SCIENCE CHINA
Information Sciences

• REVIEW •

May 2018, Vol. 61 051101:1–051101:18
doi: 10.1007/s11432-017-9189-6

Survey of recent progress in semantic image segmentation with CNNs

Qichuan GENG[1], Zhong ZHOU[1*] & Xiaochun CAO[2]

[1]*State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing* 100191*, China;*
[2]*State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing* 100093*, China*

Received 18 March 2017/Accepted 20 July 2017/Published online 17 November 2017

Abstract In recent years, convolutional neural networks (CNNs) are leading the way in many computer vision tasks, such as image classification, object detection, and face recognition. In order to produce more refined semantic image segmentation, we survey the powerful CNNs and novel elaborate layers, structures and strategies, especially including those that have achieved the state-of-the-art results on the Pascal VOC 2012 semantic segmentation challenge. Moreover, we discuss their different working stages and various mechanisms to utilize the structural and contextual information in the image and feature spaces. Finally, combining some popular underlying referential methods in homologous problems, we propose several possible directions and approaches to incorporate existing effective methods as components to enhance CNNs for the segmentation of specific semantic objects.

Keywords semantic image segmentation, CNN, Pascal VOC 2012 challenge, multi-granularity features, construction of contextual relationships

Citation Geng Q C, Zhou Z, Cao X C. Survey of recent progress in semantic image segmentation with CNNs. Sci China Inf Sci, 2018, 61(5): 051101, doi: 10.1007/s11432-017-9189-6

1 Introduction

Semantic image segmentation is one of the most challenging problems in computer vision. It is also a fundamental prerequisite for various hot topics in computer vision, such as scene understanding [1, 2], reconstruction [3] and image processing.

Before the proposal of the Pascal VOC 2007 semantic segmentation challenge, a great deal of effort had been focused on the geometric labeling of images, which also applies to semantic image segmentation. These methods can be divided into two main classes: statistics-based [4–9] and geometry-based [10, 11] methods. Most parametric statistical methods [4–7] over-segment an image based on several simple features, which are similar to image segmentation methods [12] partitioning a digital image into multiple regions based on image appearance. Then, Markov random field (MRF) methods [5] or grammar methods [8] are used to classify these super-pixels into different geometric classes by extracting complex hand-crafted features. Data-driven nonparametric statistical methods [9] without a training step find the most similar scenes from the retrieval set, on which dense alignment [13] can be implemented. Then, the aligned labels are transferred to the input. Based on constraints of parallelism and verticality between planes and lines, geometric methods calculate geometric labels directly.

* Corresponding author (email: zz@buaa.edu.cn)

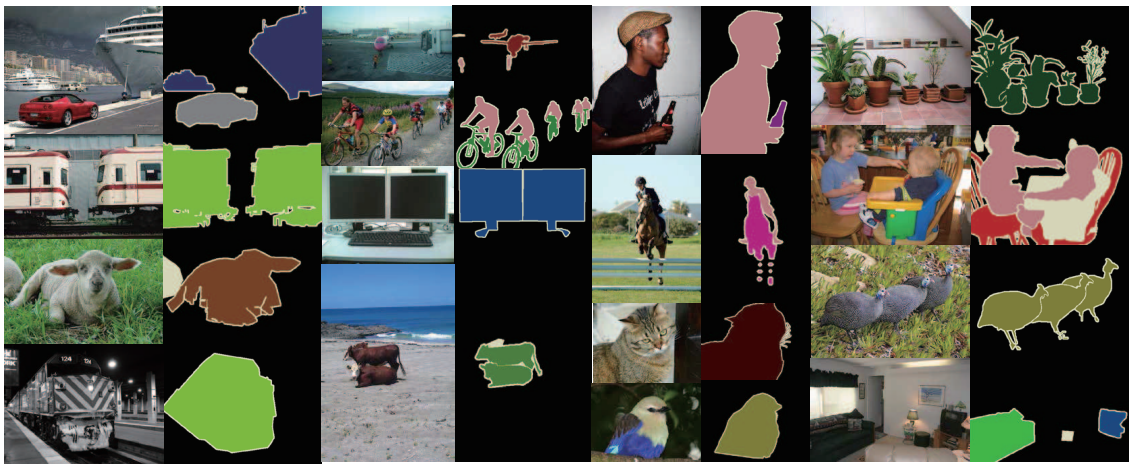


Figure 1 (Color online) Examples from the segmentation subset [17].

Recently, convolutional neural networks (CNNs) have attracted the attention of many researchers because of their ability to automatically extract more compact and meaningful features from images than classical hand-crafted ones. In fact, CNNs have demonstrated clear superiority in many tasks, including image classification and object detection.

During 2015 and 2016, the state-of-the-art results in semantic image segmentation were significantly improved by virtue of CNNs. Semantic image segmentation is equivalent to a pixel-wise classification problem. Approaches such as sliding windows or fully convolutional neural networks allow the existing structures for image classification to be directly adapted to semantic image segmentation.

The spatial-semantic uncertainty principle is the main challenge in semantic segmentation. At higher semantic levels, the resolution of feature maps rapidly decreases in general CNNs, which limits the accuracy of the segmentation results. The downsampling and invariance of CNNs that are desired in many high-level vision tasks hinder the extraction of spatial details [14]. Many recent approaches have attempted to preserve, extract or restore structural information to enhance the highly abstract features obtained from deep layers. Based on rather accurate image classification and detection procedures, the key prerequisite task is to recognize the boundary pixels of objects [15, 16].

Several datasets related to semantic segmentation are available, such as Cityscapes, PASCAL-Context, ADE20K, MS COCO, and BSDS. In this paper, we survey the latest results on the Pascal VOC 2012 semantic segmentation challenge, which is the most representative image dataset for this purpose [17].

The twenty object classes that have been selected are as follows:

- Person: person;
- Animal: bird, cat, cow, dog, horse, sheep;
- Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train;
- Indoor: bottle, chair, dining table, potted plant, sofa, TV/monitor.

This dataset can be used for object classification, detection, segmentation, action classification and a competition on large-scale recognition run by ImageNet. The segmentation subset contains 1464 images for training and 1449 images for validation. Some examples from the segmentation subset are shown in Figure 1. Each of these images contains at least one object and usually some occluded objects captured from various views.

Remarkable improvement has been achieved using several newly proposed layers, structures and strategies.

Various new layers and structures have been proposed to handle contextual information. Some of them are implementations of probabilistic graphic models (PGMs) [18–21], such as conditional random fields (CRFs) and MRFs, for modeling various contextual relationships. The generalization capability of PGMs and the learning capability of CNNs can complement each other. Other approaches for handling contextual relationships take advantage of structures such as recurrent neural networks (RNNs) [22,

23] and long short-term memory (LSTM) neural networks [24], which explicitly propagate long-range contextual information. We believe that certain correlations exist among these methods.

Features with variable resolutions, multiple scales and different levels of abstraction are passed and blended through multiple layers in several of these new structures. Unpooling layers and deconvolutional layers are used alternately to increase resolution and provide more details [25]. Dilated convolution and atrous convolution [26–28] are used to achieve a trade-off among the field of view (FOV), the number of parameters and the resolution of feature maps. Multi-scale features implicitly contain different types of contextual information, thereby enhancing the robustness of frameworks. Resized features with different levels of abstraction can be summed or concatenated to form multi-granularity features.

Because of the use of deeper layers, higher numbers of parameters and more complex structures, more effective strategies are needed to accelerate the training process. The computational cost of a single iteration should be reduced to ensure that the entire network is tractable. To avoid over-fitting caused by an unbalanced dataset, the training samples and objective functions can be augmented.

Instead of comparing the performances of various new layers, structures and strategies, we place our emphasis on revealing the reasons why these remarkable accomplishments have been achieved. We believe that by analyzing their mechanisms, we can determine how they can be combined in a mutually beneficial manner to enhance the final output.

2 Recent progress in semantic image segmentation

In recent years, significant progress has been made in semantic image segmentation. We list the most important work with the best results achieved by each architecture in Tables 1 and 2. In high-level vision tasks, the shape information is lost in the features that are extracted, layer by layer, by CNNs, with increases in the degrees of nonlinearity and semantic abstraction. However, accurate segmentation requires information of this kind to be appended to the final output.

There are two commonly applied schemes for recovering the configurations of semantic objects: combining information from different CNN layers and constructing more contextual relationships in the image or feature space.

Combining information from different CNN layers. Skip-layer architectures [29] are able to present multi-granularity information aggregated from different layers. Considering the inconsistent resolutions of feature maps, the methods used to enable the combination of feature maps are the main factors considered when designing a framework for producing fine segmentations. The feature maps are either upsampled to a given size or kept invariant in resolution. Therefore, reducing the loss of structural information between layers and improving the effectiveness of interpolation are the keys to restoring more structural information. In fact, these approaches simulate the behavior of the human mind in coping with the same problem. However, because the receptive fields of CNNs can not keep matching with observed objects, this kind of architecture cannot perform global optimization based on experience and logic in exactly the same way as human beings.

Constructing more contextual relationships. Before CNNs are applied for semantic image segmentation, PGMs, including CRFs and MRFs, are commonly used for modeling the connections among nodes, which correspond to pixels or super-pixels [4–7] in images. The inputs to the energy function are usually hand-crafted features, which may limit the effectiveness of these methods. Moreover, traditional inference algorithms, with their high computing costs, cannot be successfully adapted for application to larger datasets [30].

With the adoption of CNNs for semantic image segmentation, CRFs can be an efficient post-processing component for smoothing the output from such CNNs [25]. By virtue of the proposal of approximate algorithms [30] for inferring segmentations, CRFs can be implemented as layers in conventional CNNs. Simultaneously, the end-to-end training strategy further improves the result. Methods of converting CRFs into CNN layers and rapidly training such frameworks are the main contributions of the related work. RNNs are effective tools for depicting long-range dependencies in time and space. An initial segmentation

Table 1 PASCAL VOC 2012 Challenge Leaderboard (2015/10/2)

Architecture	Mean	Architecture	Mean
Adelaide_Context_CNN_CRF_COCO [19]	77.8	DeepLab-CRF-COCO-LargeFOV [27]	72.7
CUHK_DPN_COCO [20]	77.5	POSTECH_EDeconvNet_CRF_VOC [25]	72.5
CentraleSuperBoundaries [51]	75.7	Oxford_TVG_CRF_RNN_VOC [18]	72.0
Adelaide_Context_CNN_CRF_VOC [19]	75.3	DeepLab-MSc-CRF_LargeFOV [27]	71.6
MSRA_BoxSup [55]	75.2	DeepLab-CRF-COCO-Strong [27]	70.4
POSTECH_DeconvNet_CRF_VOC [25]	74.8	DeepLab-CRF-LargeFOV [27]	70.3
Oxford_TVG_CRF_RNN_COCO [18]	74.7	TTI_zoomout_v2 [44]	69.6
DeepLab-MSc-CRF-LargeFOV-COCO-CrossJoint [27]	73.9		

Table 2 PASCAL VOC 2012 Challenge Leaderboard (2017/2/1)

Architecture	Mean	Architecture	Mean
PSPNet [48]	85.4	CentraleSupelec Deep G-CRF	80.2
ResNet-38_COCO [39]	84.9	CMT-FCN-ResNet-CRF	80.0
Multipath-RefineNet [38]	84.2	DeepLabv2-CRF [27]	79.7
ResNet-38_MS [39]	83.1	CASIA_SegResNet_CRF_COCO	79.3
R4D_MultiScale_CRF	82.2	LRR_4x_ResNet_COCO [15]	79.3
SegModel	81.8	Adelaide_VeryDeep_FCN_VOC	79.1
HikSeg_COCO	81.4	LRR_4x_COCO [15]	78.7
DP_ResNet_CRF	81.0	CASIA_IVA_OASeg	78.3
OBP-HJLCN	80.4	Oxford_TVG_HO_CRF [46]	77.9

can be refined through repeated iteration in an RNN. RNNs can be not only implementations of PGMs but also a special kind of flexible structure for explicitly propagating contextual information. As a type of RNN that incorporates a gate function, LSTM neural networks have become popular structures that can either remember or forget inconsistent contextual information. Our experimental results show that although the accuracy on the test set is improved, the robustness of the system to images outside the training set remains an open question.

2.1 Implementations of the combination of multi-granularity features

As seen from the results produced by 16-layer VGGNet (VGG-16) [31] that are presented in Figure 2, the feature maps become sparser in both the feature dimension and the spatial dimension from one layer to the next, and higher-level semantic information becomes concentrated in fewer maps as the analysis proceeds through the entire framework. The semantic levels of various multi-granularity features arrange from color and edge to typical structures of subparts of or even whole objects. High semantic level features can be utilized to localize and classify the objects roughly. Conversely, low semantic level features related to local appearance are not insufficient to classify the observed patch but helpful to discriminate details of objects, especially which are close to boundaries.

With regard to the pixel-wise classification task, pixels can be divided into those that belong to easy and difficult regions. Difficult pixels are usually located near boundaries and are the main basis on which the effectiveness of an approach can be evaluated. Feature encoders, such as the VGG-16 and ResNet [32], output low-resolution feature maps based on stacked convolution layers, ReLu layers and pooling layers. In general CNNs, the FOV determines the accuracy of segmentation. Because of the lack of complete information on the detected objects, a small FOV may lead to false positives or false negatives. To enlarge the size of the FOV, pooling layers and downsampling layers are introduced into CNNs, thereby reducing the resolution of the output and the quality of details. Therefore, to improve the accuracy of segmentation and refine the proposed boundaries, it is natural to attempt to enlarge the FOV without losing resolution and to integrate the ability to handle incomplete objects. Notably, a large number of intermediate multi-granularity feature maps need to be used to further improve semantic segmentation

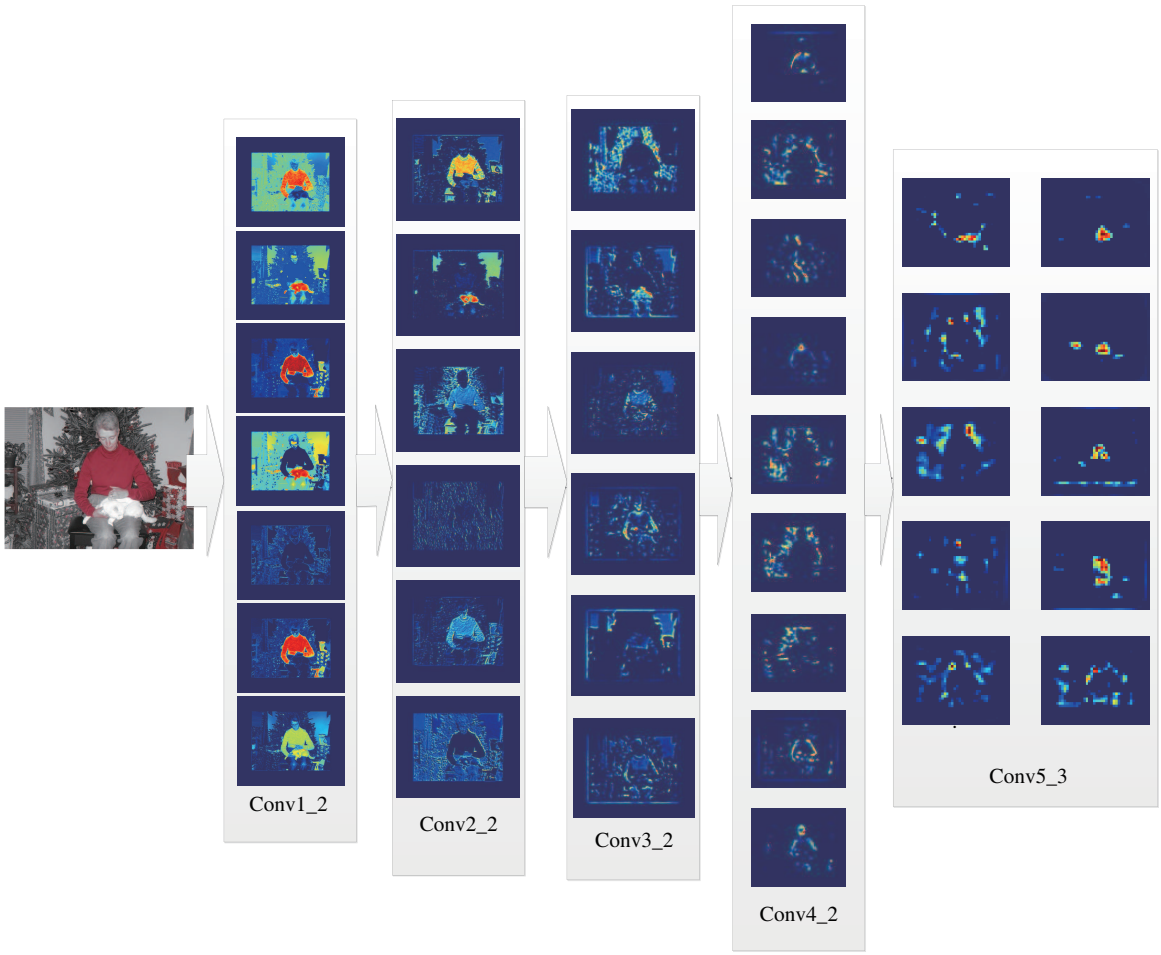


Figure 2 (Color online) Different semantic feature maps obtained from different VGG-16 layers. The sizes of the feature maps have been adjusted for ease of observation.

in various ways. However, a too large FOV may incorporate some irrelevant contexts, which lead to mislabeling, using too many parameters to overfit the data and consuming too much computing resource. Therefore, constraints on parameters and multi-scale methods are adopted to ease these issues.

Unpooling layers and deconvolutional layers [25] are efficient and critical structures for calculating more detailed feature maps at higher resolutions. Instead of classical smooth interpolation, an upsampling operation that can recover details, especially boundary information, is desired. In general, unpooling layers are placed in symmetric positions to resize the input maps to larger maps with structural patterns that are recorded in corresponding pooling layers. Deconvolutional layers usually serve as complex nonlinear interpolating filters for inferring the original shapes of semantic objects. Ghiasi et al. [15] designed their complete structural learning technique based on the idea of dividing an image into different frequency bands using classical methods. During the process of combining multi-granularity information, the details could be recovered from shallow feature maps. They artificially embedded basis functions containing prior knowledge into the deconvolution layers to effectively reconstruct high-resolution details. During end-to-end training, these basis functions were found not to vary significantly, thereby indirectly demonstrating the effectiveness of using artificial features in CNNs. This approach expands the usage of deconvolution layers to reconstruction tasks. Off-the-shelf realizations of unpooling layers and deconvolutional layers are available that are easy to configure and train. To ameliorate the problem of ill-conditioning in appearance recovery, some frameworks attempt to avoid excessive downsampling. In the Deep Parsing Network (DPN) approach, some pooling layers are removed from VGG-16 to improve the resolutions of the intermediate layers, and the kernels in the convolution layers are padded to a larger size to ensure the

usability of the existing weights. Overfeat [28] shifts the input and interlaces the output to yield denser predictions without interpolation.

Considering the number of parameters, atrous convolution and dilated convolution [26, 27] are used to reduce the loss of resolution in feature maps that is introduced by pooling layers. Experiments demonstrate that an enlarged FOV and implicitly aggregated multi-scale contextual information can significantly improve the results. Dropout, which is also a helpful module to avoid over-fitting, has been proven to reduce the Rademacher complexity in polynomial or exponential [33].

Wu et al. [34] also noted the importance of the FOV. To make a trainable network with a larger FOV, they proposed a multi-pass method that can run with limited GPU memory. In their network, multiple low-resolution score maps produced by a structure with the same parameters are stitched together to form a higher-resolution score map.

Hypercolumns [35] concatenate different responses at the same position in bilinearly interpolated heat maps from several layers to obtain a feature vector for each pixel. Considering the heterogeneous distributions of the features in the bounding boxes of detected objects, K×K logistic regression classifiers that take hypercolumns as input are interpolated to infer the results.

The Fully Convolutional Network (FCN) approach [36] is a piece of pioneering work on the realization of a skip-layer architecture with a directed acyclic graph (DAG) network. The best results are produced by FCN-8s, in which shallow predictions are added into upsampled predictions three times. Following this idea, FCN-2s [37] is equipped with a powerful function for discovering the boundaries of semantic objects. Although the results produced by FCN have now been far surpassed, it has served as a vital foundation for subsequent work. Similar to FCN, RefineNet [38] is a cascade-like multipath network. Low-level information is introduced step by step. Chained residual pooling as well as long- and short-range residual connections improve the high-resolution predictions. We consider that the residual connections present in ResNet [32] can stabilize the structure of the feature space, as proven by experiments using RefineNet.

Without explicitly combining coarse and fine information, DeconvNet [25] appends a symmetric segmentation network to the end of the VGG-16, which is treated as the feature encoder. In 2016, the more powerful ResNet was adopted as a replacement for the classical VGG-16 [15, 34, 39] to directly improve the results. The segmentation network is constructed with multiple series of unpooling, deconvolution and rectification operations. Furthermore, the results are still comparable to those of existing studies if the deconvolution layers are replaced with conventional convolution layers. We believe that the activation patterns of the pooling layers are the most meaningful component of this inference process. This scheme is extremely simple and clear, allowing it to be easily adapted into many more powerful frameworks.

Simple combinations of multiple semantic levels usually ignore the different effectiveness of different features. However, it is hoped that in addition to the learned sparse effective parameters, these specific connections can be enhanced to be sensitive to certain variations, such as variations in scale, position or even semantic category, to further constrain the activation of neural cells. Attentional models can serve as hyperparameters to softly weight different features or scores. With different considerations, attentional models learned from networks can be used to enhance the accuracy and robustness of the overall approach. Hong et al. [40] focused on an attentional model to generate category-specific saliency results for each location in an image, thereby revealing location information for each category in a coarse feature map. A dense and detailed foreground segmentation mask for each category could subsequently be obtained by the decoder. Chen et al. [41] mainly considered the problem of robustness to scale by explicitly introducing factors of this kind.

Kuen et al. [23] noted that CNNs do not work well for objects of multiple scales. A recurrent attentional convolutional-deconvolutional neural network (RACDNN) uses a spatial transformer and an RNN to perform saliency refinement. The saliency refinement is applied locally to selected subregions of the image.

Explicitly considering the completeness of segmentation, Multi-scale Patch Aggregation (MPA) [42] uses a suitable FOV to segment parts of or entire objects, which are then integrated into a complete configuration. This is a novel way to promote robustness to scale variations.

Bertasius et al. [43] also took advantage of boundary cues to enhance the FCN approach. Because

Table 3 Effects of various methods for combining multi-granularity features

Implementation	Mean/Relative improvement
FCN-8s [36]	62.2/0
DeconvNet [25]	69.6/7.4
EDeconvNet [25]	71.7/9.5
DecoupledNet-Full [45]	66.6/4.4
DPN [20]	74.1/11.9
DPN_With_COCO [20]	77.5/15.3
Hypercolumn Sys1 [35]	54.6/−7.6
Hypercolumn Sys2 [35]	62.6/0.4
Zoom-out [44]	69.6/7.4
LRR_4x_ResNet_COCO [15]	79.3/17.1
LRR_4x_COCO [15]	78.7/16.5
Multipath-RefineNet-Res152 [38]	83.4/21.2
DeepLab-CRF-COCO-LargeFOV-Attention [41]	75.1/12.9
DeepLab-CRF-COCO-LargeFOV-Attention+ [41]	75.7/13.5
TransferNet [40]	51.2/−11

of the distinct appearance of pixels near boundaries, the recognition of semantic boundaries enables the achievement of significantly better accuracy than semantic segmentation [37], and the results can be treated as a believable prior to constrain the refinement of a coarse segmentation. The coarse segmentations predicted by a semantic FCN are used to define the unary potentials of the boundary neural field (BNF). The local boundaries are then used to build pairwise pixel affinities. The pairwise potentials can be used to globally refine the initial segmentation, and these potentials are added into the CRF energy function introduced in Subsection 2.2. However, incomplete outlines and tiny false boundaries introduce failure modes that require additional research.

Because over-segmentations based on differences in appearance can yield richer statistical features, classical image segmentation approaches usually use super-pixels as inputs for the hierarchical generation of the final segmentation. Mostajabi et al. [44] presented a CNN framework for extracting features from a sequence of nested regions called zoom-out regions, which increase in size from super-pixels to even larger scales. Then, the super-pixels are classified by a multi-layer neural network based on the features assigned to these regions.

In all of the methods discussed above, the output from the feature encoder is passed as a whole to the subsequent segmentation network. However, not all feature maps will be well suited for specific semantic objects, and some may even cause the quality of the segmentation results to degrade. To select class-specific information, DecoupledNet [45] uses a bridging layer to decouple the classification network and the segmentation network. The bridging layer takes the output from the last pooling layer and uses the relevance of activations in $f(k)$ with respect to a specific class l to construct a class-specific activation map for a given semantic label. Then, the corresponding class-specific segmentation network can be trained in isolation.

The effects of these methods are summarized in Table 3, based on the results of testing on the Pascal VOC 2012 dataset. FCN-8s is chosen as the baseline. Table entries grouped by horizontal lines share similar underlying concepts.

As in the case of augmented frameworks such as EDeconvNet [25], combinations of the different methods listed in Table 3 can offer further improvements in accuracy. Many experimental results have demonstrated that these methods are, to some extent, independent. As seen in Figure 3, despite a general lack of experiments, the different methods discussed above appear to make effect in distinct stages of the whole process using different mechanisms. Regardless of the size of the dataset, the available video memory and the number of parameters, almost all of these structures can be combined to independently improve the segmentation accuracy. In the case of larger datasets, we believe that the applicable mixture of these

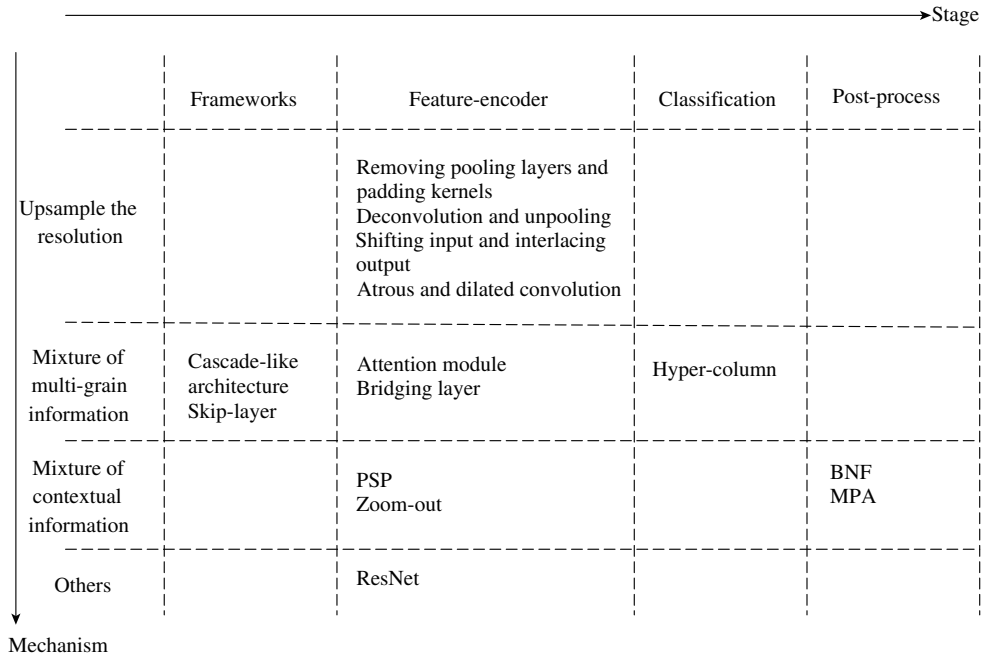


Figure 3 Distinct stages of applicability and different underlying mechanisms of the analyzed methods and structures.

methods can effectively enhance the robustness and accuracy of segmentation frameworks.

2.2 Implementations of the construction of contextual relationships

We divide contextual information into three categories: the appearance context, the high-level feature context and the semantic context.

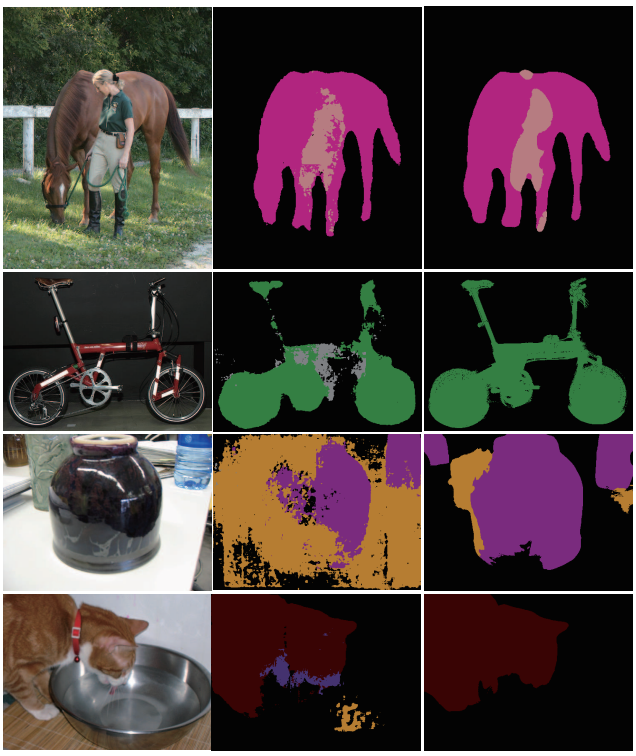
• The appearance context can be used to smooth an initial segmentation based on consistency of appearance.

• Because appearance is sometimes unreliable, high-level features, which are inherently more closely related to semantic information, can be used to further disambiguate the initial segmentation.

• The semantic context mainly describes the functional compatibility among recognized objects in an initial segmentation, which can play a large role in eliminating misunderstandings in local regions.

A CNN can propagate contextual information from inside the equivalent FOV. However, the availability of inconsistent or global information is limited in the general CNN architecture. To explicitly propagate this information, Shuai et al. [22] treat the input image as an undirected cyclic graph (UCG) structure, which can be represented as a set of DAGs with different context propagation directions. DAG-RNNs have been proposed for processing DAG-structured images, thereby allowing the network to explicitly model long-range semantic contextual dependencies. LG-LSTM extends the Grid-LSTM approach to the global context for the iterative refinement of the current segmentation.

PGMs have been adopted from classical methods for application to CNNs; they complement CNNs by providing a feasible mechanism for handling many kinds of contexts. Semantic image segmentation using PGMs can be treated as a graph cut problem. Contextual relationships can be formulated as edges in the graph model and different-order potentials in the energy function. Fully connected CRF models [30] with pairwise Gaussian edge potentials, which are conditioned on the input, are commonly used for this purpose. In a model of this kind, connections are established between all pairs of pixels in an image. These connections are able to depict various complex contextual relationships, including occlusion, spatial relations and discontinuity, as shown in Figure 4 [25]. However, the practical applicability of fully connected CRFs is limited by the rapid growth in the number of edges with increasing input resolution.



(a) (b) (c)

Figure 4 (Color online) Obvious false recognitions are evident in column (b), which can be corrected by using a CRF model. The post-processed results obtained using the CRF approach are shown in column (c).

We let x_i denote the label assigned to pixel i and let I denote the possible input images of size N. In the fully connected CRF model with potentials of order two, the corresponding Gibbs energy function [30] is

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j), \tag{1}$$

where the unary potential term $\psi_u(x_i)$ measures the cost of assigning label x_i to pixel i and the pairwise term $\psi_p(x_i, x_j)$ measures the smoothness of assigning labels x_i and x_j to pixels i and j simultaneously.

The maximum a posteriori (MAP) labeling of the random field is $x^* = \text{argmax}_{x \in L^N} P(x|I)$, where $P(x|I)$ has the following form:

$$P(x|I) = \frac{1}{Z(x)} \exp(-E(x)). \tag{2}$$

The energy functions of MRFs [20] have a form similar to that for CRFs.

Message passing (MP) algorithms [21], such as loopy belief propagation (LBP), tree-reweighted message passing and the mean-field (MF) approximation, are commonly applied to approximate inference. As demonstrated in [18–21], MP algorithms can be implemented as components of CNNs or as entire neural networks trained in an end-to-end manner.

In the case of the MF approximation, the series of independent marginal probabilities produced by minimizing the KL divergence are taken as approximations to the accurate distribution. The simpler approximated distribution function $Q(x)$ [30] can be written as follows:

$$Q(x) = \prod_i Q_i(x_i), \tag{3}$$

$$Q_i(x_i) = \frac{1}{Z(x_i)} \exp\left\{ -\psi_u(x_i) - \sum_{l'} \mu(l, l') m(l') \right\}, \tag{4}$$

Table 4 Transforming a CRF into an RNN

Step of iteration	Implementation in the RNN
Initialization	Softmax layer
Message passing	Gaussian filters
Weighting of filter outputs	1×1 convolutional layer
Compatibility transform	1×1 convolutional layer
Addition of unary potentials	Elementwise layer
Normalization	Softmax layer

$$m(l') = \sum_{m=1}^{K} \omega^m \sum_{j \neq i} k^m(f_i, f_j) Q_j(l'). \tag{5}$$

Therefore, the inference procedure is simplified to an iterative message passing procedure in the graph, which can be solved quickly and accurately. Subsequently, the unary potentials and pairwise potentials are learned, either sequentially or simultaneously.

CRF-RNN [18] reformulates the iterative algorithm presented in [30] as an RNN. Each step in the iteration is equivalent to a conventional layer used in CNNs, as shown in Table 4. Since the output of the classifier for each pixel is independent of the outputs of the classifiers for the other pixels, the unary potentials obtain inputs from the existing FCN. With the termination criterion set to a fixed number of iterations, the distribution $Q(x)$ can be assumed to converge to the actual distribution.

Context_CNN_CRF [19] won first place on the PASCAL VOC 2012 Challenge in 2015. The workflow of this framework can be divided into three parts: Network Part 1, which extracts multi-scale features; Network Part 2, which produces potentials; and a final stage for training or prediction. As a common approach to extracting hand-crafted image features, multi-scale networks [14, 37] can implicitly contain various contextual relations and enhance the robustness to scale variations of semantic objects. Network Part 2 constructs one network for the unary potential and a network for each pairwise potential function corresponding to a specific asymmetric contextual relation, such as "surrounding" or "above/below". These potential networks have different parameters but the same architecture. The procedure for label prediction can be performed by applying an efficient message passing algorithm based on the MF approximation. The prediction converges within 2 iterations, which demonstrates the effectiveness of the potentials produced by the first two parts of the framework.

Time-consuming iterations are not desirable for CNNs trained on large datasets. The DPN [20], whose pairwise potentials are generalizations of various previously proposed pairwise terms, is formed by stacking conventional layers onto a VGG-16 network. The MF approximation can be achieved within only one iteration in the DPN, reducing the computational cost while maintaining high performance.

Higher-order potentials usually describe more contextual information. However, in the work discussed above, the dimension of the output is K^a, where K is the number of semantic classes and a is the order of the potentials. Lin et al. [21] constructed a message estimator learned using a CNN architecture. The messages passed during the inference procedure, which are K-dimensional vectors, are estimated directly. This network is also compatible with other MP algorithms in addition to the MF approximation.

Arnab et al. [46] added higher-order potentials related to super-pixels and object detection into the general CRF approach. Their experiments demonstrate the connection between semantic segmentation and object detection. This approach takes advantage of the boundary consistency between super-pixels and object regions.

The continuous Gaussian CRF [47] involves a Gaussian mean field (GMF) network over a Gaussian CRF, which can be solved optimally. GCRFs model continuous quantities and can be efficiently solved using linear algebra routines. These implementations of the construction of contextual relationships represent the top five performers on the PASCAL VOC 2012 dataset in 2015, with performances ranging from 75.3% to 77.8%.

To our knowledge, CRFs continued to play an important role in enabling further improvements to the accuracy of semantic segmentation in 2016 [15, 27, 34, 47]. The cited experiments demonstrate that CNNs

show weaknesses in analyzing the connections among small ambiguous regions, which can be compensated for by using contextual information. Although these studies are not compatible with each other, they can be used to refine the combined multi-granularity features described in Subsection 2.1. We consider an ideal CRF implementation to have the following characteristics:

(1) A non-iterative workflow that can reduce the computational load;

(2) Parameters trained in an end-to-end manner to achieve better performance;

(3) Good compatibility with various contextual relationships.

Thus, a version of Context_CNN_CRF adapted to learning messages through message passing inference as proposed by Lin et al. [21] is expected to be a good choice.

In addition to the contextual information among pixels, a global prior representation [48] can also be used to eliminate the ambiguity between semantic objects and to determine the type of scene as a whole. PSPNet adds a pyramid pooling module, which consists of commonly used neural layers, back into ResNet. This module concatenates a multi-scale composite constraint with the feature maps from ResNet, thereby achieving a significant boost in accuracy.

2.3 New training strategies

The networks discussed in Subsections 2.1 and 2.2 involve additional parameters and a higher degree of nonlinearity. The parameters in these networks are expected to converge to the global optimal solution in an acceptable time with limited samples. In addition to common tricks used in other tasks, such as flipping, rotation, blur and cropping, three aspects should be considered during the end-to-end training of such deep learning frameworks:

(1) Training new layers from scratch and utilizing the existing models to fine-tune the networks;

(2) Solving the CRF objective function in the networks;

(3) Avoiding over-fitting with unbalanced samples in the dataset.

The task of semantic image segmentation requires invariance with respect to object location and scale. DeconvNet significantly reduces the search space by means of a two-stage training method. In the first stage, training samples aligned with the semantic objects of interest are cropped from the input images to significantly reduce the search space. In the second stage, training samples with more variation are used to increase the robustness of the framework.

As described in Subsection 2.2, the DPN is constructed by stacking randomly initialized layers on an adapted VGG-16 network one by one. Using a greedy strategy, only one new layer and a loss function are added to the network and are then trained while keeping all previous parameters fixed. Finally, all parameters are fine-tuned simultaneously.

We introduce two main methods of training the implementations of CRFs discussed above.

For the RNN structure, the back-propagation algorithm and the Stochastic Gradient Descent (SGD) approach are introduced. During backward passing, the error differentials propagate within the loop until the termination criterion is reached.

To reduce the time spent on inference, which is influenced by the partition function Z in (2), piecewise training is employed to jointly train the CNN and CRFs. In this method, the original graph is divided into smaller, nonadjacent sub-graphs to avoid the computation of the global partition function Z.

Insufficient samples of specific types of semantic objects [44] can lead to poor results. Hence, precautions should be taken to address this issue. For example, in the zoom-out approach, a simple method is used to rewrite the loss function according to the frequencies of the different semantic classes. In [22,49], the loss functions incorporate weights for the different categories.

Wu et al. [34] use a bootstrapping method for training. They incorporate a threshold into the loss function to guide the training of the network to focus on more difficult and more valuable pixels. In summary, because of the large size of the semantic segmentation space compared with the size of the dataset, the following strategies are used:

- Dividing the network into stages or layers;
- Dividing datasets into groups;

- Dividing PGMs into cliques;
- Fine-tuning the parameters in an end-to-end manner;
- Reformulating the loss function.

3 Progresses related to semantic segmentation

With the rapid development of deep learning, many CNN-based approaches implicitly related to semantic segmentation have been introduced for this task, such as image classication and object detection. Methods for the extraction of more meaningful features and the utilization of different types of annotations are summarized below.

3.1 Methods related to semantic segmentation

All of the studies discussed above have been based on the hypothesis that shallow features are sufficiently meaningful for semantic segmentation; however, this hypothesis has not yet been proven. We believe that although under-constrained features are generally useful for various tasks, retrained task-specific intermediate features can be more suitable for specific purposes. In particular, the features associated with boundaries are desirable for refining practical segmentations.

More robust shallow features can be introduced to produce better segmentations more efficiently. In a deeply supervised net (DSN) [50], the intermediate features and classification error are explicitly supervised simultaneously by reformulating the loss functions. It has been proven that these reformulated objective functions follow the same configuration as the original functions. Experiments have demonstrated that the parameters can converge faster with fewer training samples.

The practicality of segmentation results is more strongly affected by the boundaries of regions than by the pixels inside them. The attempt to extract boundaries with finer outlines can be seen as the foundation for many subsequent studies.

Holistically nested edge detection (HED) [37] is based on a concept similar to that of DSNs for the construction of a skip-layer architecture. Side-output layers with a classifier can be trained for boundary prediction using the desired features. The multi-scale and multi-level outputs are combined in a weighted manner to produce more accurate and meaningful boundaries.

Based on the DSN and HED approaches, Kokkinos [51] integrated multi-scale input and augmented data with a normalized cut to address the low-level vision task of boundary detection. The proposed structure is able to surpass human performance. Meanwhile, there are also many algorithms [52–54] for accelerating CNN computations that are worthy of interest. However, in this paper, we place our emphasis on the study and analysis of contributions related to the design of effective CNN architectures rather than computational methods.

3.2 Semantic image segmentation with special annotations

Annotations for semantic image segmentation require expensive labeling efforts. The workload for labeling segmentation masks is more than 15 times heavier than that for labeling image contents or object locations [14, 55], and this restricts the scale and distribution of such annotated datasets. It is well known that the use of a larger dataset can sometimes improve the results more than can be achieved by introducing a new effective layer into the network. Datasets for image classification and object detection, which are cheaper to obtain, are therefore expected to be used for the semantic segmentation task.

For the purpose of expanding the usable datasets, experiments reported by Dai et al. [55] prove that complete annotations of semantic segmentations are not necessary. Only one in ten images in their semantic segmentation dataset is fully supervised with masks, whereas the others are labeled with bounding boxes; this level of annotation can provide sufficient information to yield results comparable to the baseline. The workflow can be divided into two parts: region proposals and CNN analysis. Intuitively, this method takes advantage of the discrimination power of CNNs and the ability of unsupervised methods to find foregrounds. In algorithms such as Grabcut [56], selective search [57], MCG [58] and GOP [59],

shallow features are used to refine the boundaries. These iterative processes exploit the complementarity between semantic information and image features.

Papandreou et al. [14] adopted the expectation maximization (EM) algorithm to utilize bounding boxes and image-level labels. Three strategies were tested: Bbox-rect, Bbox-seg and Bbox-EM-fixed. In the Bbox-rect method, all pixels in a bounding box are taken as positive examples. In Bbox-seg, a CRF model is used to extract a foreground/background segmentation automatically. In Bbox-EM-fixed, the estimated segmentation is refined throughout training. The accuracy achieved on a dataset with only image-level labels was only 39.6%, whereas the accuracy on a dataset with only bounding boxes reached a level comparable to that achieved using fully supervised data. Thus, image-level labels are not sufficient to train a high-quality segmentation model. These findings demonstrate the theoretical gap between image classification and semantic segmentation and the implicit overlap between object detection and semantic segmentation.

BoxSup [55] also uses bounding boxes as annotations to train a segmentation network. The framework alternates between automatically generating region proposals and training convolutional networks. The multiscale combinatorial grouping (MCG) method is used to propose regions that are assigned semantic labels, and a CNN estimates segmentation masks from these regions. We believe that bounding boxes can enhance the ability of networks to recognize objects in images. In the iterative workflow, the candidate masks produced by supervised or unsupervised algorithms are corrected to update the parameters in the CNN. One advantage of this architecture is that it is a FCN structure during the test stage, which guarantees its efficiency.

Following the idea of decoupling the classification network from the segmentation network, Decoupled-Net [45] uses image-level labels to train the classification network and segmentation masks to train the segmentation network. The different types of datasets are input into the different parts of the framework during the training stage. This architecture fully utilizes heterogeneous annotations.

Lin et al. [60] have presented a practical method of rapidly labeling semantic segmentation masks. Meanwhile, they have proven that their sparse confident labels can be efficiently propagated to generate segmentation proposals. Their experiments prove that the labels output by the semantic segmentation network are less inaccurate than those generated by considering the labels of highly relevant confident regions.

Instance segmentation [61] is a more difficult task involving the discrimination of each distinct instance of an object during segmentation. An RNN is used to sequentially delineate instances. This architecture, which includes a spatial memory, is trained in an end-to-end manner to sequentially segment single instances. Results achieved on multiple-person segmentation and leaf counting show that this method outperforms other approaches. Dai et al. [62] have demonstrated the relationship of logical progression among the differentiation of instances, the estimation of masks and the categorization of objects. Based on knowledge of instances, the features learned by a segmentation network can be made more effective. The learning processes for instance segmentation and semantic segmentation can be mutually beneficial.

It is a meaningful task to rapidly and effectively synthesize existing semantic segmentation networks from different domains into a current domain of interest. The cost of repeatedly training such networks for different categories and datasets is prohibitively high, but transferable learning provides a way to solve this problem. Segmentation annotations in the source domain can be used to train the decoder and attentional model, whereas image-level class labels in the source and target domains can be used to train the attentional model. Thus, it possible to share the information necessary for shape generation among different categories.

4 Conclusion and possible directions of future research

We have summarized and analyzed the leading contributions on the Pascal VOC 2012 semantic segmentation challenge, the underlying concepts of which are listed in Table 5. The widespread connections between features from different layers and from the same layer can be used to significantly improve the

Table 5 Underlying concepts of the summarized studies

Concept	Related structures
Feature encoder	VGG [31], ResNet [32]
Upsampling of low-resolution features or score maps	Unpooling layers [25,36], Deconvolution layers [25,36], Reconstruction [15]
Reduction of the resolution loss	Atrous and dilated convolution [26,27], Removing pooling layers [14], Shifting input and interlacing output [28], Multi-pass method [34]
Enhancement of features	Hypercolumns [35], Attentional model [23,41], Zoom-out [44], Context_CNN_CRF [19], CentraleSuperBoundaries [51]
Selection of features	DecoupledNet [45]
Step-by-step refinement of intermediate segmentations	Skip-layer architecture [36], Cascade-like structure [38], DeconvNet [25], DSN [50]
Utilization of heterogeneous annotations	BoxSup [55], DecoupledNet [45], Weakly and semi-supervised learning [14]
Explicit propagation of context	DAG [22], LG-LSTM [24], PSPNet [48]
Learning of potentials	Context_CNN_CRF [19], GCRF [47], High-order potential CRF [21,46], DPN [20]
Solving of CRFs	CRF-RNN [18], DPN [20], Adelaide_Learning_Messages [21]

segmentation results. There is a degree of independence among these contributions, which can therefore be combined to further refine the results.

As shown in Figures 5 and 6, the state-of-the-art results produced by PSPNet [48] show remarkable performance in both the detection and localization of contours. Even in cases of varying posture, appearance and occlusion, almost all objects in the test images can be recognized. However, some failures occur as a result of the absence of sufficient context, crucial textural distinctions or in-depth functional analysis.

Despite the incredible breakthroughs achieved to date, we believe that some problems remain to be addressed:

(1) More effective strategies for extracting features.

To further improve the results, the extracted features should be more effective, robust and meaningful. In the case of specific semantic categories, more general selection policies for constructing multi-granular representations of features related to specific semantic objects are anticipated.

(2) Consideration of contextual connections throughout the entire framework.

CRFs and RNNs are implemented as post-processing components to perform joint inference with contextual information that is restricted to the output layer of the feature encoder. The connections that exist among different levels could be used to further reduce ambiguity. In addition, using general CNNs to represent context would also be a meaningful contribution.

(3) Fuller utilization of weakly and semi-supervised annotations.

Human beings have the ability to identify the outlines of new objects that they have never seen before, not merely by virtue of binocular stereo vision, which suggests the feasibility of unsupervised algorithms for segmenting foregrounds from images. We believe that it should be possible to simulate human vision by utilizing known semantic knowledge to detect objects and distinctions among features to trace contours.

Based on the recent progress reviewed here, we summarize the general framework for image segmentation in Figure 7. Below, we recommend several potential approaches to combining or enhancing current studies:

- Multi-granular information could be combined more efficiently as pixel-wise descriptions;
- Multi-scale and multi-granular information could be treated as inputs to implementations of fully connected CRFs;
- DSN-like architectures could be expanded to guide the construction of not only class-specific features but also context-specific features, among others;

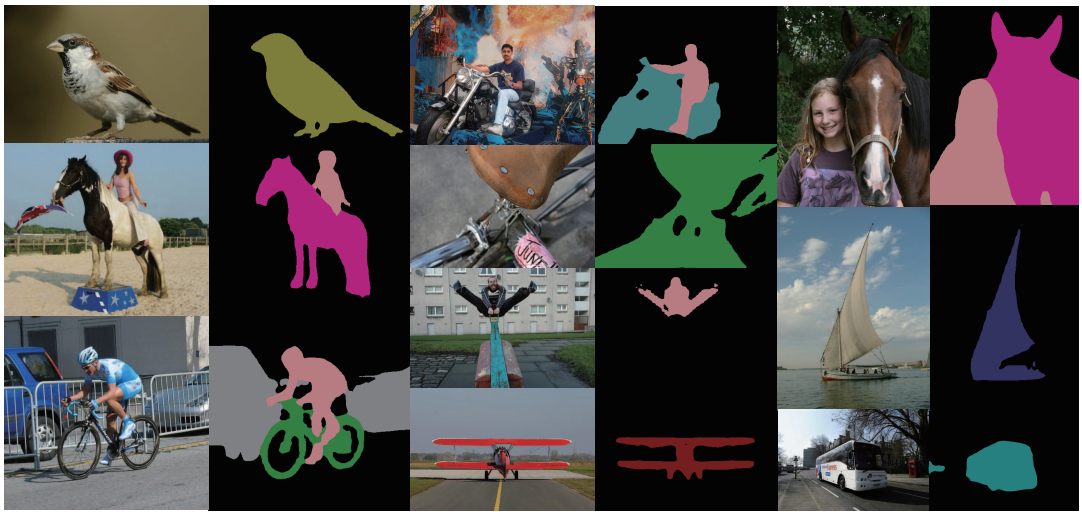


Figure 5 (Color online) Remarkable results produced by PSPNet [48].

Figure 6 (Color online) Failure cases of PSPNet [48].

• Weakly and semi-supervised annotations could be introduced to enhance the power to recognize and localize semantic objects.

Based on the above discussion, we present four possible architectures that have not been tested on the test set:

Zoom-out + CRF-RNN. Utilizing CRFs and MRFs to classify and smooth super-pixels is a conventional strategy. This idea can be applied to the present CNN architecture. The features extracted using the zoom-out approach can be treated as inputs to inference layers in a CRF-RNN to produce smoothed results.

BoxSup + DSN. We substitute a DSN-like architecture for the CNN architecture. Hence, the parameters can be updated more efficiently. We propose this architecture based on the consideration that more meaningful features can accelerate the convergence of the iterative process.

DSN + Multi-scale input + Hypercolumns + CRF-RNN. We propose this design to enhance the effectiveness and robustness of the components listed above to the greatest possible extent. Hypercolumns form the feature vector for each pixel, for which elements are obtained from the DSN-like

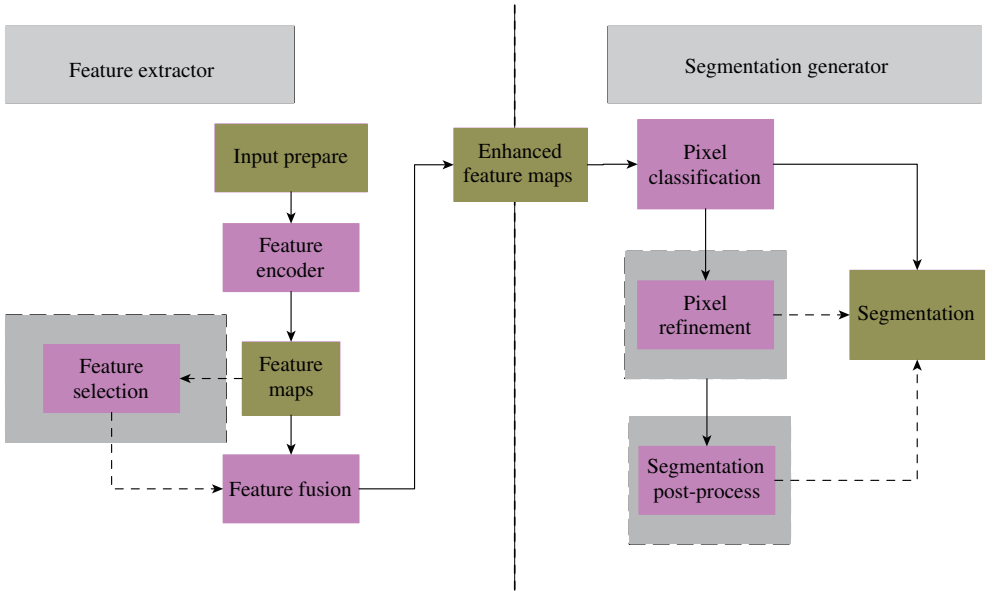


Figure 7 (Color online) The general framework for image segmentation. We divide it into two components: feature extractor and segmentation generator. The green boxes represent vital data used throughout the entire framework; the pink boxes represent functions used to process different data, where those presented with gray frames are optional structures. Different connections represent different paths.

architecture fed with multi-scale input. The subsequent CRF-RNN implementation produces smoothed results.

DSN + Cascade-like structure. Cascade-like structures can be used to refine results from a low resolution to a high resolution in a step-by-step manner. This strategy involves incorporating new feature levels at every step. Meanwhile, a DSN can supervise every step to ensure the quality of the intermediate segmentations.

We hope that further research will be conducted based on the architectures proposed above to continue to improve the performance achieved on the PASCAL VOC 2012 challenge. With the synthesis of powerful workflows and components, we trust that the current state-of-the-art results in semantic image segmentation can yet be surpassed.

Acknowledgements This work was supported by National High-tech R&D Program of China (863 Program) (Grant No. 2015AA016403) and National Natural Science Foundation of China (Grant Nos. 61572061, 61472020).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Liang G Q, Ca J N, Liu X F, et al. Smart world: a better world. Sci China Inf Sci, 2016, 59: 043401
2 Wang J L, Lu Y H, Liu J B, et al. A robust three-stage approach to large-scale urban scene recognition. Sci China Inf Sci, 2017, 60: 103101
3 Wang W, Hu L H, Hu Z Y. Energy-based multi-view piecewise planar stereo. Sci China Inf Sci, 2017, 60: 032101
4 Hoiem D, Efros A A, Herbert M. Recovering surface layout from an image. Int J Comput Vis, 2007, 75: 151–172
5 Saxena A, Sun M, Ng A Y. Make3d: learning 3d scene structure from a single still image. IEEE Trans Patt Anal Mach Intell, 2009, 31: 824–840
6 Gould S, Fulton R, Koller D. Decomposing a scene into geometric and semantically consistent regions. In: Proceedings of the IEEE International Conference on Computer Vision, Kyoto, 2009. 1–8
7 Gupta A, Efros A A, Hebert M. Blocks world revisited: image understanding using qualitative geometry and mechanics. In: Proceedings of European Conference on Computer Vision, Crete, 2010. 482–496
8 Zhao Y B, Zhu S C. Image parsing via stochastic scene grammar. In: Proceedings of the Conference and Workshop on Neural Information Processing System, Granada, 2011. 73–81
9 Liu C, Yuen J, Torralba A. Nonparametric scene parsing via label transfer. IEEE Trans Patt Anal Mach Intell, 2011, 33: 2368–2382

10 Stella X Y, Zhang H, Malik J. Inferring spatial layout from a single image via depth-ordered grouping. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Anchorage, 2008
11 Lee D C, Hebert M, Kanade T. Geometric reasoning for single image structure recovery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, 2009. 2136–2143
12 Zheng Y, Byeungwoo J, Xu D, et al. Image segmentation by generalized hierarchical fuzzy C-means algorithm. J Intell Fuzzy Syst, 2015, 28: 4024–4028
13 Liu C, Yuen J, Torralba A. SIFT flow: dense correspondence across scenes and its applications. IEEE Trans Softw Eng, 2010, 33: 978–994
14 Papandreou G, Chen L C, Murphy K P, et al. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1742–1750
15 Ghiasi G, Fowlkes C C. Laplacian pyramid reconstruction and refinement for semantic segmentation. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 519–534
16 Peng C, Zhang X Y, Yu G, et al. Large kernel matters—improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017. 4353–4361
17 Everingham M, van Gool L, Williams C K I, et al. The Pascal visual object classes (VOC) challenge. Int J Comput Vis, 2010, 88: 303–338
18 Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1529–1537
19 Lin G S, Shen C H, van den Hengel A, et al. Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3194–3203
20 Liu Z W, Li X X, Luo P, et al. Semantic image segmentation via deep parsing network. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1377–1385
21 Lin G S, Shen C H, Reid I, et al. Deeply learning the messages in message passing inference. Comput Sci, 2015, 71: 866–872
22 Shuai B, Zuo Z, Wang B, et al. Dag-recurrent neural networks for scene labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3620–3629
23 Kuen J, Wang Z H, Wang G. Recurrent attentional networks for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3668–3677
24 Liang X D, Shen X H, Xiang D L, et al. Semantic object parsing with local-global long short-term memory. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3185–3193
25 Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1520–1528
26 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. In: Proceedings of International Conference on Learning Representations, San Juan, 2016
27 Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. arXiv:1606.00915, 2016
28 Sermanet P, Fergus R, LeCun Y, et al. Overfeat: integrated recognition, localization and detection using convolutional networks. In: Proceedings of International Conference on Learning Representations, Banff, 2014
29 Zeiler M D, Fergus R. Visualizing and understanding convolutional networks. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 818–833
30 Krähenbühl P, Koltun V. Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Proceedings of Advances in Neural Information Processing Systems, Granada, 2011. 109–117
31 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations, San Diego. 2015
32 He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 770–778
33 Gao W, Zhou Z H. Dropout Rademacher complexity of deep neural networks. Sci China Inf Sci, 2016, 59: 072104
34 Wu Z F, Shen C H, Hengel A. High-performance semantic segmentation using very deep fully convolutional networks. arXiv:1604.04339, 2016
35 Hariharan B, Arbeláez P, Girshick R, et al. Hypercolumns for object segmentation and fine-grained localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 447–456
36 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 3431–3440
37 Xie S N, Tu Z W. Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1395–1403
38 Lin G S, Milan A, Shen C H, et al. RefineNet: multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017. 1925–1934
39 Wu Z F, Shen C H, Hengel A. Wider or deeper: revisiting the ResNet model for visual recognition. arXiv:1611.10080, 2016

40 Hong S, Oh J, Lee H, et al. Learning transferrable knowledge for semantic segmentation with deep convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3204–3212

41 Chen L C, Yang Y, Wang J, et al. Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3640–3649

42 Liu S, Qi X J, Shi J P, et al. Multi-scale patch aggregation (MPA) for simultaneous detection and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3141–3149

43 Bertasius G, Shi J, Torresani L. Semantic segmentation with boundary neural fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3602–3610

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

45 Hong S, Noh H, Han B. Decoupled deep neural network for semi-supervised semantic segmentation. In: Proceedings of Advances in Neural Information Processing Systems, Montreal, 2015. 1495–1503

46 Arnab A, Jayasumana S, Zheng S, et al. Higher order conditional random fields in deep neural networks. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 524–540

47 Vemulapalli R, Tuzel O, Liu M Y, et al. Gaussian conditional random field network for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3224–3233

48 Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 2017. 2881–2890

49 Yang J, Price B, Cohen S, et al. Object contour detection with a fully convolutional encoder-decoder network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 193–202

50 Lee C Y, Xie S, Gallagher P, et al. Deeply-supervised nets. In: Proceedings of Artificial Intelligence and Statistics, San Diego, 2015. 562–570

51 Kokkinos I. Pushing the boundaries of boundary detection using deep learning. In: Proceedings of International Conference on Learning Representations, San Juan, 2016

52 Giusti A, Ciresan D C, Masci J, et al. Fast image scanning with deep max-pooling convolutional neural networks. In: Proceedings of the 20th IEEE International Conference on Image Processing (ICIP), Melbourne, 2013. 4034–4038

53 Sutton C, Mccallum A. Piecewise training for undirected models. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence. Edinburgh: AUAI Press, 2005. 568–575

54 Adams A, Baek J, Davis M A. Fast high-dimensional filtering using the permutohedral lattice. Comput Graph Forum, 2010, 29: 753–762

55 Dai J F, He K M, Sun J. Boxsup: exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1635–1643

56 Rother C, Kolmogorov V, Blake A. Grabcut: interactive foreground extraction using iterated graph cuts. ACM Trans Graph, 2004, 23: 309–314

57 Uijlings J R R, van de Sande K E A, Gevers T, et al. Selective search for object recognition. Int J Comput Vis, 2013, 104: 154–171

58 Arbeláez P, Pont-Tuset J, Barron J, et al. Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 328–335

59 Krahenbhl P, Koltun V. Geodesic object proposals. In: Proceedings of European Conference on Computer Vision, Zurich, 2014. 725–739

60 Lin D, Dai J F, Jia J Y, et al. Scribblesup: scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3159–3167

61 Romera-Paredes B, Torr P H S. Recurrent instance segmentation. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 312–329

62 Dai J F, He K M, Sun J. Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 3150–3158