of similarity in visual appearance.

The emergence of terminology “semantic segmentation” can be dated back to 1970s [1]. At that time, this terminology was equivalent to image segmentation but emphasized that the segmented regions must be “semantically meaningful”. In 1990s, “object segmentation and recognition” [2] further distinguished semantic objects of all classes from background and can be viewed as a two-class image segmentation problem. As the complete partition of foreground objects from the background is very challenging, a relaxed two-class image segmentation problem: the sliding window object detection [3], was proposed to partition objects with bounding boxes. It is useful to find where the objects in the scenes with excellent two-class image segmentation algorithms such as constrained parametric min-cuts(CPMC) [4]. However, two-class image segmentation cannot tell what these objects segmented are. As a result, the generic sense of object recognition(or detection) was gradually extended to multi-class image labeling [5], i.e., semantic segmentation in present sense, to tell both where and what the objects in the scene.

In order to achieve high-quality semantic segmentation, there are two commonly concerned questions: how to design efficient feature representations to differentiate objects of various classes, and how to exploit contextual information to ensure the consistency between the labels of pixels. For the first question, most early methods [6–8] benefit from using the hand-engineered features, such as Scale Invariant Feature Transform (SIFT) [9] and Histograms of Oriented Gradient(HOG) [10]. With the development of deep learning [11,12], the using of learned features in computer vision tasks, such as image classification [13,14], has achieved great success in past few years. As a result, the semantic segmentation community recently paid lots of attention to the learned features [15–26], which are usually refer to Convolutional Neural

* Corresponding author at: National Engineering Laboratory for Robot Visual Perception and Control Technology, College of Electrical and Information Engineering, Hunan University, Changsha, China.
E-mail address: yuhongshancn@hotmail.com (H. Yu).

https://doi.org/10.1016/j.neucom.2018.03.037
0925-2312/

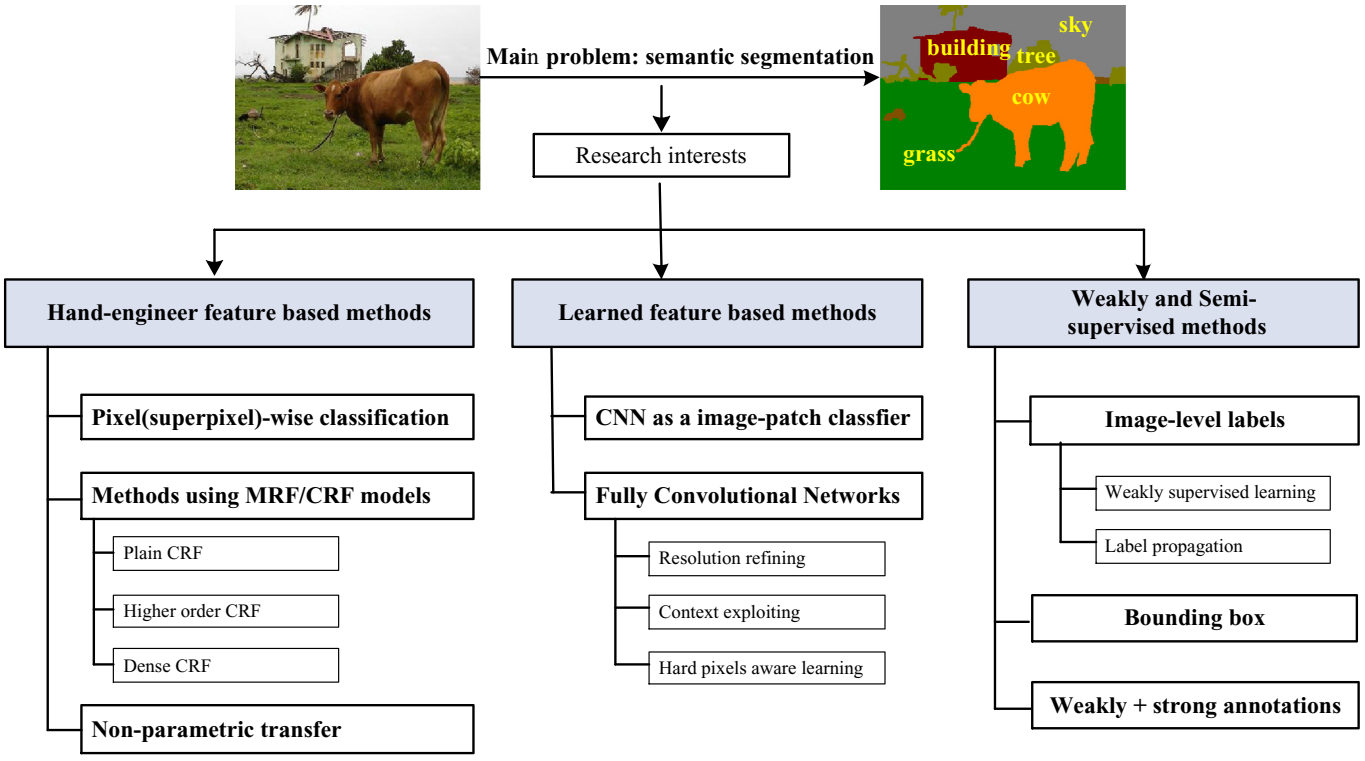

Fig. 1. Classification of existing semantic segmentation methods. According to the current research focus, existing methods can be roughly divided into three main categories. Each of them can be further classified into several sub-categories based on some key characteristics. Here we only provide a very simple description for the name of each sub-category. We refer readers to the corresponding section for more details. The image and its truth are taken from Stanford Background Dataset [32].

Networks(CNN or ConvNets) [27]. For the second issue, the most common strategy, no matter the feature used, is to use contextual models such as Markov Random Field(MRF) [28] and Conditional Random Field(CRF) [6,8,15,16,20,29–34]. These graphical models make it very easy to leverage a variety of relationships between classes via setting links between adjacent pixels. More recently, the use of Recurrent Neural Networks(RNN) [35,36] are more commonly seen in retrieving contextual information. Under the weakly supervised framework [28,37–41], another challenging issue is how to learn class models from weakly annotated images, whose labels are provided at image-level rather than pixel-level. To address this challenge, many methods resort to multiple instance learning(MIL) techniques [42].

Although there are many strategies available for addressing the problems mentioned above, these strategies are not yet mature. For example, there are still no universally accepted hand-engineered features while research on learned features has become a focus again only in recent few years. The inference of MRF or CRF is a very challenging issue in itself and often resort to approximation algorithms. Thus, new and creative semantic segmentation methods are being developed and reported continuously.

The main motivation of this paper is to provide a comprehensive survey of semantic segmentation methods, focus on analyzing the commonly concerned problems as well as the corresponding strategies adopted. Semantic segmentation is now a vast field and is closely related to other computer vision tasks. This review cannot fully cover the entire field. Since excellent reviews on research achievements on traditional image segmentation, object segmentation and object detection already exist [43,44], we will not cover these subjects. We will instead focuses on generic semantic segmentation, i.e., multi-class segmentation. Based on the observation that most works published after 2012 are CNN based, we will divide existing semantic segmentation methods into those based on hand-engineered features and learned features (see Fig. 1). We will discuss weakly supervised methods separately because this challenging line of methods is being investigated actively. It should be emphasized that there are no clear boundaries between these three categories. For each category, we further divide it into several sub-categories and then analyze their motivations and principles.

The rest of this paper is organized as follows. Before introducing recent progresses on semantic segmentation, preliminaries of commonly used theories are given in the next section. Methods using hand-engineered features and learned features are systematically reviewed in Sections 3 and 4, respectively. The efforts devoted to weakly supervised semantic segmentation are described in Section 5. In Section 6, we describe several popular datasets for semantic segmentation tasks. Section 7 compares some representative methods using several common evaluation criteria. Finally, we conclude the paper in Section 8 with our views on future perspectives. Note that semantic segmentation also called as scene labeling in literature, we will not differentiate between them in the rest of this paper.

2. Preliminaries

We start by describing the commonly used theories and technologies in the semantic segmentation community, including superpixels and contextual models.

2.1. Superpixels

As argued by Ren and Malik [45], the superpixel that consists of a set of similar and connected pixels is more appropriate for representing entities compared to the pixel. The benefits of using superpixel can be summarized into two aspects. First, the computational complexity is greatly reduced by treating a set of pixels as a single pixel, i.e., the superpixel. Second, a region is able to

provide much more cues compared with a single pixel. Thus, the use of superpixel becomes very popular in computer vision.

Generally, the results of conventional segmentation, such as mean-shift [46], Normalized cut [47], can be directly taken as superpixels. In addition, the numbers of superpixels can be controlled by tuning the parameters of these algorithms. For implementation details of popular conventional methods, we refer readers to several recent excellent reviews [43,48]. However, superpixels produced from conventional methods usually lack of shape control and lead to inefficiency in feature extraction. In recent years, some faster and better superpixel algorithms were presented in succession, such as Turbopixel [49] and Simple Linear Iterative Clustering(SLIC) [50]. Turbopixel produces superpixels through dilating a set of seed locations using geometric flow, which is computationally rooted in the curve evolution techniques. The superpixels in SLIC are generated by clustering the pixels using a localized K-means algorithm, which means that the cluster area is restricted to a local window.

2.2. Contextual models

It is well known that many successful image processing algorithms benefit from using graph. Generally, graph based methods map an image onto an undirected graph $G = \{V, E\}$, where V is the set of vertices composed of all the pixels, and E is the set of edges that connect adjacent pixels. Moreover, each edge in the graph is associated with a weight which depends on some characteristics of the two pixels it connects. In the sematic segmentation community, the graph is usually used in conjunction with Markov theory to build contextual models, such as MRF and CRF.

2.2.1. MRF

Given the number of semantic classes L and the number of pixels N, denoted the semantic label of each pixel i as y_i, the ultimate result of scene labeling can be represented by $y = \{y_1, \ldots, y_i, \ldots, y_N\}$, where y_i can take any label l from the discrete class set $\{1, 2, \ldots, L\}$. Thus, all the latent label make up a set Y with L^N elements. In the MRF framework, the prior over label $y(P(y))$ is often modeled as a MRF defined on pixel lattice with a neighborhood system ε. Thus, given the observation x of an image, according to the Bayes theory, the posterior distribution of y can be written as

$$P(y|x) \propto P(x|y)P(y). \tag{1}$$

In addition, pixels are often assumed to be independent and identically distributed to compute $P(x|y)$.

$$P(x|y) = \prod_{i=1}^{N} P(x_i|y_i) \tag{2}$$

Then, the labeling problem is equivalent to maximum a posteriori(MAP)

$$y^* = arg\max_{y \in Y} P(y|x). \tag{3}$$

MRF-MAP based methods are essentially related to generative models that need to estimate the feature distribution for each class. As argued in [51], such generative model based MRFs suffer from two main drawbacks when applying to image processing. First, the assumption of conditional independence between pixels is very restrictive. On top of that, the MAP inference may be quite complex even though the class posterior is simple. Hence, besides early studies on image restoration [52,53], only a few recent studies [28] are trying to solve scene labeling problem using the MRF-MAP framework. In contrast, the CRF-MAP is more popular in recent years.

2.2.2. CRF

A CRF [54] can be viewed as a variant of MRF. As a result, some studies do not differentiate between the two concepts. The CRF directly models the posterior distribution $P(y|x)$ as a Gibbs distribution [51]

$$P(y|x) = \frac{1}{Z} exp(-U(y)) = \frac{1}{Z} exp(-\sum_{c \in C} \Psi(y_c|x_c)), \tag{4}$$

where y is one of the label states and U is the corresponding energy. Z is a normalized constant equal to the sum of energy of all states, which ensures the distribution summed to one. The $\Psi(y_c|x_c)$ are potential functions defined over a clique $c \in C$ that contains a set of connected pixels. Thus, the solution of MAP can be formulated as a problem of energy minimization

$$y^* = arg\max_{y \in Y} P(y|x) = arg\min_{y \in Y} \sum_{c \in C} \Psi(y_c|x_c) = arg\min_{y \in Y} E(y|x). \tag{5}$$

The commonly used energy function defined on two types of cliques is given by

$$E(y) = \sum_{i=1}^{N} \varphi_i(y_i|x_i) + \sum_{(i,j) \in \varepsilon} \varphi_{ij}(y_i, y_j|x_{ij}), \tag{6}$$

where φ_i and φ_{ij} are called unary potential and pairwise potential, respectively. The unary potential reflects how appropriate the assigned label y_i for a pixel i. It is usually defined as the negative log of the likelihood of a label being assigned to pixel i [30].

$$\varphi_i(y_i) = -logP(y_i|x_i) \tag{7}$$

The pairwise potential is used to model the relationships between pixels. If there is only the unary potential, minimizing the energy is equivalent to assigning the pixel with the most appropriate label, i.e., performing pixel-wise classification. From this perspective, the pairwise potential can be understood as an additional constraint for ensuring the consistency between predictions of pixels.

2.2.3. Inference and energy minimization

RF-MAP based methods shift the focus to minimizing energy function. Unfortunately, the global minimization of an energy function is NP-hard due to the fact that there are many local minima [55]. This problem forces scholars to find an approximate solution and a number of strategies have been reported, such as Iterated Conditional Modes(ICM) [53] and Simulated Annealing(SA) [56,57]. The idea behind these approximation approaches is to iteratively decrease the energy through a standard move process(change the label of one pixel) until a convergence to a local minimum is reached. However, it is much easier to give low quality approximations with standard moves because the energy can be hardly decreased by changing a single pixel's label when falling into a bad local minimum. To solve this problem, Boykov et al. [55] reported two graph based algorithms using two types of large moves:α-expansion-moves and $\alpha - \beta$-swap-moves. Started at an arbitrary labeling y, the α-expansion algorithm iteratively expands each class label on y in a cycle, where α-expansion means α is allowed to be "expanded" to any pixel not belonging to α. If the result of any expansion, which is formulated as a bi-labeling$(\alpha, \overline{\alpha})$ problem and solved by s/t graph cuts algorithm [58], in a cycle is strictly better than previous, the algorithm continues the expansion until no further improvement happens. The $\alpha - \beta$-swap is quite similar to α-expansion in structure. The only difference lies in the process of move, where $\alpha - \beta$-swap means that pixels labeled $\alpha(\beta)$ are allowed to change to $\beta(\alpha)$.

The s/t graph cuts attempts to map two types of energy terms in (6) onto the weighted-edges so that the min-cut algorithm [59] is available for energy minimization. However, typical graph

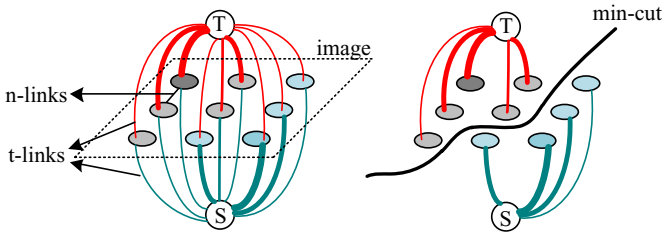

Fig. 2. Illustration of s/t graphcuts [58]. For convenience, we omit most edges between pixels and only reserve connections (t-links) between labels and pixels. The width of lines represents the probability of pixels belonging to its connected label.

includes only one type of links(n-links) that connect adjacent pixel pairs, onto which the pairwise term can be mapped. To map the unary term of bi-labeling energy, the s/t graph cuts defines another type of links(t-links) that connect pixels and virtual terminal nodes *s* and *t*. Indeed, the terminal nodes represent the label of segmented regions: "object" and "background". With this graph model and special designed weighted-edges, the energy based bi-labeling can be realized via min-cut. A simple example is shown in Fig. 2.

Besides move-making based methods outlined above, another popular strategy for minimizing energy is to use message-passing. The most well-known method in this framework is the Loopy Belief Propagation(LBP) algorithm [60], which is developed from the Belief Propagation(BP) [61]. The principle of LBP is to iteratively update the belief(approximation of marginal probability) of each node by updating messages between nodes. The message msg_{i-j} can be simply understood as the likelihoods of various states of the node *j* that supposed by the node *i*. It is usually obtained by computing the sum or the maximum of beliefs produced by all latent labels of the node *i*, which results in two main variants of LBP: the sum-product LBP and max-product LBP. The max-product version is widely used for energy minimization while the other one is usually used for computing the marginal distribution of each node in a graph. An important advantage of LBP is that its implementation is simple and can achieve a comparable performance to the α-expansion algorithm. The idea of message-passing has also been adopted by other approximation algorithms such as mean-field inference [62].

Here we introduce several most important inference algorithms used in the scene labeling community. While the study of energy minimization is still an open issue, we refer readers to a complete review of recent inference methods [63].

3. Hand-engineered features based scene labeling methods

In this section, we focus on reviewing scene labeling methods that rely on hand-engineered features, such as SIFT [9], HOG [10] and Spin images [64]. Note that we will not discuss the local features in detail owing to page limitations. We refer readers to a comprehensive review [65]. In addition, it should be emphasized that "visual words" described in this section may partially belong to learned features because, in most cases, they are learned from well-designed hand-engineered features. We treat "visual words" as hand-engineered features despite its dual statuses.

3.1. Methods using pixel(superpixel)-wise classification

Most early methods address the issue of scene labeling by performing pixel-wise classification according to the features associated with each pixel. Konishi and Yuille [66], for example, performed Bayesian classification for each pixel according to the prior probability of classes and the posterior distribution of low-level features. Schroff et al .[67] used a single histogram of visual words to model each class. The label of each pixel is then decided by the histogram distances between the pixel and all class models. Although this kind of method is easy to implement, classifying pixels separately would easily yield inconsistency in the results since it ignores the fact that pixels are usually associated with each other. The auto-context [68] proposed by Tu and Bai addressed the problem by iteratively using the pixel-wise predictions as context information, in addition to the appearance features, to train a new pixel-wise classifier. This method is similar to a multi-class object detection framework called Mutual Boosting [69], which showed that the object detection of a specific class can be improved by employing detection beliefs of other classes. Clearly, the resulting boosting framework involves a tedious learning process. By replacing single label supervision with structure label supervision, Kontschieder et al. [70] developed a more efficient way to exploit context for pixel-wise classification.

An alternative way to mitigate the inconsistency problem mentioned above is to enforce pixels belonging to a region have the same label. Thus, several methods [71–73] formulated the problem of scene labeling as a superpixel-wise classification. In this framework, features of each superpixel are usually obtained by computing certain statistics(e.g. mean) of local descriptors within the superpixel. For example, a second-order pooling technique was developed in [72]. Instead of directly pooling local descriptors, Gupta et al. [73] quantized each local feature into a visual word and then computed a HOG descriptor for each superpixel. The use of superpixel brings at least two benefits to the task of scene labeling. First, the computational complexity is significantly reduced. Second, it is more likely extract robust features from superpixels rather than pixels. Nevertheless, the region-based model has its own limitations. Such model assigns the same label to all pixels of a region with the premise that some superpixels share boundaries with objects in the image. However, even the best over-segmentation approach would produce misleading results, that is, some superpixels may contain several object classes. Using multiple over-segmentations [74] provides a strategy to alleviate this problem, however, performing multiple "segmentation and classification" processes is computationally expensive.

3.2. Methods using CRF

A more powerful way to ensure the spatial consistency is to employ a contextual model, such as MRF and CRF, to explicitly model the relations between pixels or regions.

3.2.1. Plain CRF

The most commonly used CRF model in semantic segmentation is the pairwise CRF of (6), here we called it plain CRF for convenience. In this case, the pairwise potential is usually defined in terms of a Potts model or its variants [6,29,33,75–77], which is contrast sensitive and can be written as

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

where $K(y_i, y_j)$ is a penalizing function whose output depends on the difference between adjacent pixels. If there is a large difference, the penalization should be small so as to preserve a reasonable label transition. However, the contextual relationships exploited by this pairwise potential are very restrictive and can be interpreted as "guilt by association" [78]: connected(adjacent) pixels tend to have the same label. In this situation, many works encoded contextual information into the observed data x_i when constructing the unary potential. For example, based on the fact that some objects likely to appear simultaneously within a certain range, Shotton et al. [29] counted the numbers of several textons(visual words) in a rectangular mask that offsets from each pixel and

then took them as contextual features. Other works have employed prior information for exploiting contexts. He et al. [79] learned environment-specific label distribution priors and then developed an prior-dependent CRF for scene labeling. Silberman and Fergus [6] multiplied the unary predictions with 2D and 3D absolute location priors. These methods provide opportunities for exploiting context in associative CRF, however, such context is often very limited. For examples, the inter-class correlations are not fully exploited both in prior-dependent CRF [79] and absolute location priors [6], while the offset-texton [29] do not consider global contexts.

Associative potential lack of the capability of modeling inter-class relationships (e.g., "on-top-of"). Thus, a number of works suggest to use the non-associative potential, which allows various combinations of labels between adjacent pixels (superpixels), to take much more contextual information into account. An early method [34] in this direction achieved this goal by extracting prior label patterns, i.e., the arrangements of labels in a region, from the training data. The potentials are then decided by the matching degree between test label patterns and prior patterns. However, there are usually many priors that cannot be reflected by the training data. One way to overcome this limitation is to replace the priors with likelihoods [32,80,81]. Gould et al. [32], for example, extracted features for every pair of adjacent regions and learned a multi-class logistic classifier to output the co-occurrence likelihoods of two labels. Similarly, Batra et al. [81] learned the adjacent relationships between classes based on the bag-of-words representation. Muller and Behnke [82] were among the first to apply such learned pairwise potential to indoor scene labeling.

Compared to associative potential, the non-associative potential is able to exploit much more contextual information through encoding various kinds of inter-class relationships. Nevertheless, the model also becomes more complex and hence brings much more difficulties in parameter learning and label inference. Considering the commonly used co-occurrence likelihoods, given the number of classes L and the number of edge features K, it will need to take L^2 class combinations into account and hence require for L^2K parameters. As a result, the unary model and pair-wise model are usually trained separately in non-associative CRF. Although Szummer et al. [83] have tried a max-margin based method to jointly learn the CRF parameters, the training process involves multiple approximate MAP inferences, which inevitably lead to a reduction in the accuracy of the model.

3.2.2. Higher order CRF

The main disadvantage of plain CRF, which defined on small pairwise cliques, is that it is weak to capture long-range dependencies. To alleviate this problem, a number of works have studied various higher order potentials defined on large cliques.

One well known higher order potentials is proposed by Kohli and Torr [30] and defined on the image segments in the form of robust P^n Potts model, i.e.,

$$\varphi_c(y_c) = \begin{cases} N_i(y_c)\frac{1}{Q}\gamma_{max} & if\ N_i(y_c) \le Q \\ \gamma_{max} & otherwise \end{cases}, \tag{9}$$

where $N_i(y_c)$ denotes the number of pixels in a region c not taking the dominate label; Q is a constant which controls the rigidity of the higher order potential, and γ_{max} is the sum of cost of assigning a non-dominate label to a pixel in region c. Such potential encourages pixels lie in the same superpixel to have the same label and hence is useful to produce fine boundaries in the segmentation result. As shown in [31], higher order potentials defined on superpixels can be understood as a group of pairwise potentials between a superpixel node and pixel nodes. Thus, by recursively performing superpixel segmentation and then constructing higher order CRFs over them, it is easy to build hierarchical CRFs[31,84] to exploit multi-level contexts. Such process is illustrated in Fig. 3.

Beyond superpixels, several other large cliques have also been utilized to build higher-order potentials. For example, Wojek and Schiele [85] employed connections between object hypothesis nodes and pixel nodes to ensure consistency between pixel labels and object detection results. The same authors also added a unary detection potential and achieved both object detection and segmentation in a unified CRF. Similarly, Gonfaus et al. [86] presented a harmony potential, consisting of a group of connections between the image node and pixel nodes, for joint image classification and segmentation. The allowed 2^L possible labels for the image node make the joint optimization intractable. Lucchi et al. [87] solved this problem by using class-specific image nodes, which sacrifices the capability of modeling category co-occurrence. More recently, Khan et al. [8] defined higher-order potentials on 3D planar regions constructed from depth images. These multi-task based higher-order potentials have the advantage that they can be easily extended by incorporating various constraints from other tasks. Nevertheless, the implementation of multi-task is usually very tedious. More importantly, it is possible to produce misleading constraints since the implementation of other vision tasks such as object detection is very challenging in itself.

3.2.3. Dense CRF

Higher-order CRF defined on superpixels suffers from a drawback that superpixels produced from unsupervised segmentation may contain misleading boundaries. The dense CRF is able to avoid this problem while capturing long-range dependencies. Not limited to adjacent pixels, the dense CRF also considers interactions between long-range pixels. For example, Torralba et al. [88] established potentials over pixel pairs that have shown correlations in the training set. This method needs to learn the correlation patterns for each class and hence lack of generality when more classes are considered. The generic dense CRF [62,89] instead connects all pixel pairs no matter whether they are correlated or not. Since the inference of dense CRF is computationally expensive, [89] still resorted to superpixels to ensure computational efficiency. By expressing the massage passing step as a convolution with a Gaussian kernel, Krahenbuhl and Koltun [62] developed an efficient mean-field inference algorithm for a dense CRF in which the pairwise potentials are defined by a linear combination of Gaussian kernels. This approach provided a generic and efficient strategy for exploiting long-range dependencies and thus became popular in subsequent works, including methods using learned features introduced later.

At this point, we have critically reviewed scene labeling methods based on contextual models. For the sake of clarity, we summarize them in the Table 1 according to several key characteristics, including scene type, model structure, etc. The abbreviation of superpixel algorithms used in region based methods are described as follow: Mean-shift(MS) [46], Normalized Cuts(Ncut) [47], Felzenszwalb–Huttenlocher(FH) [91], Turbopixels(TP) [49], Simple Linear Iterative Clustering(SLIC) [50] and globalized probability of boundary-oriented watershed transform(gPb-OWT) [92].

3.3. Non-parametric methods

All methods outlined so far work with a fixed number of classes and need to use sophisticated parametric models, especially the non-associative CRF method. When it is necessary to take new classes into account, all model parameters need to be adjusted, which would be very tedious. To avoid this problem, Liu et al. [93] developed a non-parametric approach for scene labeling. The core idea of this method can be described as "label transfer" (see Fig. 4), which means the labels of test pixels are transferred from

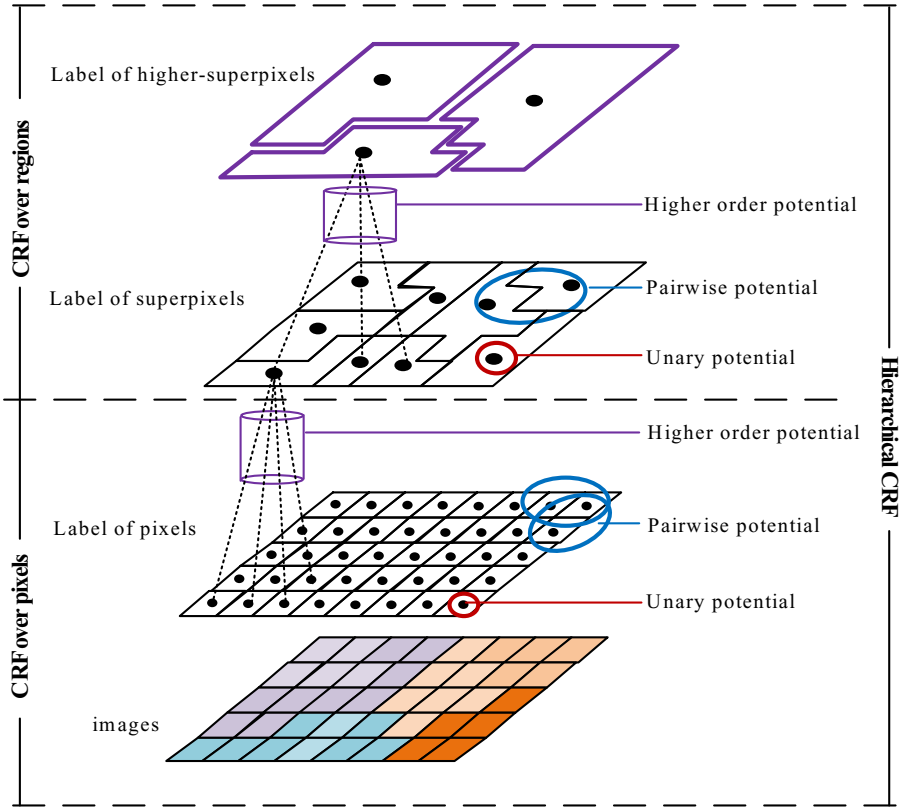

Fig. 3. Three architectures of CRF model used in scene labeling. For CRF over pixels , graph nodes (second layer) are composed of labels of each pixel. On the other hand, the graph nodes (third layer) in CRF over regions are the labels of superpixels. The unary potential and pairwise potential are similarly defined in both models. Viewing the pixel as an elementary superpixel, several CRF over regions models can be organized into a hierarchical CRF [31].

Table 1
Summary of contextual model based scene labeling methods.

Papers	Scene type	Model structure		Contextual information	Inference
		CRF structure	others		
Shotton et al. 2006 [29]	outdoor	Plain CRF	pixel-based	Smoothness preference; Relative texture pattern; 2D location prior	graph cuts
Fulkerson et al. 2009 [75]	outdoor		region-based	Smoothness preference	ditto
Zhang et al. 2010 [76]	outdoor		region-based(TP)	ditto	ditto
Silberman et al. 2011 [6]	indoor		pixel-based	Smoothness preference 2D/3D location prior geometry location prior;	ditto
Ren et al. 2012 [33]	indoor		region-based(gPb-OWT)	Smoothness preference	ditto
Cadena et al. 2013 [77]	indoor		region-based(SLIC)	ditto	LBP
Gould et al. 2009 [32]	outdoor		region-based(MS)	co-occurrence likelihoods	ICM
Batra et al.2008 [81]	outdoor		region-based(FH)	ditto	LBP
Muller et al. 2014 [82]	indoor		region-based(SLIC)	ditto	–*
Kumar et al. 2015 [90]	outdoor		region-based(MS)	ditto	ICM
Gould et al. 2008 [80]	outdoor		region-based(Ncut and FH)	relative location prior	LBP
Kohli et al. 2008 [30]	outdoor	Higher-order CRF	superpixel clique	Smoothness preference;	graph cuts
Ladicky et al. 2009 [31]	outdoor		hierarchical(MS)	ditto;	ditto
Plath et al. 2009 [84]	outdoor		hierarchical(FH)	global consistency;	–
Wojek et al. 2008 [85]	outdoor		object clique	local/object consistency ;	LBP
Gonfaus et al. 2010 [86]	outdoor		image clique	local/global consistency	–
Lucchi et al. 2011 [87]	outdoor		ditto	ditto	–
Khan et al. 2014 [8]	indoor		3D plane clique	local consistency	–
Torralba et al et al. 2005 [88]	outdoor	Dense CRF	boosting framework	object correlations	BP
Rabinovich et al. 2007 [89]	outdoor		region-based(NCut)	co-occurrence prior	–
Krahenbuhl et al. 2011 [62]	outdoor		pixel-based;	local/global dependency	mean-field

*Inference algorithms that not introduced in this paper.

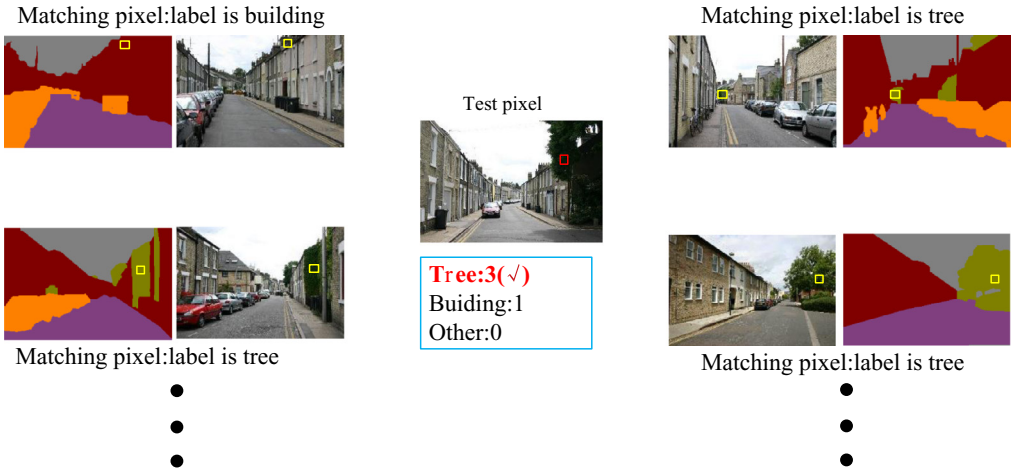

Fig. 4. Illustration of the idea of label transfer. The red rectangle is the test pixel and its matching pixels found in the training set are shown by the yellow rectangles. The labels of matching pixels can be naturally transferred to the test pixels. The color images and theirs' truth are taken from Stanford Background Dataset [32] (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

their matching pixels found in labeled images. To accomplish this, scene retrieval is first executed to find a set of most similar labeled images of a test image. Then, dense scene correspondences(pixel to pixel) between input and each of the similar images is established by matching local SIFT descriptors. The images with bottom matching energy are chosen as the ultimate candidates for label transfer. Instead of using pixel-level transferring, Tighe and Lazebnik [94] transferred labels at the level of superpixels and achieved better performance compared with [93].

Non-parametric methods provide a more generic framework for scene labeling. There is no need for tedious training no matter how many classes must be taken into account. One can simply add sufficient images including these classes into retrieval sets. The main limitation of non-parametric methods is that the computational complexity increases remarkably with the increase of image size and retrieval sets. To alleviate this problem, without retrieving similar labeled images for a test image, Gould et al. [95] presented an algorithm for directly finding good matches for test superpixels. This is achieved by first constructing an optimum graph, in which nodes represent superpixels and edges represent match cost defined by a learned distance, from training images. Given a set of superpixel of a test image, the graph can be rapidly reconstructed to find their matches.

3.4. 3D scene labeling methods

With the development of 3D sensors, such as Light Detection and Ranging(LIDAR), 3D laser scanners and RGB-D cameras, it is feasible to obtain 3D point clouds of the environment or objects [96]. As a result, the concept of image labeling has been extended to 3D point clouds labeling. Like methods outlined in 3.2, most 3D scene labeling methods are formulated as the problem of finding the most appropriate labeling of a CRF. The representative methods are summarized in the Table 2.

Unlike 2D scene labeling, in which the using of associative potential is very common, the non-associative potential is more popular in 3D point cloud labeling. This can be explained by the difference between the information provided by 2D images and 3D point clouds. With the explicit and discriminative visual appearance(e.g., color), the unary classifier is able to provide a good(albeit imperfect) result. Hence, majority of CRF based 2D scene labeling methods set the pairwise potential with a simple Potts model, merely aim for enforcing local smoothness. However, the information that can be exploited from 3D point clouds is more of implicit geometric structure, within which considerable geometry contextual are hidden. Consequently, the research community has shifted its focus to non-associative potential to employ as much geometric relationships as possible.

Another significant difference between 2D and 3D CRF models lies in the selection of learning strategy. Because of the Potts model based potential is simply decided by the difference between pixels(superpixels) and several manual tuned parameters. Most 2D scene labeling methods only need to train a unary classifier, such as neural networks [6], Random Forests [76] and SVM [33]. A bit more complicated learning strategy used in 2D scene labeling is piece-wise learning, which involves training different types of potential models independently and adding weights to each of them through cross-validation [29,30,32]. In contrast, most 3D scene labeling methods use structural learning algorithms to obtain parameters of the CRF model, that is, all the parameters are jointly learned from training samples. As we know, a common approach to learn parameters is to perform maximum likelihood estimation. Unfortunately, the computation of the true likelihood is intractable. Instead, the pseudolikelihood approach is widely used in the literature [100–102]. Another popular structural learning algorithm is to use a max-margin based objective, which aims to maximize the margin of confidence between the true labels and any other predictions [98,103–105].

4. Learned features based scene labeling methods

In this section, we review the representative scene labeling methods based on learned features, which usually refer to convolutional neural networks(CNN). Typical CNNs take inputs of fixed-size and produce a single prediction for whole-image classification. However, scene labeling aims to assign a class label to each pixel of arbitrary-sized images. In other words, when applying typical CNNs to scene labeling, the main problem is how to produce pixel-dense predictions. To this end, we first describe the principle of the CNN and then analyze different strategies of producing dense-predictions.

4.1. Principle of CNN

As most literatures do, we refer standard CNN to the LeNet5, which was proposed by LeCun et al. [27,107] and initially designed for handwriting recognition. CNN is a variant of multilayer neural networks(NN) and biologically-inspired by an early research

Table 2
Summary of influential 3D scene labeling methods.

CRF model structure		paper	3D region	Learning strategy
associative	point-based	Lu et al. 2012 [97]	–	Piecewise learning
		Anguelov et al. 2005 [98]	–	Max-margin structural learning
	3D region-based	Valentin et al. 2013 [99]	Meshes	Piecewise learning
non-associative	point-based	–	–	–
	3D region-based	Xiong et al. 2010 [100]	Planar patches	Pseudolikelihood structural learning
		Kahler et al. 2013 [101]	SLIC-like supervoxels	ditto
		Pham et al. 2015 [102]	ditto	ditto
		Anand et al. 2012 [103]	region growing based supervoxels	Max-margin structural learning
		Kim et al. 2013 [104]	Small cubic	ditto
		Wolf et al. 2015 [105]	SLIC-like supervoxels	ditto
		Najafi et al. 2014 [106]	K-means based supervoxels	Piecewise learning

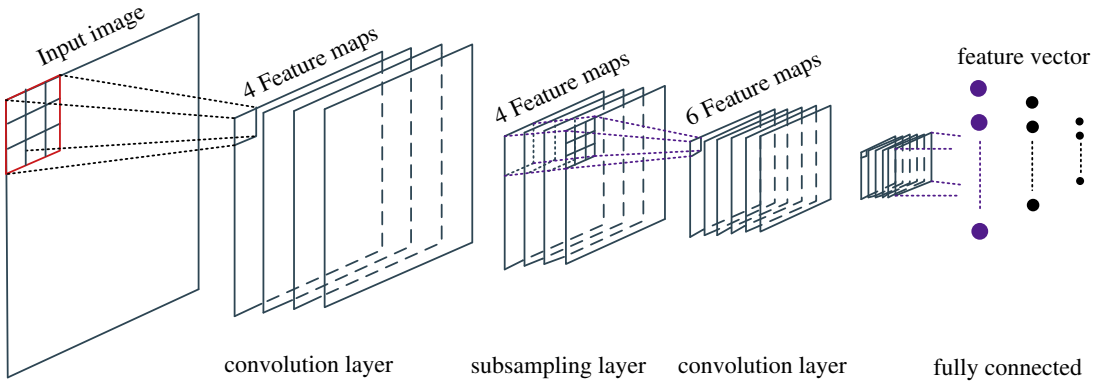

Fig. 5. Simplified architecture of CNN. Different feature maps in the first convolution layer are obtained by performing convolution over the input image with different templates. Each feature map in the intermediate convolution layer are obtained by performing convolution over several or all the maps of previous layer with different templates and then summing all the results.

[108] on the visual cortex of cats. A simplified architecture of CNN is shown in Fig. 5. Unlike the standard neural networks, which use fully connectivity between layers, the CNN considers a sparse connectivity structure based on the local perception mechanism of visual cortex. In other words, the inputs of a unit are from a set of units located in a small window(called receptive field) in the previous layer. In addition, all the units in the same layer share the identical weight mask and the outputs of these units are organized into a so-called feature map. However, a specific weight mask can only extract one type of feature. In order to form a richer representation of the input, a complete convolutional layer usually consists of several feature maps constructed from different weight masks. Let l^k be the kth feature map in an convolution layer l. Then, without padding zeros at image boundaries, the feature value of each unit l^k_{ij} is given by

$$l^k_{ij} = tanh\left(\sum_{h=1}^{H}\left(w^{kh} * x^h_{i+\lfloor s/2\rfloor, j+\lfloor s/2\rfloor} + b_{kh}\right)\right), \tag{10}$$

where w^{kh} and b_{kh} are the weight vector and bias between l^k and $(l-1)^h$, respectively. H is the number of feature maps of layer $l-1$. $x^h_{i+\lfloor s/2\rfloor, j+\lfloor s/2\rfloor}$ is a local vector of feature map $(l-1)^h$ centered at $(i+\lfloor s/2\rfloor, j+\lfloor s/2\rfloor)$, where s is the size of weight mask w^{kh}. $tanh$ is a well-known nonlinear activation function. It should be emphasized that a feature map is not required to connect to all maps in the previous layer although the equation indicates so (just for convenience).

In the real world, the positions of features vary for different instances of the same object and likely to change with shifts and distortions of input. As a consequence, the convolutional layer is usually followed by a subsampling (also called pooling) layer, which performs local averaging or maximizing over each feature map, to achieve robustness to shifts and distortions.

The CNN received great attention after its presentation. However, with the development of other efficient algorithms such as SVM, it fell out of fashion due to computational concerns. In recent years, with the rise of deep learning [11], the interest in CNN is rekindled by Krizhevsky et al. [13]. They achieved a considerably higher image classification accuracy on the ImageNet Large Scale Visual Recognition Challenge(ILSVRC) 2012 through using a deep CNN(called AlexNet). Indeed, there is no difference between the AlexNet and the LeNet5 in structure. However, the scale of the AlexNet is significantly larger and thus many strategies were adopted for efficient training and over-fitting prevention, such as non-saturating activation function(e.g., $max(x, 0)$), parallel computing with Graphics Processing Unit(GPU) and "dropout" regularization.

4.2. Scene labeling based on CNN

4.2.1. Naive approach

A straightforward way, what we called naive approach, to produce dense predictions is to take CNNs as image patch classifiers or feature extractors and then perform pixel/region-wise classification.

In the framework of pixel-wise classification, CNNs are usually applied to image patches of fixed-size centered at each pixel [15,20,23,109] (see the top of Fig. 6.). However, labeling each pixel by just observing a small region around it is insufficient. It is necessary to take a wide context into account to make a local decision. This can be achieved simply by increasing the size of image patches. Unfortunately, this solution would increase the number of parameters of CNN drastically and thus lead to much higher computational complexity. In order to take wide contexts into account while ensure computational efficiency, multi-scale strategies were successively proposed. Farabet et al. [15], for instance, applied CNN

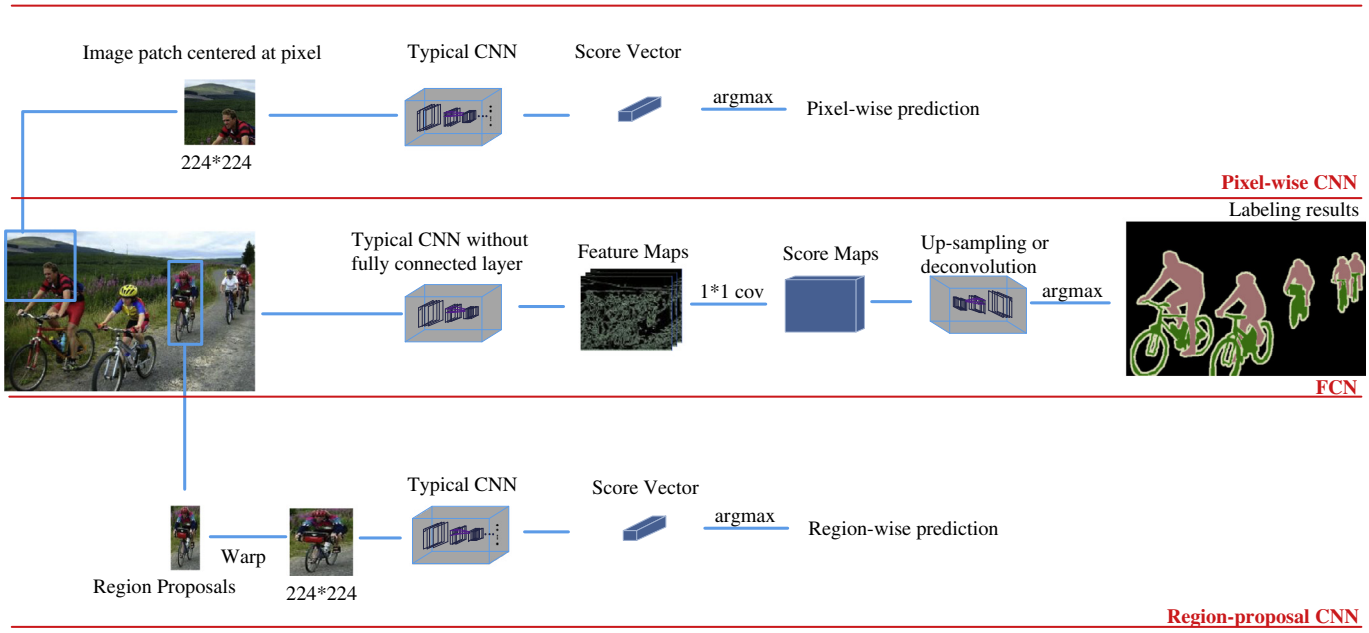

Fig. 6. Three common CNN architectures for semantic segmentation. The color image and its truth are taken from PASCAL VOC dataset [113].

to patches of images in a multi-scale pyramid of the input. The feature maps generated from different scales were then concatenated after a size matching procedure. As a result, each pixel is associated with a feature vector encoded from multiple patches with the same sizes but increasing visual fields. This multi-scale architecture was also adopted by Couprie et al. [20] to extract features from the depth information provided by RGB-D sensors. Similarly, in [23], more scales were considered.

Although it is structurally straightforward to feed image patches to a typical CNN to produce dense predictions, this scheme is computationally inefficient. There are a large amount of redundant convolutions because patches centered at adjacent pixels are significantly overlapped. One way to alleviate this problem is to perform region-wise classification by utilizing the concept of region proposals [17,21]. This type of methods has the advantage that the region proposals, which are usually in the form of rectangular shape, can be directly fed into a CNN for classification after warping. In addition, object detection and segmentation can be achieved simultaneously in this framework. However, rectangular region proposals contain not only the object class but also several other classes(see the bottom of Fig. 6). As a result, many bottom-up region proposals must be taken into account to determine the label of a single pixel.

Mostajabi et al. [110] took a CNN as a feature extractor and produced pixel-level features by up-sampling the last feature maps. Superpixel features were then extracted by pooling pixel-level features. In addition, the authors proposed to concatenate the features of intermediate layers to capture multi-scale information. However, up-sampling these multi-scale feature maps to image resolution would lead to memory sensitive since there are usually hundreds of feature maps in each scale. In contrast, [111,112] used a recursive context propagation network, which consists of a stage of bottom-up feature aggregation and a process of top-down context propagation, to enrich each superpixel with contextual information.

4.2.2. Methods using fully convolutional networks

Recent advances in semantic segmentation are mostly achieved by using fully convolutional networks(FCN). As described in Fig. 6, the basic idea of FCN is to replace the fully connected layer of a typical CNN with a 1×1 convolution layer to produce low-resolution predictions. To obtain pixel dense predictions, a trivial approach [109] called shift-and-stitch stitches predictions mapped from multiple shifted versions of the input. This method does work because the output of the CNN is shifted with the shift of the input. Considering output score maps with input stride s, the shift-and-stitch method needs to process s^2 shifted versions of the input image. A more efficient strategy performs up-sampling(e.g., bilinear interpolation) over the coarse predictions. Here we called it UP_FCN for convenience. For examples, [19,22] adopted a deconvolution layer to up-sample the low-resolution predictions. The UP_FCN based method enables end-to-end training and thus becomes a main stream method in scene labeling. The end-to-end training means that the feature representation and pixel-wise classifiers are jointly learned. Based on the structure of FCN, the strategies designed for improving semantic segmentation can de divided into three categorises.

Resolution Refining: Directly up-sampling predictions of low-resolution(typically 1/32 of the original input) to image resolution could cause a great loss. The SegNet [114] instead used a hierarchy of decoders to recover the feature maps. To control the scale of the model, the authors performed dimensionality reduction over the feature maps, which scarifies the model accuracy. One can also remove several down-sampling operations to obtain finer-resolution predictions before up-sampling. Unfortunately, doing so could decrease the receptive field sizes of the last layer and thus lead to a loss of global and contextual information crucial to semantic segmentation. To solve this problem, several methods [16,25,115,116] introduced atrous convolution to output finer-resolution(typically 1/8) score(or feature) maps. A similar method called dilated convolution was reported in [117]. The basic idea is to dilate(insert zeros) the convolution filters behind the removed down-sampling layers. The atrous convolution introduces zeros only inside of the original convolution mask, while the dilated convolution also inserts zeros outside(see Fig. 7).

The most notable advantage of atrous_FCN model is that it can produce fine-resolution predictions without introducing other parameters. However, the large receptive filed used in this model can hardly capture low-level visual information, which introduces another way of producing finer-resolution outputs that adds skip connections to middle layer features. For examples,

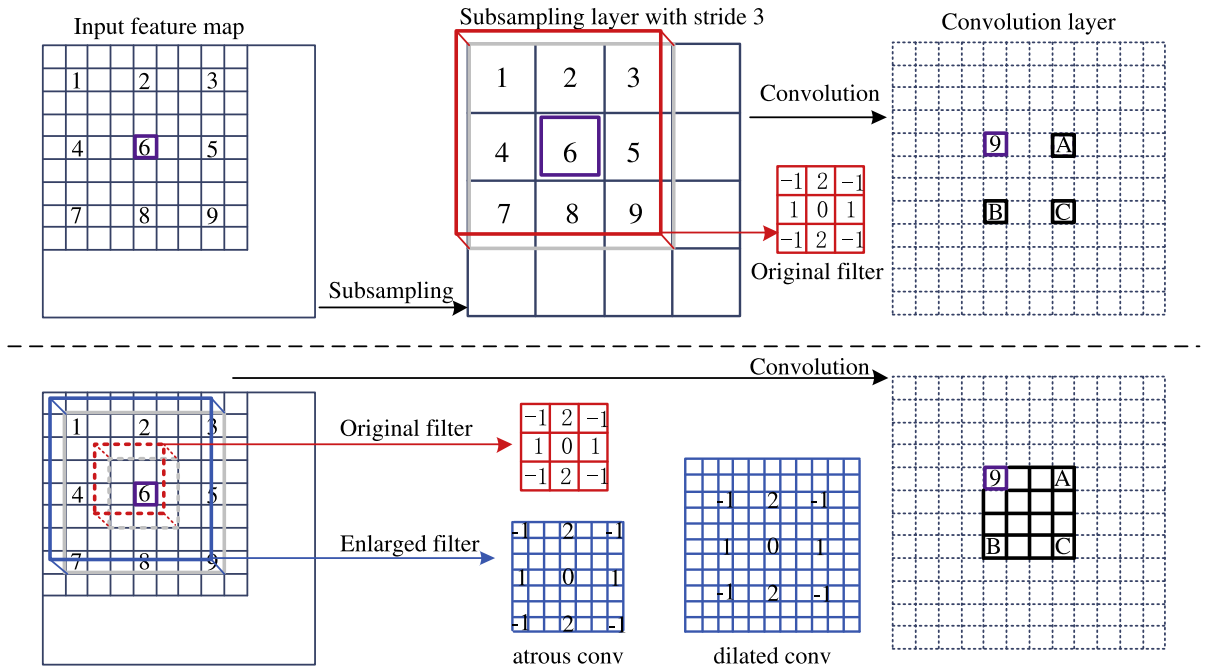

Fig. 7. Principle of *atrous* convolution. The top describe conventional pipeline of producing feature maps in FCN model. The bottom is the pipeline of producing finer-resolution feature maps by employing atrous convolution. The dash girds of the outputs indicate the resolution of the input. The purple boxes at different layers represent different level features of a specific pixel. We can see that the atrous pipeline produces finer-resolution output by removing the sub-sampling layer in the conventional pipeline. In order not to reduce the receptive fields, the convolutional mask is dilated accordingly (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

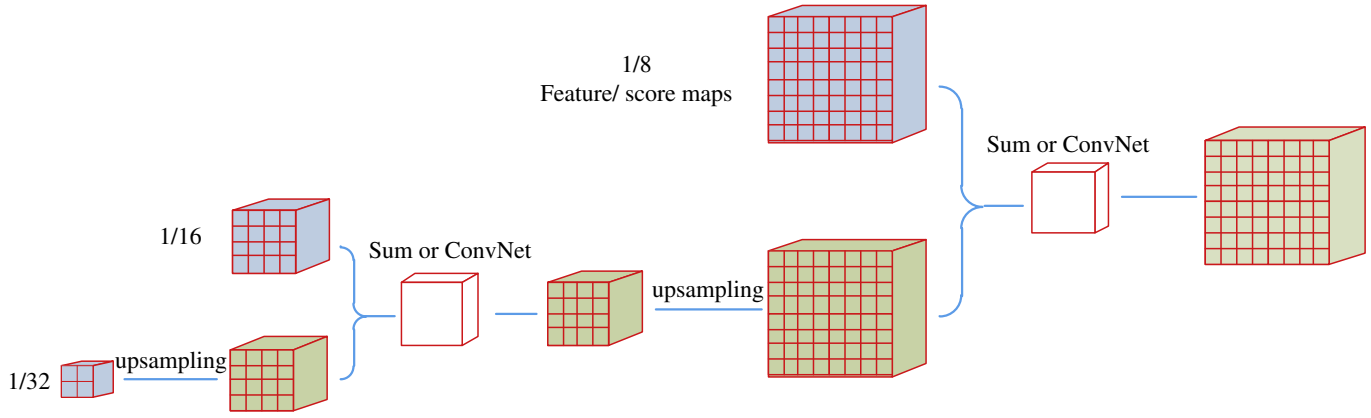

Fig. 8. Fusion of multi-resolution feature maps. The blue boxes depict three feature layers of typical FCNs. From left to right, different resolution feature maps are continuously fused by performing sum operation after size matching or feeding into a ConvNet (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

[19,26,118] added score(or score-like) operation to intermediate feature maps and then fused the multi-resolution scores to produce the final prediction. [119] argued that up-sampling feature maps is better than score maps. All these approaches fused the multi-resolution score or feature maps by a summing operation after up-sampling. By contrast, [18] employed a CNN based fusion strategy to continuously refine low-resolution features to finer resolution. The same idea has been adopted by several recent methods such as the RefineNet [120], RDFNet [121] and Label Refinement Network(LRN) [122]. Fig. 8 shows the fusion of multi-resolution feature maps.

Context exploiting: Although theoretically the receptive fields of many deep CNNs are close to or even larger than the entire input, it has been shown by Zhou et al. [123] that the valid receptive field of deep CNN is much smaller than the theoretical one, which indicates that higher-level contexts are not sufficiently exploited in many segmentation networks. To this end, Zhao et al. [25] further enlarged the receptive field of FCN to different sizes by pyramid pooling(multiple parallel pooling). Features extracted from multiple receptive fields were then fused to capture context in both low and high levels. A similar but more efficient method called atrous spatial pyramid pooling(ASPP) was reported in [116]. Recently, Lin et al. [124] presented a superpixel based receptive field and generated multiple receptive fields by adjusting the size of superpixels.

A number of methods [16,24,125] have incorporated CRF into UP_FCN architecture for further improvements. Chen et al. [16] defined a dense CRF over up-sampled predictions and obtained sharp boundaries using mean-field inference. Instead of using low-level pairwise potentials defined on color contrast, [24] computed pairwise pixel affinities using semantic boundaries produced from a trained CNN. Lin et al. [125] adopted a CNN based non-associative potential. In order to combine the strengths of the CNN and CRF

in one unified framework, Zheng et al. [126] converted several iterations of CRF inference to a recurrent neural network(RNN) and achieved the goal of training CNN and CRF jointly. Following this work, Arnab et al. [127] demonstrated that the CRF with higher order potentials can also be embedded into the FCN and produced significant improvements over the CRF-RNN. These CRF embedded FCNs, though effective, have much higher complexity than the basic FCN [128]. To solve this problem, the Deep Parsing Network(DPN) [129] approximated the inference of CRF as a convolutional network by casting the pairwise potential to a contextual classifier over unary scores. Given the discrete CRF based methods often resort to approximate algorithms for inference, [130,131] used Gaussian Conditional Random Fileds(GCRF) to pursue exact inference. While [130] used the similar idea of unrolling fixed inference steps to a deep network as [126] for training, [131] casted the inference as a simplified quadratic optimization module.

In recent years, the RNN have achieved great success in dependency modeling for time-sequenced data [132]. Built on this insight, several scene labeling approaches incorporated RNN into CNN to capture long-range spatial dependencies. The basic philosophy of RNN is to memorize some information generated in current prediction and then pass it to the next prediction. As a result, each prediction can be influenced by historical information based on this recurrent process. When applied to scene labeling, the concept of time sequence is extended to spatial sequence. For instance, Visin et al. [35] constructed four 1D spatial sequences by scanning the image horizontally and vertically in both directions. In [36], a 2D spatial sequence is formed by treating an image as a directed acyclic graph.

We have described three strategies for incorporating contextual information into FCN. All these strategies share certain advantages and disadvantages. Enlarging the receptive filed requires a simpler model and is easier to implement compared with using CRF/RNN. However, the structural correlations between pixels are implicitly and completely modeled by a hard-to-understand learning process. On the other hand, using CRF/RNN requires considerable efforts and tricks to train the model although it can exploit more accurate structure correlations.

Hard pixels aware learning: Most semantic segmentation datasets suffers from a class-imbalance problem, which results in performance differences in recognizing pixels of different classes. In addition, pixels in the same image are not equally recognizable. For example, as shown in [133], pixels located on the boundaries between objects are harder to recognize than those lie in the central part of objects. All these observations suggest that “hard” pixels [134] should be distinguished from “easy” pixels. To this end, Wu et al. [134] implicitly forced networks to focus on hard pixels during training. This is achieved by ignoring pixel samples if they have been “correctly recognized”, gauged by examining whether the predicted probability of the true label surpasses a threshold. At the beginning of training, pixels are equally treated because initial predictions for most pixels are of low confidences with respect to ground truths. As the training proceeds, easy pixels are gradually recognized and ignored. Instead of fully ignoring easy pixels, the Deep Layer Cascade [133] ignores easy pixels only in deep layers. A more advanced work [135] was introduced to adaptively evaluate the contributions of each pixel, where pixels with higher loss are weighted more than pixels with a lower loss. Note that these methods can be also regarded as cost-sensitive learning with hard(class)-specific weights. Although it is difficulty to design hard(class)-aware loss function, this kind of method can improve the performance of FCN without changing the model structure. Thus, it can be used as a generic strategy for improving semantic segmentation.

Table 3 summarizes learned features based methods described in this section. We can see from the table that recent methods mostly focus on making improvements over two basic FCN architectures: skip connected UP_FCN and atrous UP_FCN. These improvements include employing more powerful CNN models(e.g. ResNet), incorporating various CRF or higher level contexts and using cascaded methods.

From the discussions above, we can see that the FCN can be used as a general model for producing semantic segmentations. A current and common problem faced by FCN model is that it is easy to reach memory limits during training if using finer-resolution feature maps. That’s why the commonly used finer-resolution is typically less or equal to 1/8 of the input resolution. This implies a future study to further increase the resolution of the output.

5. Weakly and semi- supervised scene labeling methods

The methods mentioned so far require a large number of densely annotated(one pixel with one label) images for training or providing candidate labels for transferring in non-parametric framework. It is well known that such annotation task could be very tedious and time expensive. This has introduced a more challenging research field, namely the weakly supervised scene labeling, where the ground truth is given by only image level labels or bounding boxes.

5.1. Methods using image-level labels

5.1.1. Weakly supervised learning

Given the class set $\{1, 2, \ldots, L\}$, image-level label means that each training image I_i is labeled by an L dimensional indicator vector $y_i =< y_i^1, y_i^l, \ldots, y_i^L >$, where each element y_i^l takes the value 1 or 0 to indicate whether the image has the label l within it. In this context, learning class models would be very challenging due to the absence of direct correspondences between patterns and labels. In [28,138], the authors tried to recover these correspondences by using a variant expectation-maximization(EM) algorithm with the following iterative steps: correspondences estimation using fixed model parameters and model parameters optimization using estimated correspondences.

A recent trend to learn class model from weakly annotated data is to use multiple instance learning(MIL) techniques. The MIL usually refers to a generalized binary classification problem, in which the label is attached to a so called bag(a set of samples) but not to each sample or instance. The bag is positive only if there is a positive instance within it and otherwise negative. The MIL is usually formulated as a maximum margin problem. MI-SVM [139] and mi-SVM [139] are two classic methods in this framework. The key idea of the two methods can be simply described by an iteration process with following two steps: (1) fix labels for several or all instances in the positive bags and then learn the parameters of SVM; (2) reassign labels for several or all instances in the positive bags according to the SVM classifier learned in step 1. From this perspective, these two methods have the flavor of EM mentioned above. The difference between the two methods lies in the margin they used. The margin used in mi-SVM is same to the one used in standard SVM, which is defined by the distance between individual instance and the hyperplane. Thus, the labels of all instances in each positive bag are updated in the second step in mi-SVM. For a positive bag, if there is no positive instance after updating, the instance with the largest margin is labeled positive. In MI-SVM, the notion of instance-margin is extended to bag-margin, which is defined by the maximum instance-margin of a positive bag. In other words, only one instance per positive bag matters during learning in step 1.

Table 3
Summary of learned feature based methods.

Papers	Year	Description	Basic CNN
Alvarez et al. [23]	2012	Pixel-wise multi-scale CNN +assoc* CRF as post-processing	none
Farabet et al. [15]	2013	Pixel-wise multi-scale CNN +superpixels+assoc CRF as post-processing	none
Couprie et al. [20]	2013	ditto	none
Girshick et al. [17]	2014	Region-wise CNN	AlexNet [13]
Gupta et al. [21]	2014	ditto	ditto
Mostajabi et al. [110]	2015	Superpixel classification using multi-scale CNN features	VGG [14]
Sharma et al. [111]	2014	Superpixel classification using features learned from a two-stage process	Multi-scale CNN [15]
Pinheiro et al. [109]	2013	Shift-and-Stitch	none
Long et al. [19]	2015	Skip connected UP_FCN	VGG
Noh et al. [22]	2015	UP_FCN on regions	VGG
Badrinarayanan et al. [114]	2017	UP_FCN+decoder network	VGG
Hariharan et al. [118]	2015	Skip connected UP_FCN on regions	unknown
Zheng et al. [126]	2015	Skip connected UP_FCN+embedded CRF(as RNN)	VGG [19]
Liu et al. [129]	2015	Atrous UP_FCN+embedded CRF(as CNN)	VGG
Arnab et al. [127]	2016	Skip connected UP_FCN+embedded higher order CRF	VGG
Peng et al. [26]	2017	Skip connected UP_FCN+large receptive field	ResNet [136]
Lin et al. [125]	2016	Skip connected UP_FCN+non-assoc CRF using CNN based pair-pot@	VGG
Chen et al. [16]	2014	Atrous UP_FCN+ dense assoc CRF as post-processing	VGG
Chen et al. [116]	2016	Atrous UP_FCN + ASPP + dense assoc CRF as post-processing	ResNet
Bertasius et al. [24]	2016	Atrous UP_FCN+assoc CRF using pair-pot computed from learned boundaries	deeplab [16]
Zhao et al. [25]	2016	Atrous UP_FCN+pyramid pooling for considering high-level contexts	ResNet
Wu et al. [115]	2016	Atrous UP_FCN+better basic CNN	ResNet
Vemulapalli et al. [130]	2016	Atrous UP_FCN+embed Gaussian CRF	deeplab
Chandra [131]	2016	Atrous UP_FCN+Gaussian CRF	ditto
Visin et al. [35]	2016	CNN+RNN	VGG
Shuai et al. [36]	2016	ditto	VGG
Eigen et al. [18]	2015	Skip connected UP_FCN+CNN based feature fusion	AlexNet/VGG
Lin et al. [120]	2016	ditto	ResNet
Park et al. [121]	2017	ditto	ResNet
Islam et al. [122]	2017	Skip connected UP_FCN+CNN based feature fusion+deeply supervision	VGG
Li et al. [133]	2017	Atrous UP_FCN+hard-aware learning	Inception-ResNet [137]
Bulo et al. [135]	2017	Atrous UP_FCN+ASPP+hard-aware learning	ResNet

* associative. @ pairwise potential.

Vezhnevets and Buhmann [37] were among the first to extend binary MIL framework to multi-class weakly supervised segmentation. They learned a Random Forest inspired by the mi-SVM methods. Notice that the CRF model is able to improve the consistency of labeling. The same authors took their weakly supervised classifier as the unary model and presented a multi-image graphical model(MIM) [38], which is essentially a region based CRF but allows for links between similar superpixels of different images. Later, they [140] generalized this MIM model to allow for more types of pairwise potential between different images.

More recently, several works try to incorporate the CNN into the MIL for scene labeling task. For example, Pathak et al. [40] used a multi-class MIL loss, in which only the max scoring pixel of each class presented in an image matters, to fine-tune the pre-trained image-level classifiers of [14]. Papandreou et al. [39] developed an EM based algorithm for training CNN from weakly annotated images. It is well known that such heuristic algorithm is very sensitive to the initialization. To alleviate this problem, instead of directly estimating the labels of pixels in each iteration, Pathak et al. [141] computed an optimal probability distribution over the image labeling under the weak annotation constraints.

Inspired by the observation that different graphlets(small graph composed of several similar superpixels) from the same object often share similar appearance and spatial structures, Zhang et al. [41] learned a graphlet-wise classifier from weakly annotated images. This is achieved by first transforming graphlets of different sizes(number of supepixles) into equal-length feature vectors through a proposed manifold embedding algorithm. A multi-class SVM was then learned by treating each training image as a 1-sized graphlet, i.e., a single feature vector. Thereafter, they further incorporated this learned classifier into a hierarchical Bayesian Network(BN) to model the semantic associations between graphlets.

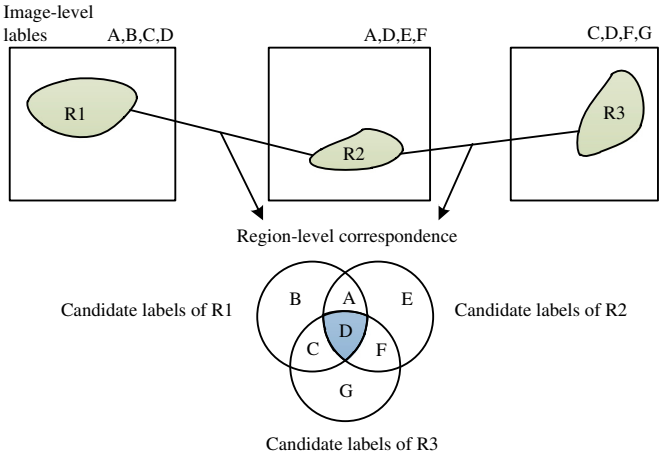

Fig. 9. Label propagation based on region-level correspondence. R1, R2, and R3 are three semantic regions of different images. The region-level correspondence indicates that two regions stand a good chance to share the same label. Thus, the label shared by these three regions can be obtained through inferring the label shared by their candidate labels, i.e. the image-level labels.

5.1.2. Label propagation

Besides learning parametric models, another type of method focuses on propagating the image-level labels to regions based on the assumption that visually similar regions with common image-level labels are likely to have the same label. For example, supposed that one label of an image only characterizes a single semantic region, Liu et al. [142] formulated the label-to-region problem as a problem of uncovering the region-level correspondence between all the image pairs. A simple illustration of their idea is shown in Fig. 9. However, it is difficult to directly partition an

image into several regions where each region is corresponding to a unique image-level label. To this end, they proposed a bi-layer sparse coding formulation to reconstruct a semantic region from a group of atomic patches. Thus, the reconstruction coefficients, which in some sense act as patch-level correspondence between image pairs, can be used to propagate the image-level labels to image patches.

Recently, Liu et al. [143] presented a graph based method for the task of label propagation. They first constructed a graph to model the contextual relationships, including consistency as well as the incongruity relationships, among patches extracted from the image collection. The consistency indicates two patches share the same label with high probability while the incongruity indicates opposite. With these two types of contextual constraints and the image labels serving as weak supervision, the image-level labels were then propagated to patches by solving a constrained optimization problem.

5.2. Methods using bounding box annotations

An early work on applying bounding box annotations(coarse object locations) to semantic segmentation was presented by Xia et al. [144]. Based on the fact that each bounding box usually contains only one object, this method firstly performed object-background segmentation for each bounding box. These binary semantic segmentations are then merged and post-processed to obtain the final result. One main limitation of this method lies in its assumption that each image is associated with several bounding boxes provided by either manually or well-designed object detectors. Nevertheless, it provides an intuition that the bounding box annotations can be easily transformed to dense annotations. Such estimated annotations can be then used to conduct fully supervised learning. Following this direction, Papandreou et al. [39] explored three methods(Bbox-Rect,Bbox-Seg and Bbox-EM-fixed) to estimate dense annotations and then trained a FCN based semantic segmentation model. The Bbox-Rect and Bbox-Seg produced the dense annotations as a pre-processing step, while the Bbox-EM-fixed continuously refined the estimated annotations during subsequent fully supervised training. Similar to the Bbox-EM-fixed, the BoxSup [145] applied the region proposal mechanism to the process of annotations refinement and produced a better performance.

From the discussions above, we can see that most weakly supervised scene labeling methods follow a standard two-stage pipeline: estimating dense annotations and then conducting fully supervised training. In addition, this two-stage process is usually iterated many times to refine the estimations and parametric models. Since it is very easy to obtain a good estimation using the bounding box, the estimating stage can be also taken as a pre-processing step in this case.

5.3. Semi-supervised methods

Learning semantic segmentation model using only image-level labels is very challenging, yet producing pixel-level labels requires tremendous efforts. To compromise, the semi-supervised methods [39,146–149] have been proposed to use both weakly and strong annotations.

The commonly used EM-like weakly supervised learning can hardly produce a satisfactory segmentation because the estimated truth is often full of wrong labels even after a number of iterations. Thus, Papandreou et al. [39] suggested to use a few strongly annotated images in each iteration of their EM-based algorithm and obtained significant improvements. To produce more accurate truth for weakly annotated images, the methods in [147,149] first conducted a fully supervised training using a small set of strong annotations before applying EM-like semi-supervised learning. In addition, multi-task learning, such as object recognition [147,149], image reconstruction [147] and image description [149], were also used in these two methods to further improve the accuracy of segmentation.

Instead of adopting the EM based algorithm, Hong et al. [146] developed a novel approach for exploiting weak annotations for semi-supervised semantic segmentation. In this approach, the weak annotations were only used to train an image classification network. The probabilities of classes obtained from the classification network were then back-propagated to produce class-specific activation maps. Taking the activation maps and the final feature maps produced from the classification network as inputs, a deconvolutional network was then learned to produce semantic segmentation with strong annotations. This method is close to fully supervised learning if regarding the training of multi-label image classification as a pre-training process. Its success demonstrated that the multi-label image classification network trained from a huge number of weakly annotated images can also yield discriminative features appropriate for semantic segmentation.

6. Public datasets for scene labeling

To inspire new methods and facilitate the comparison between different approaches, many public datasets for scene labeling have been proposed(see Table 4). Sowerby [150] is one of the earliest datasets which mainly includes a variety of urban and rural scenes with small sizes(96×64). MSRC [29] is another early dataset and consists of 591 photographs in 23 classes.

It is well known that a large number of training images are more useful for learning models of object categories. Motivated by the needs of training images, several large-scale datasets had been established in recent years. LabelMe [151] provides an web-based annotation tool through which a large number of annotated images can be collected from a large population of web-users. It adopts bounding polygon annotations and users are free to label as many objects in the chosen images. PASCAL is derived from a well-known competition: the PASCAL Visual Object Classes (VOC) challenge [113]. It provides a large-scale dataset with high quality annotation, including 2913 images with pixel-level labeling, and has been enlarging continuously. Objects in the PASCAL are divided into 20 classes and organized into four major categories: vehicles, household, animals and other(person). ImageNet [152] is very similar to PASCAL but contains more than an order of magnitude in number of object classes and images [153]. The huge amount of annotated images(nearly 15 million) are collected from the internet and labeled by Amazons Mechanical Turk. The recently released COCO dataset [154] contains more than 328,000 images with 91 object classes. The most notable characteristic of this dataset is that each category is associated with thousands(27,000 on average) of instances. Hence, it will be very powerful for learning discriminative models of classes. With these large-scale datasets, many derived datasets with smaller number of images were also presented. For example, SIFT-flow [93] contains 2688 images obtained from LabelMe system. Similarly, Stanford Background Dataset(SBD) [32] collected 715 images from LabelMe, PASCAL and MSRC.

More recently, Cordts et al. [155] developed a dataset called Cityscapes, which consists of a large amount of street scenes and mainly designed for urban scene understanding. Cityscapes contains 5000 densely labeled images and 20,000 coarse annotated images that can be used for fully supervised methods and weakly supervised methods respectively. Another very recent dataset called ADE20K [156] provides a generic and more challenging benchmark for semantic segmentation. It contains more than 20,000 images of various scene types and considers 150 classes for dense labeling.

Table 4
Popular datasets for 2D/3D semantic segmentation.

Dataset	Scene type	Number		Annotation	Website
		images/scenes	classes		
MSRC [29]	outdoor	591	23	pixel-level labeling	https://www.microsoft.com/en-us/research/project/image-understanding/
LabelMe [151]	hybird	187,240	unlimited	bounding polygon	http://labelme2.csail.mit.edu/Release3.0/
PASCAL VOC [113]	hybird	21,738	20	hybrid	http://host.robots.ox.ac.uk/
ImageNet [152]	hybird	1,431,167	1000	ditto	http://image-net.org/download
MSC COCO [154]	hybird	328,000	91	object masks	http://mscoco.org/dataset/
SIFT-FLOW [93]	outdoor	2,688	33	pixel-level labeling	http://people.csail.mit.edu/celiu/LabelTransfer/code.html
SBD [32]	outdoor	715	8	pixel-level labeling	http://dags.stanford.edu/projects/scenedataset.html
CityScapes [155]	street	25,000	19	hybrid	https://www.cityscapes-dataset.com/
ADE20K [156]	hybird	20,000+	150	pixel-level labeling	http://groups.csail.mit.edu/vision/datasets/ADE20K/
NYU Depth V2 [157]	indoor	1,449	894	pixel-level labeling	http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html
Washington RGB-D [158]	indoor	250,000	300	segmented object	http://rgbd-dataset.cs.washington.edu/
Cornell RGB-D [159]	indoor	52(point cloud)	undetailed	point level labeling	http://pr.cs.cornell.edu/sceneunderstanding/
Sun RGB-D [160]	indoor	10,335	undetailed	hybrid	http://rgbd.cs.princeton.edu/
ScanNet [161]	indoor	1513	20	voxel level labeling	http://www.scan-net.org/
VMR-Okaland-V2 [162]	outdoor	36(3D blocks)	7	point level labeling	http://www.cs.cmu.edu/%CB%9Cvmr/datasets/

With the advent of consumer RGB-D sensors, many RGB-D datasets have been released in succession. NYU Depth dataset [157] is the first and the most popular one. It consists of densely labeled 1449 RGB-D images captured from 464 diverse indoor scenes. Washington RGB-D Object dataset [158] contains visual and depth images of 300 common daily objects with which personal robots are supposed to interact. These objects are organized into a hierarchical category structure that contains 51 classes and each object is associated with different instances taken from multiviews. Apart from these isolated objects, the dataset also includes 22 annotated video sequences of natural scenes.

Turn to 3D point cloud labeling, Cornell RGB-D dataset [159] provided 52 labeled 3d point clouds stitched from color and depth images. The Sun RGB-D dataset [160] contains 10,335 RGB-D images taken from four types of RGB-D sensors. All these images are annotated with polygons in 2D and bounding boxes in 3D. A very recent dataset called ScanNet [161] contains 2.5M views in 1513 scenes annotated with dense 3D voxel labeling. VMR-Okaland-V2 [162] is one of the most popular urban 3D datasets. It consists of 36 blocks collected from a terrestrial laser scanner. These blocks contain nearly 3 million 3D points labeled by seven common outdoor objects: wire, pole, leaves, tree trunk, building and vehicle.

There also exist many datasets designed for specific applications of semantic segmentation. For example, Yamaguchi et al. [163] collected a dataset called Fashionista for clothes parsing, aiming for segmenting garment pieces (shoes, socks, etc.). The human parsing dataset [164] contains labels of not only clothes but also body parts(arm, leg, etc.). We do not intend to list more of such application-specific datasets because this review focuses on generic semantic segmentation for natural scene understanding.

7. Evaluation and comparison

7.1. Evaluation of scene labeling methods

It is well-known that image segmentation is an ill-defined problem, how to evaluate the property of an algorithm has always been a critical issue. For the case of semantic segmentation, evaluation is usually defined on the comparison between the algorithm's output and the ground-truth. Examples including pixel accuracy, class accuracy, Jaccard index, precision and recall.

Pixel Accuracy(P-ACC) and Class-average Accuracy(C-ACC)

The P-ACC is the most widely used evaluation in scene labeling. It defines the accuracy of pixel-wise prediction. Let $\hat{l}_i$ and l_i be the predicted label and ground truth of pixel i, the P-ACC can be written as

$$P-ACC=\frac{\sum_i^N \delta(\hat{l}_i, l_i)}{N}, \tag{11}$$

where N is the total number of pixels of all test images and $\delta(x, y)$ is an indicator function written by

$$\delta(x,y)=\begin{cases}1 & if\ x=y\\ 0 & otherwise\end{cases}. \tag{12}$$

For each class k, the accuracy of prediction is defined as

$$ACC(k)=\frac{\sum_i^N (\delta(l_i,k)\&\delta(\hat{l}_i,k))}{\sum_i^N \delta(l_i,k)}, \tag{13}$$

where the denominator represents the total number of pixels with ground truth label k, the numerator is the statistics of correctly predicted pixels of class k. Given the total number of classes L, the class average accuracy is given by

$$C-ACC=\frac{\sum_k^L ACC(k)}{L}. \tag{14}$$

Jaccard Index(JI)

Jaccard Index(JI) is a well-known statistic used for measuring the similarity of two sets. As shown in (15), JI is defined as the size of the intersection divided by the size of the union of two sets. Hence, it is also called as intersection over union(IU or IOU) in some literatures [18,19].

Table 5
Comparison of CRF model and non-CRF model.

Models	P-ACC (%)	mean JI(IOU) (%)
Region-wise classifier [33]	82.85	63.75
Region based CRF [33]	84.18	65.00

Table 6
Comparison of pixel based model and region based model.

Models	P-ACC (%)	mean JI(IOU) (%)
Pixel-wise predictions of FCN [19]	90.97	60.14
Region-wise predictions of FCN [19]	91.11	60.73

$$JI = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{15}$$

Adopt the same notes mentioned above, the JI for scene labeling of a specific class k can be written as

$$JI(k) = \frac{\sum_i^N (\delta(l_i, k) \& \delta(\hat{l}_i, k))}{\sum_i^N (\delta(l_i, k) | \delta(\hat{l}_i, k))}, \tag{16}$$

As in the case of class average accuracy, the mean JI can be obtained by averaging JI over all classes.

Precision and Recall(PR)

PR is an evaluation frequently used in information retrieval. In the context of classification, the precision(see (17)) of a class k is defined as the proportion of elements that actually belong to k among all elements predicted to be k. The recall of k is identical to class accuracy defined in (13).

$$precision(k) = \frac{\sum_i^N (\delta(l_i, k) \& \delta(\hat{l}_i, k))}{\sum_i^N \delta(\hat{l}_i, k)} \tag{17}$$

7.2. Comparison of some scene labeling methods

7.2.1. Comparison of representative models

Besides the model structure, there are many other differences between scene labeling methods, such as features, superpixel algorithms and classify strategies. Thus, it is difficult to determine which model is better by directly comparing those published methods. To this end, we choose several methods and change the model structures they used to check out the advantages and drawbacks of different models.

Non-CRF v.s. CRF

To evaluate the performance of commonly used CRF model. We implemented the CRF model described in [33] using the region-wise classifier code provided by the author and the graph cut code of Boykov [55]. Then, we evaluated the original classifier and its CRF counterpart on the Stanford Background dataset [32]. The qualitative comparison of the two models is shown in Fig. 10. We can see that non-CRF models easily produce inconsistency(white rectangle) in labeling. On the other hand, associative CRF could lead to over-smooth(black rectangle) sometimes. Nevertheless, the CRF model performs better than non-CRF model in total according to the quantitative comparison shown in Table 5.

Pixel v.s. superpixel

In the second experiment, we conduct a comparison between pixel based model and region based model. Toward this goal, we chose 577 images from PASCAL VOC2012 validation set and produced pixel-wise predictions for each of them. The pixel-wise predictions are obtained by using the FCN-8s caffe (Convolution Architecture For Feature Extraction) model provided by the author of FCN [19]. Then, we implemented a region-based model by assigning pixels within a superpixel with the same label if there is a label accounting for more than 70 percent. With this simply modification, we found a slight improvement in the boundary details of output. However, the region based model may also produce worse results due to the inaccuracy of superpixel algorithms. Fig. 11 shows both of the two situations. But in general, we can see from Table 6 that the region based model achieve a minor improvement on the quantitative performance.

7.2.2. Comprehensive comparison of fully supervised methods

With the available of the commonly used datasets and evaluation methods, it is very easy to conduct a direct comparison between scene labeling methods based on the results they reported. Thus, we choose five typical indoor scene labeling methods, in which hand-engineered feature based method, CRF based method and learned feature based method all included, to give further insight into the advantages and disadvantages of different models. The results of comparisons is shown in Table 7. Since the results reported may be evaluated on different datasets, we compare them through choosing a reference method (denoted by "#" in table) that experimented with several datasets or tasks. For instance, Gupta et al. [73] validated the performance of their method on two tasks: the 4-class task and the 40-class task.

In the Table 7. we find that [73], which use hand-engineered features and non-CRF model, performs better than both a learned feature based method [20] and a CRF based method [157]. This is reasonable since [73] takes lots of geometric information from RGB-D images into account. In addition, the authors adopted a learned based over-segmentation algorithm [92] that can produce superpixels with accurate boundaries. In other words, the success of [73] is mainly resulted from the careful designed feature representations as well as the high accuracy superpixel algorithm. However, its performance is still far behind those achieved by FCN models [19,120].

In order to demonstrate the state-of-the-arts of semantic segmentation, Table 8 presents top 10 performances achieved on PASCAL VOC 2012 segmentation task. The numbers in the brackets represent the rankings(up to Nov. 2017) in performance. Several rankings are missing because some works listed in the leader board[1] do not provide corresponding technical reports. We can see that these methods have three characters in common: using a very deep CNN(ResNet, Inception-ResNet) to produce feature maps, using additional data(e.g. COCO, JFT-300M [165]) for pre-training and using FCN based model to produce dense predictions. It is worth noting that most of them achieve the state-of-the-art even without using CRF. This is mainly due to the strength of the deep CNN on learning discriminative features. In addition, although the CRF is waived, the contextual information is also encoded by other strategies(e.g. multi-scale).

7.2.3. Weakly supervised methods v.s. fully supervised methods

We also conduct a comparison between weakly supervised methods and fully supervised methods to evaluate the performance of weakly supervised scene labeling. As shown in the Table 9, a simple weakly supervised learning strategy [40], in which only the maximum scoring pixel of each class matters, can only achieving 25.7%(mIoU) on the PASAL-VOC 2012 test. By taking all of pixels into account and using an EM-based training algorithm, [39] achieved much higher performance. However, it still significantly lags behind the performance of fully supervised methods. On the other hand, weakly supervised methods using bounding box annotations can produce competitive results. Moreover, the bounding boxes are far easier to collect than pixel-level labels.

[1] http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6.

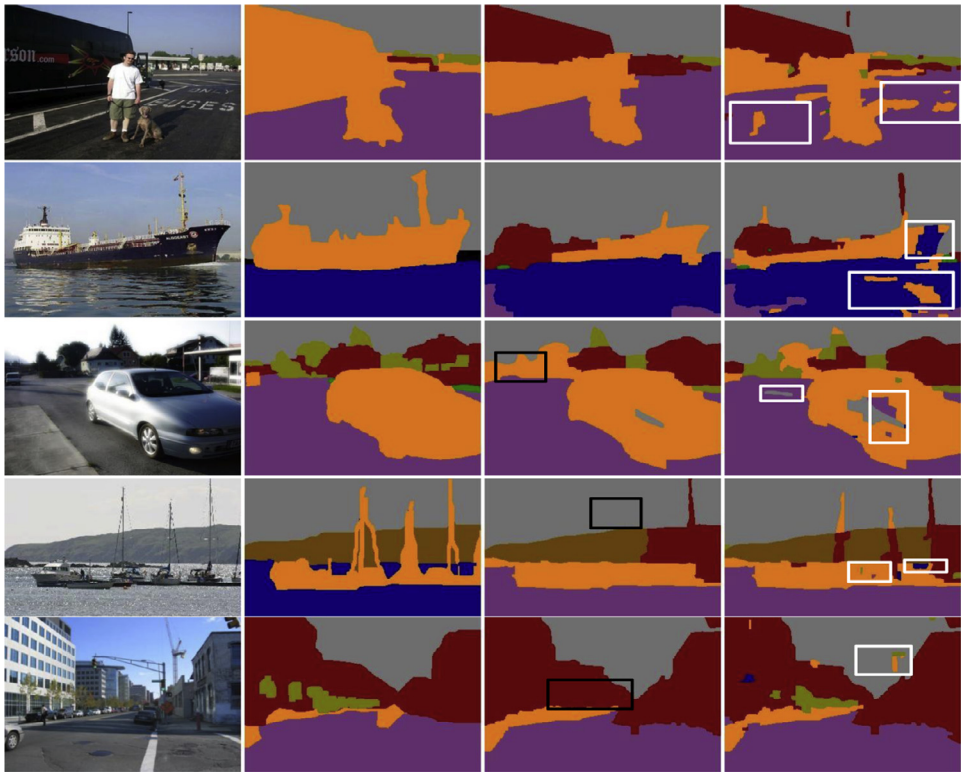

Fig. 10. Qualitative comparison of non-CRF model and CRF model. The first and second columns are the input images and ground truths, respectively. The third and fourth columns are the labeling produced by CRF model and non-CRF model, respectively.

Table 7
Comparison of indoor scene labeling methods using NYU datasets.

Methods	Descriptions	4-class P-ACC	40-class P-ACC
Silberman et al. 2012 [157]	hand-engineered features+region-based non-assoc* CRF	58.6%	
Couprie et al. 2013 [20]	Scan image with typical CNN+region-based assoc CRF	64.5%	
Gupta et al. 2013 [73]#	hierarchical oversegmentation+hand-engineered features+region-wise classifier	78.1%	58.3%
Long et al. 2013 [19]	FCN+multi-scale features		65.4%
Lin et al. 2016 [120]	FCN+cascaded refinement		73.6%

* associative.

Table 8
State-of-the-arts on PASCAL VOC 2012 test dataset.

Methods	ResNet	COCO	Improvement over basic FCN		Others	mIoU
			atrous conv.	skip connection		
DeepLabv3-JFT(1) [166]	yes	yes	yes	no	pre-trained on JFT-300M	86.9
DIS(2) [147]	yes	yes	n/a	n/a	multi-task learning	86.8
IDW-CNN(4) [149]	yes	yes	yes	no	multi-task learning	86.3
DeepLabv3(5) [166]	yes	yes	yes	no	n/a	85.7
PSPNet(6) [25]	yes	yes	yes	no	n/a	85.4
ResNet-38_COCO(7) [115]	yes	yes	yes	no	n/a	84.9
Multipath-RefineNet(8) [120]	yes	yes	no	yes	n/a	84.2
Large Kernel Matters(9) [26]	yes	yes	no	yes	n/a	83.6

Table 9
Comparison between weakly supervised methods and fully supervised methods.

Methods	Annotations	Descriptions	mIoU(VOC2012) (%)
Pathak et al. 2015 [40]	Image-level labels	MIL+Fully CNN	25.66
Papandreou et al. 2015 [39]	ditto	EM based training +Deep CNN	39.6
Dai et al. 2015 [145]	Bounding box	bounding box to pixel-level labeling+FCN	64.6
Chen et al. 2015 [16]	Pixel-level labeling	FCN+CRF	66.4
Zhao et al. 2017 [25]	Pixel-level labeling	FCN+Multi-scale	85.4

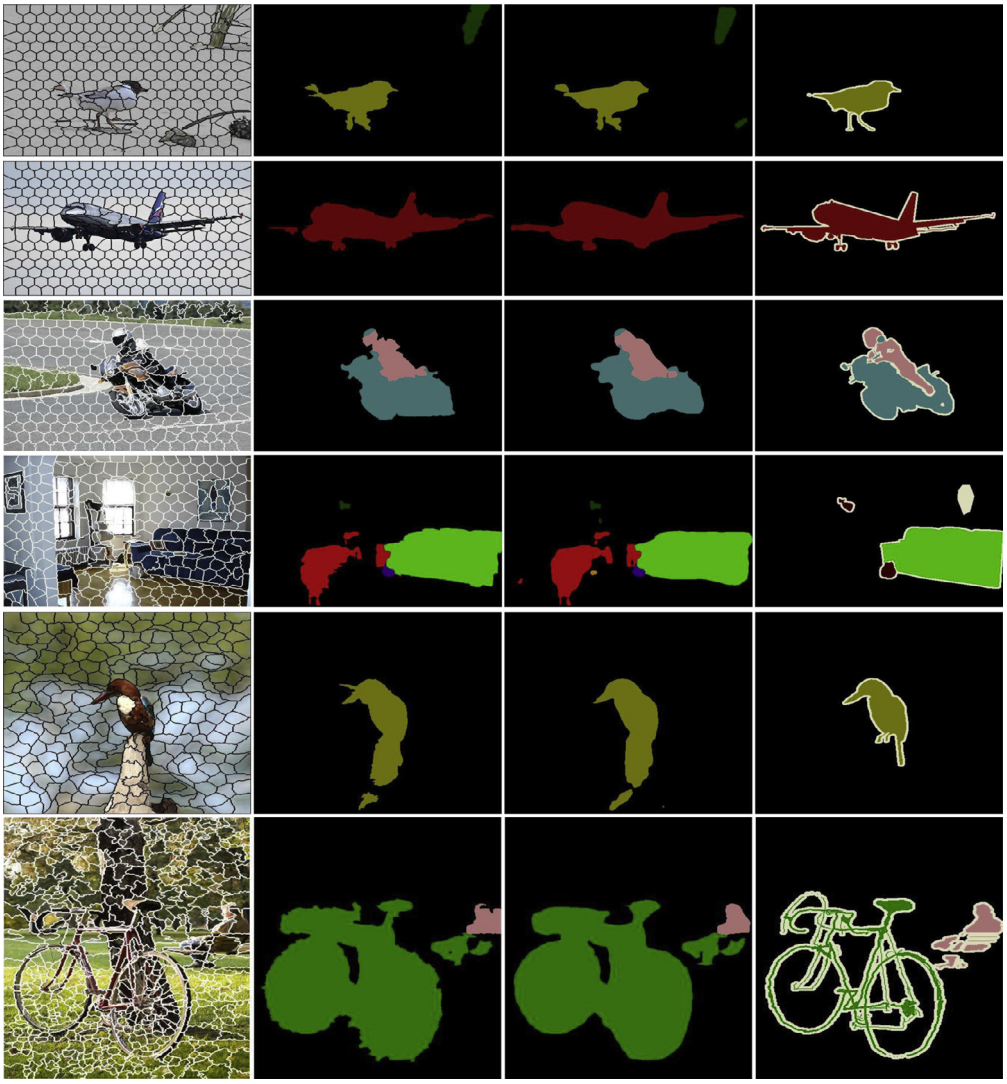

Fig. 11. The comparison of pixel based model and region based model. The first column are superpixels generated from a variant of SLIC algorithm (http://www.peterkovesi.com/projects/segmentation/). The number of superpixels is set to 300. The middle two columns are the predictions of region based model and pixel based model, respectively. The last column are the ground truth. We can see from the first four rows that the region based model can give much more boundary details, such as beak, airplane wheels and sofa edges. The last two rows show that the inaccuracy of superpixels may result in worse performance in the region based model.

Based on these, we argue that box annotations will receive more attention in the future.

8. Conclusion and future directions

In this paper, we critically reviewed existing scene labeling methods. The simplest way to solve this challenge issue is to perform pixel-wise classification using hand-engineered features, such as color and texture. However, this line of work easily produces inconsistency in results. Performing superpixel-wise classification can somehow mitigate this problem. Through assigning the same label to a superpixel, the spatial consistency is ensured, but only limited to a local range. A more efficient way to ensure consistency is to exploit contextual information of the scene using the concept of MRF or CRF. For the methods based on hand-engineered features, the main challenge in performance improvement is how to design more feature representations. However, such heuristic designs could be very difficult. An alternative way is to use the learned features that have been proven very useful in computer vision.

Although methods reviewed in this paper have achieved great success on semantic segmentation, there is still a long way to the goal of enabling computers or robots understanding theirs surrounding scene as we human do. According to the recent developments in both semantic segmentation and other related vision communities, here we suggest several future directions.

Better understanding of CNN: Despite the success achieved by CNNs, there is a limited understanding why they work so well and how to improve them. Thus, the CNNs have long been used as “black boxes” and the development of better models is usually resort to trial-and-error. It is well known that the cost of every trial is huge since the training of modern CNNs usually requires a large amount of time. For example, it takes five days for Long et al. [19] to fine-tune the VGG model to the fcn-8s model on a single GPU(NVIDIA Tesla K40). Hence, it is necessary to have deeper insights into the inner working of the CNNs to provide scientific criteria for designing better models. Although there have been a number of works [167–169] trying to visualize the learned features to gain intuition about the CNN and have achieved significant progress, there are still a lot of rooms for better understanding of CNN. For example, how the number of feature channels in each

layer influences the performance of CNN. This may be very useful for FCN based semantic segmentation in view of its huge requirements for memory.

3D CNNs for point cloud labeling: In recent years, reports on applying CNNs to 3D vision tasks have increased significantly as well. Two common CNN models used in these reports including Volumetric CNN [170–172] and Multi-view CNN [172–174]. In the Volumetric CNN, the 3D data are converted into a 3D binary voxel, which is then fed into a 3D convolution network for feature extraction. The Multi-view CNN utilizes multiple 2D views of the 3D scenes and feeds them into 2D CNNs. These CNN based 3D vision algorithms mostly focus on object recognition/detection. A very recent work [175] on point cloud labeling developed an architecture that directly consumes points. Following this direction, it can be expected more efforts of applying the CNN to 3D semantic segmentation, i.e. 3D point cloud labeling.

Weakly supervised semantic segmentation: It has been shown that the performance of semantic segmentation can be improved by including additional valid training images. However, it requires tremendous efforts to produce pixel-level annotations. Although weakly supervised methods can drastically reduce the annotation cost, their performances are still far from satisfactory. Thus, more novel ideas, such as collecting images [176] or videos [177] from the Internet as supervision, are needed to boost the development of weakly supervised semantic segmentation.

Evaluated on more complex datasets: Most recent methods tested their models on two common datasets: PASCAL [113] and Cityscapesv [155]. Both of the two datasets use very simple samples and make it easy to arrive at the bottleneck. For example, most annotations in PASCAL have a large background and make the semantic segmentation almost equivalent to object segmentation. Compared to PASCAL, Cityscapes provides annotations with more details, however, it contains only urban street scenes and is not appropriate for generic semantic segmentation. With the newly released ADE20K [156], which considers 150 semantic classes and contains densely labeled 20,000 scene images of various types, testing with ADE20K-like complex datasets will become a mainstream in semantic segmentation.

Aiming at applications: In the past several years, great improvements have been made on semantic segmentation by using deep learning. Consequently, there will be greater interest in applying semantic segmentation models to practical applications, such as service or industrial robots, unmanned vehicle, medical image segmentation [178], human parsing [164,179] and so on. Among them, the application to autonomous mobile systems could be in a huge demand. In this context, how to simultaneously guarantee the accuracy and computational efficiency of the model could be the most challenge issue.

At last, it is worth noting that many image segmentation methods are inspired by the theories and methods studied in cognitive science. For instance, most clustering based methods are inspired by the theory of gestalt [180]. Similarly, the CNN is motivated from the research on the visual cortex of cats. We predict that the next quantum leap in segmentation community will be closely related to theories in cognitive science, neuroscience or their closely related scientific fields.

Acknowledgment

This work has been supported by National Natural Science Foundation of China (Grant No. 61573135), National Key Technology Support Program(Grant No. 2015BAF11B01), National Key Scientific Instrument and Equipment Development Project of China (Grant No. 2013YQ140517), Hunan Key Laboratory of Intelligent Robot Technology in Electronic Manufacturing (Grant No.2018001), Science and Technology Plan Project of Shenzhen City (JCYJ20170306141557198), Key Project of Science and Technology Plan of Guangdong Province(Gran No. 2013B011301014), Open foundation of State Key Laboratory of Robotics of China (Grant No. 2013O09), and the National Institutes of Health of the United States (Grant No.R01CA165255 and R21CA172864).

References

[1] Y.-i. Ohta, T. Kanade, T. Sakai, An analysis system for scenes containing objects with substructures, in: Proceedings of the Fourth International Joint Conference on Pattern Recognitions, 1978, pp. 752–754.
[2] S. Edelman, T. Poggio, Integrating visual cues for object segmentation and recognition, Opt. News 15 (5) (1989) 8–13.
[3] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001, pp. 511–518.
[4] J. Carreira, C. Sminchisescu, Cpmc: automatic object segmentation using constrained parametric min-cuts, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1312–1328.
[5] P. Carbonetto, N.D. Freitas, K. Barnard, A statistical model for general contextual object recognition, in: Proceedings of the European Conference on Computer Vision, 2004, pp. 350–362.
[6] N. Silberman, R. Fergus, Indoor scene segmentation using a structured light sensor, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2011, pp. 601–608.
[7] S. Gupta, P. Arbelaez, J. Malik, Perceptual organization and recognition of indoor scenes from RGB-D images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 564–571.
[8] S.H. Khan, M. Bennamoun, F. Sohel, R. Togneri, Geometry driven semantic labeling of indoor scenes, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 679–694.
[9] D.G. Lowe, Object recognition from local scale-invariant features, in: Proceedings of the IEEE International Conference on Computer Vision, 1999, pp. 1150–1157.
[10] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.
[11] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[12] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in neural information processing systems, 2012, pp. 1097–1105.
[14] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556(2014).
[15] 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.
[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected CRFs, arXiv:1412.7062(2014).
[17] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[18] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2650–2658.
[19] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[20] C. Couprie, C. Farabet, L. Najman, Y. LeCun, Indoor semantic segmentation using depth information, arXiv:1301.3572(2013).
[21] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-D images for object detection and segmentation, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 345–360.
[22] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[23] J.M. Alvarez, Y. LeCun, T. Gevers, A.M. Lopez, Semantic road segmentation via multi-scale ensembles of learned features, in: Proceedings of the European Conference on Computer Vision, 2012, pp. 586–595.
[24] G. Bertasius, J. Shi, L. Torresani, Semantic segmentation with boundary neural fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3602–3610.
[25] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, arXiv:1612.01105(2016).
[26] C. Peng, X. Zhang, G. Yu, G. Luo, J. Sun, Large kernel matters–improve semantic segmentation by global convolutional network, arXiv:1703.02719(2017).
[27] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[28] J. Verbeek, B. Triggs, Region classification with Markov field aspect models, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[29] J. Shotton, J. Winn, C. Rother, A. Criminisi, Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation, in: Proceedings of the European conference on computer vision, 2006, pp. 1–15.

[30] P. Kohli, P.H. Torr, Robust higher order potentials for enforcing label consistency, Int. J. Comput. Vis. 82 (3) (2009) 302–324.

[31] L. Ladick, C. Russell, P. Kohli, P.H. Torr, Associative hierarchical CRFs for object class image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 739–746.

[32] S. Gould, R. Fulton, D. Koller, Decomposing a scene into geometric and semantically consistent regions, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 1–8.

[33] X. Ren, L. Bo, D. Fox, RGB-D scene labeling: Features and algorithms, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2759–2766.

[34] X. He, R.S. Zemel, M.Á. Carreira-Perpiñán, Multiscale conditional random fields for image labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2004, pp. II–695.

[35] F. Visin, M. Ciccone, A. Romero, K. Kastner, K. Cho, Y. Bengio, M. Matteucci, A. Courville, Reseg: A recurrent neural network-based model for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 41–48.

[36] B. Shuai, Z. Zuo, B. Wang, G. Wang, Dag-recurrent neural networks for scene labeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3620–3629.

[37] A. Vezhnevets, J.M. Buhmann, Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3249–3256.

[38] A. Vezhnevets, V. Ferrari, J.M. Buhmann, Weakly supervised semantic segmentation with a multi-image model, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 643–650.

[39] G. Papandreou, L.-C. Chen, K.P. Murphy, A.L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1742–1750.

[40] D. Pathak, E. Shelhamer, J. Long, T. Darrell, Fully convolutional multi-class multiple instance learning, arXiv preprint arXiv:1412.7144 (2014).

[41] L. Zhang, Y. Yang, Y. Gao, Y. Yu, C. Wang, X. Li, A probabilistic associative model for segmenting weakly supervised images, IEEE Trans. Image Process. 23 (9) (2014) 4150–4159.

[42] J. Amores, Multiple instance classification: Review, taxonomy and comparative study, Artif. Intell. 201 (2013) 81–105.

[43] H. Zhu, F. Meng, J. Cai, S. Lu, Beyond pixels: a comprehensive survey from bottom-up to semantic image segmentation and cosegmentation, J. Vis. Commun. Image Represent. 34 (2016) 12–27.

[44] D.K. Prasad, Survey of the problem of object detection in real images, Int. J. Image Process. 6 (6) (2012) 441–466.

[45] X. Ren, J. Malik, Learning a classification model for segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2003, pp. 10–17.

[46] D. Comaniciu, P. Meer, Mean shift: a robust approach toward feature space analysis, IEEE Trans. Pattern Anal. Mach. Intell. 24 (5) (2002) 603–619.

[47] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.

[48] B. Peng, L. Zhang, D. Zhang, A survey of graph theoretical approaches to image segmentation, Pattern Recognit. 46 (3) (2013) 1020–1038.

[49] A. Levinshtein, A. Stere, K.N. Kutulakos, D.J. Fleet, S.J. Dickinson, K. Siddiqi, Turbopixels: Fast superpixels using geometric flows, IEEE Trans. Pattern Anal. Mach. Intell. 31 (12) (2009) 2290–2297.

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

[51] S. Kumar, M. Hebert, Discriminative random fields, Int. J. Comput. Vis. 68 (2) (2006) 179–201.

[52] S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. (6) (1984) 721–741.

[53] J. Besag, On the statistical analysis of dirty pictures, J. R. Stat. Soc. Ser. B (Methodol.) (1986) 259–302.

[54] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the Eighteenth International Conference on Machine Learning, ICML, 2001, pp. 282–289.

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

[56] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simmulated annealing, Science 220 (4598) (1983) 671–680.

[57] V. Černỳ, Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm, J. Optim. Theory Appl. 45 (1) (1985) 41–51.

[58] Y.Y. Boykov, M.-P. Jolly, Interactive graph cuts for optimal boundary & region segmentation of objects in nd images, in: Proceedings of the IEEE International Conference on Computer Vision, 2001, pp. 105–112.

[59] 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.

[60] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient belief propagation for early vision, Int. J. Comput. Vis. 70 (1) (2006) 41–54.

[61] J. Pearl, Reverend bayes on inference engines: a distributed hierarchical approach, in: Proceedings of the Second National Conference on Artificial Intelligence, 1982, pp. 133–136.

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

[63] J.H. Kappes, B. Andres, F.A. Hamprecht, C. Schnörr, S. Nowozin, D. Batra, S. Kim, B.X. Kausler, T. Kröger, J. Lellmann, N. Komodakis, B. Savchynskyy, C. Rother, A comparative study of modern inference techniques for structured discrete energy minimization problems, Int. J. Comput. Vis. 115 (2) (2015) 155–184.

[64] A.E. Johnson, Spin-images: a representation for 3-D surface matching, Carnegie Mellon University, 1997 Ph.D. thesis.

[65] J. Li, N.M. Allinson, A comprehensive review of current local features for computer vision, Neurocomputing 71 (10) (2008) 1771–1787.

[66] S. Konishi, A.L. Yuille, Statistical cues for domain specific image segmentation with performance analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2000, pp. 125–132.

[67] F. Schroff, A. Criminisi, A. Zisserman, Single-histogram class models for image segmentation, in: Computer Vision, Graphics and Image Processing, Springer, 2006, pp. 82–93.

[68] Z. Tu, X. Bai, Auto-context and its application to high-level vision tasks and 3d brain image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 32 (10) (2010) 1744–1757.

[69] M. Fink, P. Perona, Mutual boosting for contextual inference, in: Proceedings of the International Conference on Neural Information Processing Systems, 2003, pp. 1515–1522.

[70] P. Kontschieder, S.R. Bulo, H. Bischof, M. Pelillo, Structured class-labels in random forests for semantic image labelling, in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 2190–2197.

[71] L. Yang, P. Meer, D.J. Foran, Multiple class segmentation using a unified framework over mean-shift patches, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[72] J. Carreira, R. Caseiro, J. Batista, C. Sminchisescu, Semantic segmentation with second-order pooling, Comput. Vis.–ECCV 2012 (2012) 430–443.

[73] S. Gupta, P. Arbeláez, R. Girshick, J. Malik, Indoor scene understanding with RGB-D images: bottom-up segmentation, object detection and semantic segmentation, Int. J. Comput. Vis. 112 (2) (2015) 133–149.

[74] C. Pantofaru, C. Schmid, M. Hebert, Object recognition by integrating multiple image segmentations, in: Proceedings of the European Conference on Computer Vision, 2008, pp. 481–494.

[75] B. Fulkerson, A. Vedaldi, S. Soatto, Class segmentation and object localization with superpixel neighborhoods, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 670–677.

[76] C. Zhang, L. Wang, R. Yang, Semantic segmentation of urban scenes using dense depth maps, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 708–721.

[77] C. Cadena, J. Košecka, Semantic parsing for priming object detection in RGB-D scenes, in: Proceedings of the Third Workshop on Semantic Perception, Mapping and Exploration, Citeseer, 2013.

[78] B. Taskar, V. Chatalbashev, D. Koller, Learning associative Markov networks, in: Proceedings of the Twenty-First International Conference on Machine Learning, 2004, p. 102.

[79] X. He, R.S. Zemel, D. Ray, Learning and incorporating top-down cues in image segmentation, in: European Conference on Computer Vision, Springer, 2006, pp. 338–351.

[80] S. Gould, J. Rodgers, D. Cohen, G. Elidan, D. Koller, Multi-class segmentation with relative location prior, Int. J. Comput. Vis. 80 (3) (2008) 300–316.

[81] D. Batra, R. Sukthankar, T. Chen, Learning class-specific affinities for image labelling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[82] A.C. Müller, S. Behnke, Learning depth-sensitive conditional random fields for semantic segmentation of RGB-D images, in: Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014, pp. 6232–6237.

[83] M. Szummer, P. Kohli, D. Hoiem, Learning crfs using graph cuts, in: Proceedings of the European Conference on Computer Vision, Springer, 2008, pp. 582–595.

[84] N. Plath, M. Toussaint, S. Nakajima, Multi-class image segmentation using conditional random fields and global classification, in: Proceedings of the International Conference on Machine Learning, 2009, pp. 817–824.

[85] C. Wojek, B. Schiele, A dynamic conditional random field model for joint labeling of object and scene classes, in: Proceedings of the European conference on computer vision, 2008, pp. 733–747.

[86] J.M. Gonfaus, X. Boix, J.V.D. Weijer, A.D. Bagdanov, J. Serrat, J. Gonzalez, Harmony potentials for joint classification and segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 3280–3287.

[87] A. Lucchi, X. Boix, K. Smith, P. Fua, Are spatial and global constraints really necessary for segmentation? in: Proceedings of the IEEE International Conference on Computer Vision, 2011, pp. 9–16.

[88] A. Torralba, K.P. Murphy, W.T. Freeman, Contextual models for object detection using boosted random fields, in: Proceedings of the Advances in Neural Information Processing Systems, 2005, pp. 1401–1408.

[89] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, S. Belongie, Objects in context, in: Proceedings of the IEEE International Conference on Computer Vision, 2007, pp. 1–8.

[90] M.P. Kumar, H. Turki, D. Preston, D. Koller, Parameter estimation and energy minimization for region-based semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 37 (7) (2015) 1373–1386.

[91] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient graph-based image segmentation, Int. J. Comput. Vis. 59 (2) (2004) 167–181.

[92] P. Arbelaez, M. Maire, C. Fowlkes, J. Malik, Contour detection and hierarchical image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 898–916.

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

[94] J. Tighe, S. Lazebnik, Superparsing: scalable nonparametric image parsing with superpixels, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 352–365.

[95] S. Gould, J. Zhao, X. He, Y. Zhang, Superpixel graph label transfer with learned distance metric, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 632–647.

[96] R.B. Rusu, S. Cousins, 3d is here: Point cloud library (pcl), in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2011, pp. 1–4.

[97] Y. Lu, C. Rasmussen, Simplified Markov random fields for efficient semantic labeling of 3d point clouds, in: Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 2690–2697.

[98] D. Anguelov, B. Taskarf, V. Chatalbashev, D. Koller, D. Gupta, G. Heitz, A. Ng, Discriminative learning of Markov random fields for segmentation of 3d scan data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 169–176.

[99] J.P. Valentin, S. Sengupta, J. Warrell, A. Shahrokni, P.H. Torr, Mesh based semantic modelling for indoor and outdoor scenes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2067–2074.

[100] X. Xiong, D. Huber, Using context to create semantic 3d models of indoor environments., in: Proceedings of the BMVC, 2010, pp. 1–11.

[101] O. Kahler, I. Reid, Efficient 3d scene labeling using fields of trees, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3064–3071.

[102] T.T. Pham, I. Reid, Y. Latif, S. Gould, Hierarchical higher-order regression forest fields: an application to 3d indoor scene labelling, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2246–2254.

[103] A. Anand, H.S. Koppula, T. Joachims, A. Saxena, Contextually guided semantic labeling and search for three-dimensional point clouds, Int. J. Robot. Res. 32 (1) (2013) 19–34.

[104] B.-S. Kim, P. Kohli, S. Savarese, 3d scene understanding by voxel-crf, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1425–1432.

[105] D. Wolf, J. Prankl, M. Vincze, Fast semantic segmentation of 3d point clouds using a dense CRF with learned parameters, in: Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 4867–4873.

[106] M. Najafi, S.T. Namin, M. Salzmann, L. Petersson, Non-associative higher-order Markov networks for point cloud classification, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 500–515.

[107] Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.

[108] D.H. Hubel, T.N. Wiesel, Receptive fields and functional architecture of monkey striate cortex, J. Physiol. 195 (1) (1968) 215–243.

[109] P. Pinheiro, R. Collobert, Recurrent convolutional neural networks for scene parsing, arXiv:1306.2795 5(2013).

[110] M. Mostajabi, P. Yadollahpour, G. Shakhnarovich, Feedforward semantic segmentation with zoom-out features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3376–3385.

[111] A. Sharma, O. Tuzel, M.-Y. Liu, Recursive context propagation network for semantic scene labeling, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2447–2455.

[112] A. Sharma, O. Tuzel, D.W. Jacobs, Deep hierarchical parsing for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 530–538.

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

[114] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[115] Z. Wu, C. Shen, A.v. d. Hengel, Wider or deeper: revisiting the resnet model for visual recognition, arXiv:1611.10080(2016).

[116] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, arXiv preprint arXiv:1606.00915 (2016).

[117] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv:1511.07122(2015).

[118] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Hypercolumns for object segmentation and fine-grained localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 447–456.

[119] G. Ghiasi, C.C. Fowlkes, Laplacian pyramid reconstruction and refinement for semantic segmentation, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 519–534.

[120] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: multi-path refinement networks with identity mappings for high-resolution semantic segmentation, arXiv:1611.06612(2016).

[121] S.-J. Park, K.-S. Hong, S. Lee, Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4980–4989.

[122] M.A. Islam, S. Naha, M. Rochan, N. Bruce, Y. Wang, Label refinement network for coarse-to-fine semantic segmentation, arXiv:1703.00551(2017).

[123] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene CNNs, arXiv:1412.6856(2014).

[124] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, H. Huang, Cascaded feature network for semantic segmentation of RGB-D images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1311–1319.

[125] G. Lin, C. Shen, A. van den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[126] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P.H. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1529–1537.

[127] A. Arnab, S. Jayasumana, S. Zheng, P.H.S. Torr, Higher order conditional random fields in deep neural networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 524–540.

[128] G. Bertasius, L. Torresani, S.X. Yu, J. Shi, Convolutional random walk networks for semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 858–866.

[129] Z. Liu, X. Li, P. Luo, C.C. Loy, X. Tang, Semantic image segmentation via deep parsing network, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1377–1385.

[130] R. Vemulapalli, O. Tuzel, M.-Y. Liu, R. Chellapa, Gaussian conditional random field network for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3224–3233.

[131] S. Chandra, I. Kokkinos, Fast, exact and multi-scale inference for semantic image segmentation with deep Gaussian CRFs, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 402–418.

[132] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 6645–6649.

[133] X. Li, Z. Liu, P. Luo, C.C. Loy, X. Tang, Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade, arXiv:1704.01344(2017).

[134] Z. Wu, C. Shen, A.v. d. Hengel, High-performance semantic segmentation using very deep fully convolutional networks, arXiv:1604.04339(2016).

[135] S. Rota Bulo, G. Neuhold, P. Kontschieder, Loss max-pooling for semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2126–2135.

[136] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[137] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, arXiv:1602.07261(2016).

[138] P. Duygulu, K. Barnard, J.F.G.D. Freitas, D.A. Forsyth, Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, in: Proceedings of the European Conference on Computer Vision, 2002, pp. 97–112.

[139] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Proceedings of the Advances in Neural Information Processing Systems, 2002, pp. 561–568.

[140] A. Vezhnevets, V. Ferrari, J.M. Buhmann, Weakly supervised structured output learning for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 845–852.

[141] D. Pathak, P. Krahenbuhl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1796–1804.

[142] X. Liu, B. Cheng, S. Yan, J. Tang, T.S. Chua, H. Jin, Label to region by bi-layer sparsity priors, in: Proceedings of the Seventeenth ACM International Conference on Multimedia, 2009, pp. 115–124.

[143] S. Liu, S. Yan, T. Zhang, C. Xu, J. Liu, H. Lu, Weakly supervised graph propagation towards collective image parsing, IEEE Trans. Multimed. 14 (2) (2012) 361–373.

[144] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, S. Yan, Semantic segmentation without annotating segments, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2176–2183.

[145] J. Dai, K. He, J. Sun, Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1635–1643.

[146] S. Hong, H. Noh, B. Han, Decoupled deep neural network for semi-supervised semantic segmentation, arXiv:1506.04924(2015).

[147] P. Luo, G. Wang, L. Lin, X. Wang, Deep dual learning for semantic image segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2718–2726.

[148] N. Souly, C. Spampinato, M. Shah, Semi supervised semantic segmentation using generative adversarial network, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5688–5696.

[149] G. Wang, P. Luo, L. Lin, X. Wang, Learning object interactions and descriptions

for semantic image segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[150] S. Konishi, A.L. Yuille, J. Coughlan, S.C. Zhu, Fundamental bounds on edge detection: An information theoretic evaluation of different edge cues, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1999.
[151] B.C. Russell, A. Torralba, K.P. Murphy, W.T. Freeman, Labelme: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (1–3) (2008) 157–173.
[152] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[153] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[154] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: common objects in context, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 740–755.
[155] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[156] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ade20k dataset, arXiv:1608.05442(2016).
[157] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGB-D images, in: Proceedings of theEuropean Conference on Computer Vision, 2012, pp. 746–760.
[158] K. Lai, L. Bo, X. Ren, D. Fox, A large-scale hierarchical multi-view RGB-D object dataset, in: Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA), 2011, pp. 1817–1824.
[159] H.S. Koppula, A. Anand, T. Joachims, A. Saxena, Semantic labeling of 3d point clouds for indoor scenes, in: Proceedings of the Advances in Neural Information Processing Systems, 2011, pp. 244–252.
[160] S. Song, S.P. Lichtenberg, J. Xiao, Sun RGB-D: a RGB-D scene understanding benchmark suite, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 567–576.
[161] A. Dai, A.X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: richly-annotated 3d reconstructions of indoor scenes,arXiv:1702.04405(2017).
[162] X. Xiong, D. Munoz, J.A. Bagnell, M. Hebert, 3-d scene analysis via sequenced predictions over points and regions, in: Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA), 2011, pp. 2609–2616.
[163] K. Yamaguchi, M.H. Kiapour, L.E. Ortiz, T.L. Berg, Parsing clothing in fashion photographs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3570–3577.
[164] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, L. Lin, S. Yan, Deep human parsing with active template regression, IEEE Trans. Pattern Anal. Mach. Intell. 37 (12) (2015) 2402–2414.
[165] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv:1503.02531(2015).
[166] L.C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv:1706.05587(2017).
[167] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 818–833.
[168] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization, arXiv:1506.06579(2015).
[169] A. Mahendran, A. Vedaldi, Salient deconvolutional networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 120–135.
[170] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3d shapenets: a deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1912–1920.
[171] D. Maturana, S. Scherer, Voxnet: a 3d convolutional neural network for real–time object recognition, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),, IEEE, 2015, pp. 922–928.
[172] C.R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, L.J. Guibas, Volumetric and multi-view CNNs for object classification on 3d data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5648–5656.
[173] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3d shape recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 945–953.
[174] X. Chen, H. Ma, J. Wan, B. Li, T. Xia, Multi-view 3d object detection network for autonomous driving, arXiv:1611.07759(2016).
[175] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, arXiv:1612.00593(2016).
[176] B. Jin, M.V. Ortiz Segovia, S. Susstrunk, Webly supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3626–3635.
[177] S. Hong, D. Yeo, S. Kwak, H. Lee, B. Han, Weakly supervised semantic segmentation using web-crawled videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7322–7330.
[178] J.U. Kim, H.G. Kim, M.R. Yong, Iterative deep convolutional encoder-decoder network for medical image segmentation, in: Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, 2017, pp. 685–688.
[179] X. Liang, Y. Wei, Y. Chen, X. Shen, J. Yang, L. Lin, S. Yan, Learning to segment human by watching youtube., IEEE Trans. Pattern Anal. Mach. Intell. 39 (7) (2016) 1462–1468.
[180] M. Wertheimer, Laws of organization in perceptual forms, in: A Source Book of Gestalt Psychology, London: Routledge & Kegan Paul, 1923.

Hongshan Yu received the B.S., M.S. and Ph.D. degrees of Control Science and Technology from electrical and information engineering of Hunan University, Changsha, China, in 2001, 2004 and 2007 respectively. From 2011 to 2012, he worked as a postdoctoral researcher in Laboratory for Computational Neuroscience of University of Pittsburgh, USA. He is currently an associate professor of Hunan University and associate dean of National Engineering Laboratory for Robot Visual Perception and Control. His research interests include autonomous mobile robot and machine vision.

Zhengeng Yang received the B.S. and M.S. degrees from Central South University, Changsha, China, in 2009 and 2012 respectively. He is currently a Ph.D. candidate at the Hunan University, Changsha, China. His research interests include computer vision, image analysis and machine learning, especially focus on the problems of semantic segmentation.

Lei Tan received the B.S. and M.S. degrees of Control Science and Technology from electrical and information engineering of Hunan University, Changsha, China, in 2007 and 2010 respectively. From 2013 to 2015, he visited the Robotics Institute at Carnegie Mellon University, USA as a joint training Ph.D. student sponsored by the China Scholarship Council. His research interests mainly focus on computer vision and mobile robot.

Yaonan Wang received the B.S. degree in computer engineering from East China Technology Institute in 1981, and the M.S. and Ph.D. degrees in electrical engineering from Hunan University, Changsha, China in 1990 and 1994 respectively. From 1994 to 1995, he was a Postdoctoral Research Fellow with the National University of Defense Technology. He is currently a Professor at Hunan University. His research interests are image processing, pattern recognition, and robot control.

Wei Sun received the M.S. and Ph.D. degrees of Control Science and Technology from the Hunan University, Changsha, China, in 1999 and 2002, respectively. He is currently a Professor at Hunan University. His research interests include artificial intelligence, robot control, complex mechanical and electrical control systems, and automotive electronics.

Mingui Sun received the B.S. degree in instrumental and industrial automation from Shenyang Chemical Engineering Institute, Shenyang, China, in 1982, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pittsburgh, Pittsburgh, PA, USA, in 1986 and 1989, respectively. He is currently a Professor of neurosurgery, electrical and computer engineering, and bioengineering. His current research interests include advanced biomedical electronic devices, biomedical signal and image processing, sensors and transducers, artificial neural networks.

Yandong Tang received the B.S. and M.S. degrees in mathematics from Shandong University, China, in 1984 and 1987, respectively. He received the Ph.D. degree in applied mathematics from the University of Bremen, Germany, in 2002. He is a Professor at the Shenyang Institute of Automation, Chinese Academy of Sciences. His research interests include numerical computation, image processing, and computer vision.