World Wide Web (2019) 22:555–570
https://doi.org/10.1007/s11280-018-0556-3

Multi-scale deep context convolutional neural networks for semantic segmentation

Quan Zhou[1,2] · Wenbing Yang[1,2] · Guangwei Gao[3] · Weihua Ou[4] · Huimin Lu[5] · Jie Chen[6] · Longin Jan Latecki[7]

Received: 16 August 2017 / Revised: 16 March 2018 / Accepted: 27 March 2018 /
Published online: 19 April 2018

Abstract Recent years have witnessed the great progress for semantic segmentation using deep convolutional neural networks (DCNNs). This paper presents a novel fully convolutional network for semantic segmentation using multi-scale contextual convolutional

This article belongs to the Topical Collection: *Special Issue on Deep vs. Shallow: Learning for Emerging Web-scale Data Computing and Applications*
Guest Editors: Jingkuan Song, Shuqiang Jiang, Elisa Ricci, and Zi Huang

✉ Quan Zhou
quan.zhou@njupt.edu.cn

✉ Weihua Ou
ouweihuahust@gmail.com

Wenbing Yang
1129924198@qq.com

Guangwei Gao
csggao@gmail.com

Huimin Lu
dr.huimin.lu@ieee.org

Jie Chen
nethuhost@163.com

Longin Jan Latecki
latecki@temple.edu

[1] National Engineering Research Center of Communications and Networking, Nanjing University of Posts, Telecommunications, Nanjing, China

[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China

[3] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science, Technology, Nanjing, China

[4] School of Big Data and Computer Science, Guizhou Normal University, Guiyang, China

features. Since objects in natural images tend to be with various scales and aspect ratios, capturing the rich contextual information is very critical for dense pixel prediction. On the other hand, when going deeper in convolutional layers, the convolutional feature maps of traditional DCNNs gradually become coarser, which may be harmful for semantic segmentation. According to these observations, we attempt to design a multi-scale deep context convolutional network (MDCCNet), which combines the feature maps from different levels of network in a holistic manner for semantic segmentation. The segmentation outputs of MDCCNets are further enhanced using dense connected conditional random fields (CRF). The proposed network allows us to fully exploit local and global contextual information, ranging from an entire scene to every single pixel, to perform pixel-wise label estimation. The experimental results demonstrate that our method outperforms or is comparable to state-of-the-art methods on PASCAL VOC 2012 and SIFTFlow semantic segmentation datasets.

Keywords Multi-scale context · MDCNNs · Semantic segmentation · CRF

1 Introduction

Image semantic segmentation is a classic and challenging visual task in the field of computer vision. It aims to assign a semantic label to each image pixel to achieve object recognition and segmentation tasks synchronously. Semantic segmentation provides complete understanding of the scene. It predicts the label, location, as well as accurate shape information for each image element. Therefore, this topic is of broad interest for potential real applications, such as automatic driving, robot navigation, and robot manipulation, etc. On the other hand, it is also associated with many high-level vision tasks, including image classification [15, 18, 39], edge detection [23], object recognition [11, 14, 42] and detection [5, 12, 30], and video segmentation and action recognition [36, 44].

In recent years, deep convolutional neural networks (DCNNs) have gained a lot of attention and become very popular in the computer vision community. Due to its superiority in modeling high-level visual concepts, DCNNs substantially advance the performance for the task of semantic segmentation [3, 4, 24, 49]. Specifically, the recent state-of-the-art semantic segmentation frameworks are mostly based on the fully convolutional network (FCN) [24], where the architecture of DCNN originally developed for image classification has been successfully repurposed for dense pixel prediction in the framework of end-to-end learning. When performing dense estimation using FCNs, a common architecture of DCNN is to successively reduce the spatial size of the feature maps using pooling operations and/or strided convolutions. Such operations significantly increase the size of the receptive field, which refers to the extent of data that are path-connected to a neuron [25]. Although the FCN-based methods boost the performance of semantic segmentation, they still face some challenges. The major limitation lies in the fact that the low resolution feature maps always lead to the loss of spatial statistics for object instance. In addition, the receptive fields is not

[5] Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Kitakyushu, Japan

[6] Huawei Technologies Co. Ltd., ShenZhen, China

[7] Department of Computer and Information Sciences, Temple University, Philadelphia, USA

adaptive when using the traditional FCN-based network, where a small receptive field may lead to inconsistent segmentation results on large objects while a large receptive field often ignores small objects and classifies them as background [28].

In order to further advance the performance of semantic segmentation, the context cues are widely employed for image semantic segmentation [1, 14, 28, 31, 48, 52]. For instance, Noh et al. [28] and Vijay et al. [1] proposed a coarse-to-fine structure with deconvolution network to learn the segmentation mask. Olaf et al. [31] employed a U-structure network to capture context cues by contracting pooling layers and corresponding deconvolution layers. Yu and Koltun [48] introduced dilated convolutions to reduce the effect of pooling in their pre-trained network. Zhao et al. [49] exploited the global context using pyramid parsing network. An alternative approach to investigate context clues is to adopt Conditional Random Fields (CRFs) as post-processing steps [4, 50]. In spite of achieving promising results, there are still two primarily important issues that need to be considered to integrate context cues in the fashion of FCN for semantic segmentation:

- How to learn a powerful representation to capture the wide variety visual context in a given scene?
- How to design a simple and efficient network to ensure a globally consistent semantic segmentation outputs?

Motivated by spatial pyramid pooling [14, 19], in this paper, we make an effort to address these two questions based on embedding multi-scale context information in the architecture of FCN. Precisely, a *multi-scale deep context convolutional network* (MDCCNet) is introduced that explores wide scale contextual cues using the feature maps from different convolutional layers. The main idea is inspired by two common observations: firstly, similar to previous FCN-based methods [7, 24], the spatial statistics of pixels is gradually lost with going deeper of convolutional layers; secondly, compared with shallower layers, the deeper ones have large receptive fields, which transmit large area of image content. Intuitively, the feature maps with different scales can provide sufficient spatial statistics and contextual cues. Our main contribution lies in direct integrating the feature maps from multiple convolutional layers. Thereafter, a fully connected CRF is applied to refine the segmentation result and achieve the delineated object boundaries. A high-level illustration of the proposed method is shown in Figure 1. Under this paradigm, the local and global contextual clues, ranging from an entire scene to every single pixel, are taken into account *jointly* to assign semantic label for each pixel. We compare the performance of our model with the mainstream models, which include traditional CRF-based contextual formulation [33, 52], FCN-based approaches [4, 7, 24, 26, 48], and deconvolution network [1, 28]. These are top-ranked models that previous studies have shown to significantly improve the results on semantic segmentation datasets. In summary, the main contributions of this paper are three folds:

- We propose a MDCCNet to capture multi-scale context features in the framework of FCN for dense pixel prediction.
- The outputs of MDCCNet are further enhanced using dense connected CRF, where the object shapes and boundaries are well rectified.
- The experimental results demonstrate that our method outperforms or is comparable to state-of-the-art models on PASCAL VOC 2012 [6] and SIFTFlow datasets [24].

The remainder of this paper is organized as follows. After a brief discussion of related work in Section 2, we describe our MDCCNet in Section 3. Implemented details and

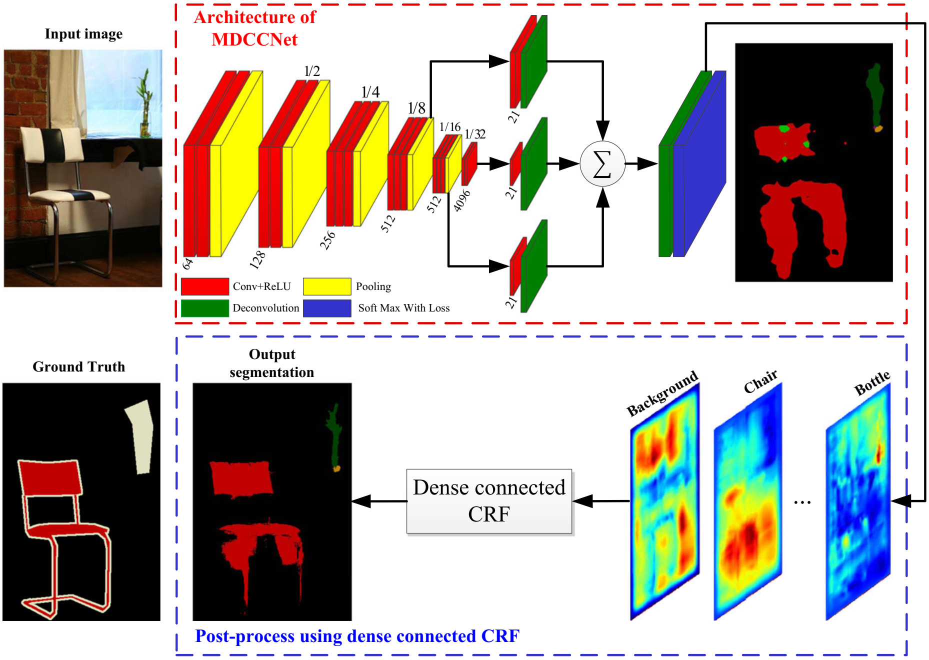


Figure 1 Overview of our semantic segmentation method. The whole pipeline of our approach is consist of two stage: MDCCNet construction and dense estimation using fully connected CRF as post-processing. Given an input image in first stage, we first use VGG-16 network to produce the hierarchical feature maps of intermediate pooling layers, which carry both local and global contextual cues and spatial statistics within different scales. Then the associated score maps are upsampled and concatenated to get the per-pixel prediction. In the second stage, the score maps of each category are fed into a dense connected CRF to achieve better delineated object boundaries. Note in the second stage, the score maps are represented by the heat maps, where red color denotes high probability, while blue color indicates low probability. (best viewed in color)

experimental results are given in Section 4. Finally, we give concluding remarks in Section 5.

2 Related work

In this section, we briefly review the advances in recent related work for semantic segmentation, which are roughly divided into three categories: generative graphical model, DCNN-based model, and FCN-based model.

Typically, the previous semantic segmentation models capture contextual cues using generative graphical models, such as CRFs [4, 33, 43] or markov random field (MRF) [2, 51], where the unary potential is formulated based on local image features, and pairwise potential is encoded based on short or long-ranged interactions. Precisely, for unary potential, the local appearance features, such as intensities, colors, gradients and textures, are first extracted to describe different object instances. Then these hand-craft features are fed into the well trained classifiers to identify the category label for each image pixel, including regression boosting [33, 41], random forests [32], or support vector machines (SVM) [8]. Recently, Gao et al. [10] utilize attention-based LSTM to investigate the global context for semantic consistency of video captioning. Song et al. [35] propose to explore the contextual cues from the partial available tags for image and video annotation. On the other hand, for pairwise potential, the contextual information is encoded in terms of intersections [33, 52] or top-down scene structured predictions [2, 43] to improve the global label consistency. However, the performance of these systems has always been compromised by the limited expressive power of the hand-craft features.

Recently, the most successful methods for semantic segmentation are based on DCNNs. Compared with graphical-based approaches, DCNN-based models have shown great potential and outstanding performance for the task of semantic segmentation. As the pioneer work, LeCun et al. [7] employ the DCNNs at multiple image resolutions to compute image features, resulting in smooth predictions using a segmentation tree as postprocess. Another elegant work was proposed in [12], where the bounding box proposals and masked regions are used as inputs to train a DCNN. In the stage of classification, object shape information is taken into account within the trained DCNN. Except the representation of bounding box proposals, the DCNN models can be also trained based on different image representation, such as superpixels [26]. In this work, the authors extracted the zoom-out spatial features, which are embedded into DCNNs to classify a superpixel. Although these methods can benefit from the delineated boundaries produced from a good segmentation, the object accurate shapes may be not always recovered well when there are some errors in segmentation results.

An alternative approach to investigate contextual clues relies on fully convolutional networks (FCN), where an end-to-end learning paradigm is adopted to train DCNNs [4, 24, 28]. Motivated from the DCNNs for image classification task, these methods directly target on estimating category-level per-pixel labels. The most representative work is [24], where the last fully connected layers of the DCNN are transformed into convolutional layers. In [4], Chen et al. proposed to learn a DeepLab model for semantic segmentation, where the receptive fields with different scales are employed using atrous convolution in deep convolutional networks. In [48], the authors append a series of dilated convolutional layers after a FCN backbone to expand receptive field. However, the repeated operation of max-pooling and striding at consecutive layers significantly reduces the spatial resolution of the resulting feature maps. In order to remedy such problem, Noh et al. [28] and Vijay et al. [1] proposed a coarse-to-fine structure with deconvolution network to learn the segmentation mask, yet with the cost of huge number of memory requirement and computing time. Unlike these approaches, our MDCCNet directly learns the context representation by integrating the intermediate feature maps, and upsampling the combined outputs to the original image resolution. This all-in-once fashion allows us to learn multi-scale context *jointly*, thus yielding more robust and reliable results.

Another two related works are [27] and [17]. In [27], the authors formulate the CRF as a DCNN layer, therefore, it is pluggable into any layer of a DCNN. However, our approach adopts a two-stage scheme, where the MDCCNet is first trained to generate rough semantic segmentation results, and then dense connected CRF is used for detailed rectification. In [17], the authors employ dilated convolution for decreasing network parameters, and explore all the convolutional feature maps to produce semantic outputs. On the contrary, our MDCCNet only integrates the feature maps from the deepest three pooling layers, and demonstrates it is enough to achieve good segmentation results.

An early version of this work was first published in [47]. This journal version extends previous one in two aspects: besides exploring mid-layer features to capture context information, we also employ dense connected CRF to encode short and long-ranged pixel-based interactions; we have implemented more complete experiments, and reported more comparisons and improved results.

3 Our method

In this section, we first elaborate on architectural details of our MDCCNet to investigate the multi-scale context information, and then give the details to further capture contextual information in terms of pixel-wised interaction using dense connected CRF.

3.1 The architecture of MDCCNet

As shown in the upper panel of Figure 1, we define our MDCCNet based on VGG 16-layer network, which consists of four basic components, including convolution, rectified linear unit (ReLU), pooling (downsampling) and deconvolution (upsampling). In the main stream structure, we borrow the network architecture widely used in FCN-based model [24], where the final fully connected layers are transfered to convolution layers. It consists of the repeated convolution with 3 × 3 filter kernels, followed by a ReLU and a 2 × 2 max pooling operation with stride 2 for downsampling. The resolution of feature maps are reduced to $1/2^N$ of the original one after N max pooling operation. Here we set $N = 5$, leading to the 1/2, 1/4, 1/8, 1/16 and 1/32 of the original resolution, and denote the final three ones as MDCCNet-8s, MDCCNet-16s, MDCCNet-32s, respectively. The detail hyper parameter setting and each side-output layer of our main stream structure are summarized in Table 1.

Similar to previous FCN-based approaches [4, 24, 28], the major limitation of our main stream structure lies in the spatial statistics of pixels is gradually lost with going deeper of our MDCCNet. On the other hand, compared with shallower layers, the deeper ones have large receptive fields and are able to see more pixels. Intuitively, the feature maps with different scales can provide sufficient spatial statistics and contextual cues, which complement each other to get more reliable predictions. To this end, we integrate the main stream with two additional streams from the max pooling layers of MDCCNet-8s and MDCCNet-16s, as shown in the upper panel of Figure 1. Once augmented, MDCCNet allows us to fuse predictions from three streams that are learned jointly in an end-to-end architecture. More specifically, the final layers of these three streams are first convoluted with a 1 × 1 filter kernel to map them to three score maps, representing the confidence for each individual classes. However, these layers are with different resolution, they are thus required to be aligned by scaling and cropping. The deconvolution layers in our MDCCNet utilizes

Table 1 Details of hyper parameter setting and each side-output. $(c, k \times k)$ means that there are c convolutional channels using the filter kernels with size $k \times k$

Cov Layer No.	cov1	cov2	cov3	cov4	cov5	cov6
Side output	$(64, 3 \times 3)$	$(128, 3 \times 3)$	$(256, 3 \times 3)$	$(512, 3 \times 3)$	$(512, 3 \times 3)$	$(4096, 3 \times 3)$
Cov Channels	2	2	3	3	3	2
stride	1	1	1	1	1	1

the nonlinear upsampling to scale score maps to the resolution with respect to MDCCNet-8s, where the upsampling filter kernels are learned with the initialized weights of bilinear interpolation [24]. Subsequently, a cropping operation is performed. Cropping removes any portion of the upsampled layer which extends beyond the other layer, resulting in layers of equal dimensions for exact fusion. Finally, the three score maps are fused to a single score map, which is further upsampled to obtain the semantic output with the same resolution of original image. In Figure 2, we illustrate the outputs of MDCCNet of some visual examples on PASCAL VOC 2012 validation set [6].

At first glance of our MDCCNet, it might look similar to the skip version of FCN [24], but in fact their network architectures are quite different. Essentially, FCN relies on *gradually* learning finer-scale prediction from lower layers in a stage-by-stage manner. That is, in each stage, the net is trained using the initialization of the previous stage net, where the contextual features are explored *independently*. In contrast, our MDCCNet employs a all-in-once fashion to fuse the computed intermediate feature maps, where the contextual cues are investigated *jointly* to make final estimation. In addition, compared with stage-by-stage scheme used in FCN model, our all-in-once fashion results in computational efficiency and is less tedious in training process.

3.2 Dense connected CRF enhancement

As illustrated in Figure 2, our MDCCNet is able to estimate the presence and the rough position of objects, yet their accurate borders are not well delineated. In this section, we employ the dense connected CRF framework [43] to solve the fine-grained localization problem, yielding accurate semantic segmentation results and recovering object boundaries at a detail level.

Let $\boldsymbol{x} \in \mathcal{X}$ denotes the input image and $\boldsymbol{y} \in \mathcal{Y}$ is the labeling results which describe the label configuration of each pixel in the dense connected CRF graph. Our goal is to find the optimal label configurations $\boldsymbol{y}^*$ that minimizes the following energy function:

$$E(\boldsymbol{y}, \boldsymbol{x}; \theta) = \sum_i E_i(y_i, x_i; \theta_i) + \sum_{ij} E_{ij}(y_i, y_j, x_i, x_j; \theta_{ij}) \quad (1)$$

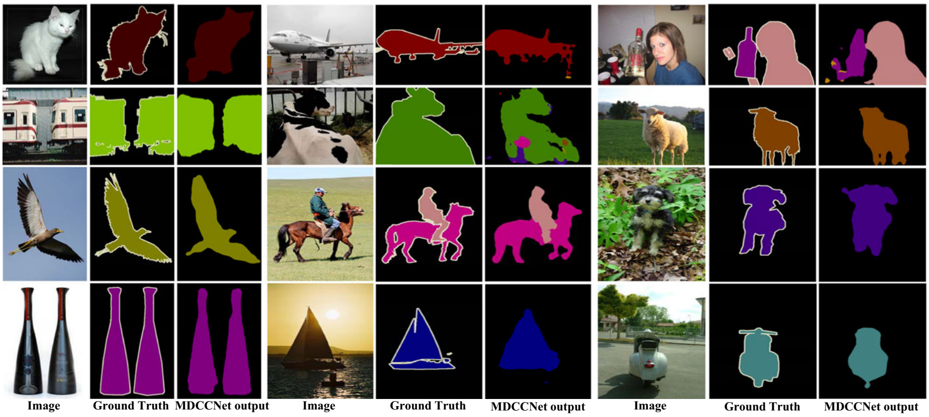


Figure 2 Some visual examples of semantic segmentation results on PASCAL VOC 2012 validation set using our MDCCNet, where different semantic categories are coded using different color. (best viewed in color)

where $\theta = \{\theta_i, \theta_{ij}\}$, and the first and second terms are unary potential and pairwise potential, respectively. More specifically, the unary potential energy function is defined as:

$$E_i(y_i, x_i; \theta_i) = -\log p(y_i, x_i; \theta_i) \tag{2}$$

where $p(y_i, x_i; \theta_i)$ is the label assignment probability (denoted as heat map in Figure 1) for pixel $x_i \in \boldsymbol{x}$ and the parameters θ_i is trained using our MDCCNet. According to [43], the dense connected graph requires each image pixel pair x_i and x_j has to be connected, leading to a huge number of connected edges in our CRF formulation. In order to perform efficient inference, the Potts model is adopted in our pairwise potential, which has the form as follows:

$$E_{ij}(y_i, y_j, x_i, x_j; \theta_{ij}) = \delta(y_i, y_j)\left[\exp\left\{-\frac{|x_i - x_j|^2}{2\sigma_\alpha^2}\right\} + \lambda \exp\left\{-\frac{|p_i - p_j|^2}{2\sigma_\beta^2}\right\}\right] \tag{3}$$

where $\theta_{ij} = \{\sigma_\alpha, \sigma_\beta, \lambda\}$. $\delta(y_i, y_j)$ is a characteristic function that penalizes the nodes with distinct labels as $\delta(y_i, y_j) = 1$ if $y_i \neq y_j$, and zero otherwise. The remaining expression uses two Gaussian kernels in RGB color space and position space. Preciously, the first kernel constrains that the connected pixels with similar color appearance tend to be assigned with same category label. Likewise, the second kernel considers spatial proximity when enforcing smoothness. Intuitively, the pixels with short distance are prone to be with same semantic label, while different labels should be assign to the pixels when they are far away. The hyper parameters σ_α and σ_β control the scale of two Gaussian kernels, and λ is a trade-off parameter to balance the energy cost in in RGB color space and position space.

The optimal solution $\boldsymbol{y}^*$ of our CRF model can be achieved using efficient mean field approximation inference [43], where a high dimensional filtering algorithm significantly speeds up the computation in message passing, resulting in less that 0.5 second on average to parse a Pascal VOC image.

4 Experimental evaluation

The purpose of our experiments is to evaluate the effectiveness of our method, and better understand the behavior of our system. We have conducted several experiments on PASCAL VOC 2012 [6] and SIFTFlow datasets [7, 24].

4.1 Dataset descriptions

The PASCAL VOC 2012 dataset [6] contains 21 category classes, including 20 categories for foreground object classes and additional one class for background. The original dataset contains 1464, 1449, and 1456 images for training, validation, and testing, respectively, where each image in the training and validation subset has accurate pixel-level annotated ground truth. The dataset is augmented by the extra annotations [13], resulting in 10582 training images.

The SIFT flow dataset [7, 24] is composed of 2688 images, that have been thoroughly labeled by LabelMe users. Most images are with resolution of 256 × 256. The authors used synonym correction to obtain 33 semantic categories. It is also a fully annotated dataset, most of which are outdoor scenes including street, beach, mountains and fields. The pixels labeled as "unlabeled" class are not considered during the training and testing for direct comparison.

4.2 Evaluation metrics

We evaluate our labeling models based on the following widely-used criteria, named pixel accuracy, class average accuracy, and mean intersection over Union (mIoU). Let N_{mn} be the number of pixels of category m labeled as class n, where there are $\mathcal{C}$ different object classes, then the three evaluation metrics are defined as follows.

Pixel accuracy pays the most attention to frequently occurring objects and penalizes infrequent objects. It refers to overall accuracy among all categories:

$$\frac{\sum_m N_{mm}}{\sum_{m,n} N_{mn}} \tag{4}$$

Average accuracy evaluates the recognizable accuracy per category:

$$\frac{1}{\mathcal{C}} \frac{\sum_m N_{mm}}{\sum_n N_{mn}} \tag{5}$$

mIoU is always used to penalize both over- and under-segmentation for scene labeling, which is defined as the ratio of true positives to the sum of true positive, false positive and false negative, averaged over all object classes:

$$\frac{1}{\mathcal{C}} \frac{\sum_m N_{mm}}{\sum_n N_{mn} + \sum_n N_{nm} - N_{mm}} \tag{6}$$

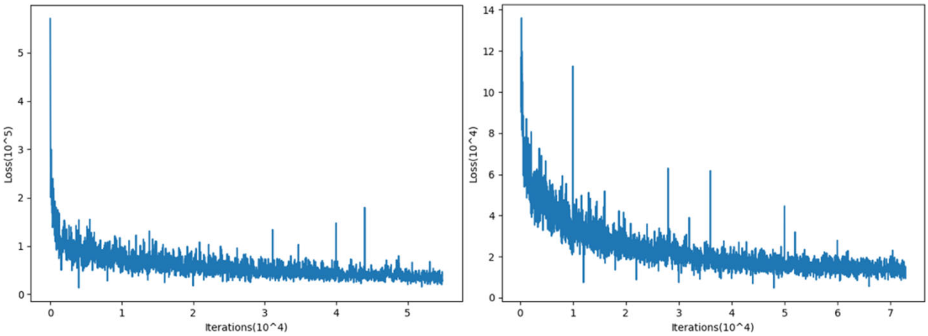
4.3 Implementation details

We use the split setting criteria of [4] for PASCAL VOC datasets, where our MDCCNet is first trained using 10582 training subset, and then augmented trained using 12031 images from tra-val subset. While for SIFT flow dataset, we use the evaluation procedure introduced in [7]: 2488 images used for training and 200 images used for testing.

Specifically, our implementation is based on the public platform Caffe [16] using the hardware with Intel Xeon E5-2680 CPU at 2.4GHz and 160GB memory and NVIDIA Tesla P40 GPU with 12GB memory. We minimize the soft max loss averaged over all image positions with stochastic gradient descent algorithm. The parameters of initialized network are borrowed from the pre-trained VGG-16 model using ImageNet dataset [34]. Then we fine-tune our MDCCNet model using training images of PASCAL VOC 2012 and SIFT-Flow dataset, respectively. Inspired by [4], we use the "poly" learning rate policy where the learning rate γ in iteration T equals to the base B multiplied by a factor:

$$\gamma = B \cdot \left(1 - \frac{T}{M}\right)^{power} \tag{7}$$

where M denotes the total number of training iterations. The base learning rate is set as $B = 0.01$ and power is set as 0.9. Figure 3 plots the curves of loss function as the iteration number increases on PASCAL VOC and SIFTFlow dataset, where some drastic fluctuations of loss function are observed, especially on SIFTFlow dataset. This is probably because the small batchsize (set as 1 in our experiments) is adopted to train our MDCCNet. If a training image contains a class that rarely appears in previous training images, the current trained network may have bad predictions according to the corresponding ground truth, resulting in unstable fluctuations in Figure 3. Even so, we can observe that the loss function is eventually convergent on two datasets, indicating that our MDCCNet is well trained. The performance of our MDCCNet can be further improved by increasing the iteration number. We found that



(a) PASCAL VOC 2012 dataset

(b) SIFTFlow dataset

Figure 3 The convergence curve of our MDCCNet learning process on (**a**) PASCAL VOC 2012 and (**b**) SIFTFlow dataset.(best viewed in color)

our best performance is achieved when total iteration number is set to $M = 20K$, and any refinement of this parameter will result in no more significant improvement of performance.

After the MDCCNet has been fine-tuned and pre-trained, we cross-validate the CRF parameters following [43]. 100 images from validation set of PASCAL VOC dataset are used to search best values of σ_α, σ_β, and λ defined in (3). In practice, we employ a switch-fixed searching scheme, where one parameter is updated, while the others are fixed. Three hyper parameters are initialized as $\sigma_\alpha = 30$, $\sigma_\beta = 1$, and $\lambda = 3$, and then updated with steps of 10, 1, 1, respectively. The searching procedure for each parameter terminates after 10 mean field iterations. The searched best values of σ_α, σ_β, and λ are directly used for SIFTFlow dataset.

4.4 Results on PASCAL VOC 2012 dataset

We report the results in Table 2, and compare with the baselines in terms of mIoU. The results clearly demonstrate that our MDCCNet outperforms prior state-of-the-art methods, including FCN-based models [4, 22, 24, 26, 48], deconvolution network [28], and the networks using contextual formulation [20, 50]. Only using PASCAL VOC 2012 training subset, our MDCCNnet achieves 71.4% mIoU among all 21 categories. Following the setting in recent work [4, 24], we then demonstrate that our method scales nicely when augmenting the number of training images from the trainval subset. The results are shown as *MDCCNet*† in Table 2. It is observed that our performance improves 1.7% mIoU. This is probably because more training data is able to provide more richer context cues to train our MDCCNet, where the conclusion is also consistent with the observations of [4, 20]. After applying the dense connected CRF method [43] for boundary refinement, our final model gains further improvement, achieving the highest accuracy on the classes of “bird”, “boat”, “bottle”, “car”, “cat”, “chair”, “dog”, “person”, “sheep”, “train” and “TV”. This, once again, demonstrates that adding our pairwise potentials to encode pixel-based interactions brings significant improvement, for which we achieve the best performance in PASCAL VOC 2012 dataset.

It is intriguing that the results of MDCCNet† are superior to the existing methods [4, 50] that employ CRF for further improving performance. This indicates our MDCCNet is able to capture wide scale context clues, allowing us to predict more accurate object localizations. Another interesting result is our approach also outperforms deconvolution network [28].

Table 2 Individual category results on the PASCAL VOC 2012 test set in terms of mIoU scores

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow
FCN [24]	91.2	76.8	34.4	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5
DCM [48]	90.7	82.2	37.4	72.7	57.1	62.7	82.8	77.8	78.9	28.0	70.0
DZF [26]	89.8	85.6	37.3	83.2	62.5	66.0	85.1	80.7	84.9	27.2	73.2
DLN [4]	87.5	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1
CRFRNN [50]	–	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2
CDN [28]	92.7	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0
DPN [22]	–	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9
DCSM [20]	–	90.6	37.6	80.0	67.8	74.4	92.0	85.2	86.2	39.1	81.2
MDCCNet	93.2	84.1	39.0	82.1	67.7	78.4	87.4	83.4	85.8	38.2	77.2
MDCCNet†	94.6	85.2	41.0	83.4	69.4	80.4	89.5	85.1	87.1	40.3	78.5
MDCCNet†+CRF	95.4	87.6	43.7	85.3	72.3	83.0	91.7	86.5	89.9	43.8	80.5

Method	table	dog	horse	mbike	person	planet	sheep	sofa	train	tv	mIoU
FCN [24]	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	67.2
DCM [48]	51.6	73.1	72.8	81.5	79.1	56.6	77.1	49.9	75.3	60.9	67.6
DZF [26]	57.5	78.1	79.2	81.1	77.1	53.6	74.0	49.2	71.7	63.3	69.6
DLN [4]	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRFRNN [50]	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
CDN [28]	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
DPN [22]	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
DCSM [20]	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
MDCCNet	45.6	80.9	75.2	77.3	82.8	56.7	81.5	51.8	82.4	70.5	71.4
MDCCNet†	47.8	82.2	76.9	79.0	84.5	58.8	83.2	53.3	84.2	72.1	73.1
MDCCNet†+CRF	50.6	84.2	79.7	81.0	86.6	61.5	85.7	55.6	86.3	74.8	75.5

The underline indicates the best performance among all approaches for each category

This is probably because the proposed MDCCNet has more powerful generalization ability than [28] due to its simple network architecture.

Since the ground truth labels are not available for the test images of PASCAL VOC dataset, we evaluate our MDCCNet on validation subset to get the qualitative results. Some visual pleasing results of simultaneous recognition and segmentation are shown in Figure 4, including the results before and after CRF enhancement, and the comparison with baseline approaches [4, 24]. Each example shows both the original image and the color coded output. Except the boundary pixels that exhibit relative higher confusion, nearly all pixels are correctly classified. It is evident that our MDCCNet can handle large appearance variations of object classes ("person", "table", and "chair", etc.) and efficiently prohibit the clutter background. The visualization results obtained before CRF already yields excellent segmentation results, while employing the CRF further improves the performance by removing false positives and refining object boundaries.

4.5 Results on SIFTFlow dataset

We then demonstrate that our method scales nicely when augmenting the number and classes on SIFTFlow dataset in Table 3. Compared with PASCAL VOC dataset, due to

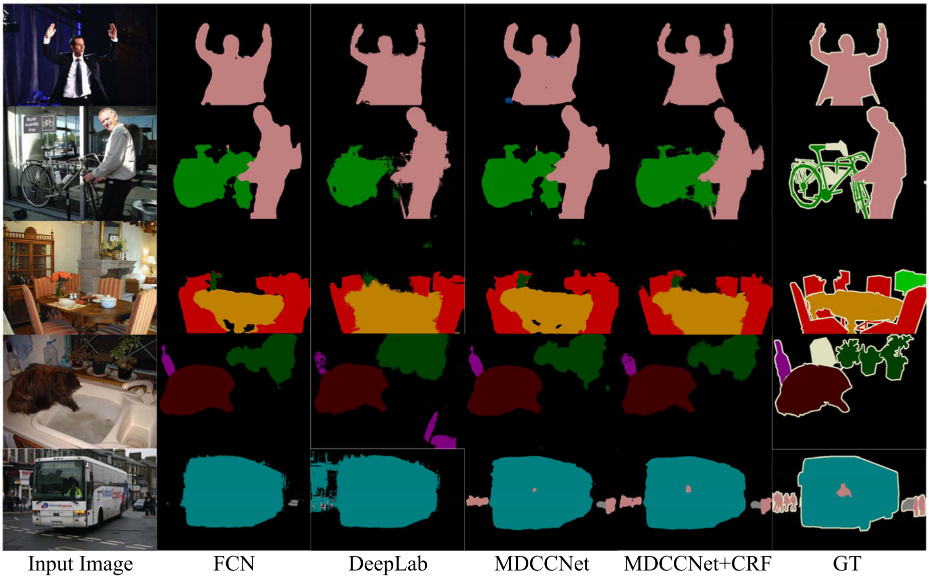


Figure 4 The comparison of some estimation examples of our method on the PASCAL VOC 2012 val dataset. (best viewed in color)

the small number of training data, our method only achieves 44.5% mIoU, and small improvement (0.7% mIoU) by employing CRF. Even so, it still demonstrates the superior performance of the proposed approach in comparison to the state-of-the-art methods. A remarkable fact is that our approach also outperforms the recent FCN-based method [24], and achieves comparable results to the latest FCN-CRF based approach such as DCSM [20]. Some qualitative results are exhibited in Figure 5. We also observe that object boundaries, for instance, "building", "car", and "window", are well recovered and delineated using CRF

Table 3 Segmentation results on SIFTFlow dataset (33 classes)

Method	Pixel accuracy	Mean accuracy	mIoU
Liu et al. [21]	76.7%	–	–
Tighe et al. [40]	75.6%	41.1%	–
Tighe et al.(MRF) [40]	78.6%	39.2%	–
Farabet et al.(natural) [7]	72.3%	50.8%	–
Farabet et al.(balance) [7]	78.5%	29.6%	–
Pinheiro et al. [29]	77.7%	29.8%	–
FCN [24]	85.9%	53.9%	41.2%
DCSM [20]	88.1%	53.4%	44.9%
MDCCNet	87.7%	54.2%	43.8%
MDCCNet+CRF	88.5%	55.6%	44.5%

The underline indicates the best performance among all approaches for each evaluation metric

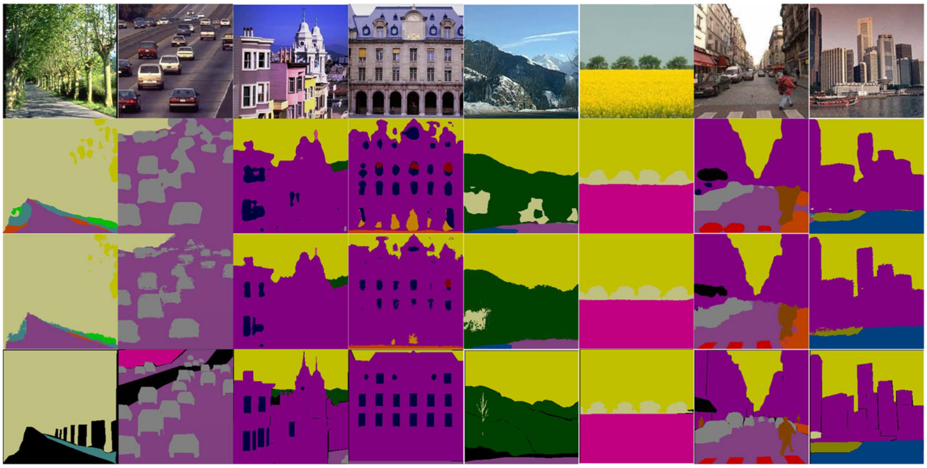


Figure 5 Some visual examples of semantic segmentation results on SIFTFlow dataset. From top to bottom are original images, MDCCNet output, MDCCNet+CRF results, and corresponding ground truth. (best viewed in color)

as post process. In addition, it is interesting that our method can provide accurate estimation for some generic categories, such as the "road" as shown in the first example, while the corresponding annotation is missing in ground truth.

4.6 Other aspect

One important factor is how the performance is affected by the intermediate feature maps to explore multi-scale context. In Table 4, we identify the contribution by sequentially adding the middle level layers in our MDCCNet. As can be seen, more levels of intermediate feature maps result in higher accuracy, which shows that using multi-scale context plays the most critical role in dense pixel classification. We also consider to use shallow layers of feature maps, and found it gives negligible improvement without refining the visual quality of outputs. Figure 6 shows the score maps per each category from MDCCNet-8s, MDCCNet-16s, MDCCNet-32s, and the combination thereof, respectively. It illustrates that the performance is dominant by the results of MDCCNet-32s, which is consistent with the the conclusion of Table 4.

Table 4 Results by gradually adding intermediate feature maps to capture multi-scale context on PASCAL VOC 2012 test dataset

Method	Pixel accuracy	Mean accuracy	mIoU
MDCCNet-32s	66.2%	52.3%	60.5%
MDCCNet-32s + MDCCNet-16s	78.6%	67.1%	66.8%
MDCCNet-32s + MDCCNet-16s + MDCCNet-8s	87.7%	73.6%	71.4%

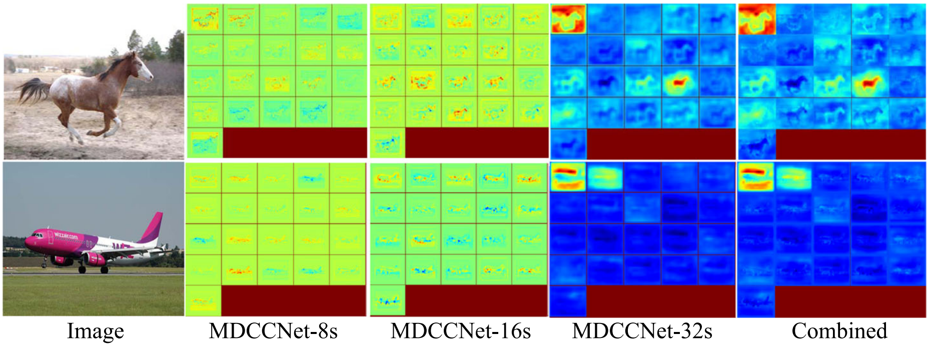


Figure 6 Side outputs of scores of two examples from PASCAL VOC validation set. The score maps are represented by the heat maps, where red color denotes high probability, while blue color indicates low probability (best viewed in color)

5 Conclusions and future work

This paper describes a MDCCNet to explore the multi-scale context information for semantic segmentation problem. Combining fine layers and coarse layers provides a more powerful representation with different receptive fields, allowing us to produce semantically accurate predictions and detailed segmentation maps. In order to further improve the performance, we also employ dense connected CRF to eliminate false positives and achieve delineated object shapes and boundaries. Our experimental results show that the proposed method outperforms or is comparable to state-of-the-art methods on PASCAL VOC 2012 and SIFTFlow semantic segmentation datasets.

In the future, we hope to demonstrate the generality of our method for other visual tasks and applications, such as video analysis [36, 44], image segmentation [9], video retrieval [37, 38] and cross-modal retrieval [45, 46].

Acknowledgements The authors would like to thank all the anonymous reviewers for their valuable comments and suggestions. This work was partly supported by the National Science Foundation (Grant No. IIS-1302164), the National Natural Science Foundation of China (Grant No. 61401228, 61402238, 61762021, 61571240, 61501247, 61501259, 61671253, 61402122), China Postdoctoral Science Foundation (Grant No. 2015M581841), Natural Science Foundation of Jiangsu Province (Grant No. BK20150849, BK20160908), Postdoctoral Science Foundation of Jiangsu Province (Grant No. 1501019A), Open Research Fund of National Engineering Research Center of Communications and Networking (Nanjing University of Posts and Telecommunications) (Grant No. TXKY17009), Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University) (Grant No. MJUKF201710), Open Fund Project of Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education (Nanjing University of Science and Technology) (Grant No. JYB201709, JYB201710), Natural Science Foundation of Guizhou Province (Grant No.[2017]1130), and the 2014 Ph.D Recruitment Program of Guizhou Normal University.

References

1. Badrinarayanan, V., Alex, K., Roberto, C.: SegNet: A deep convolutional encoder-decoder architecture for scene segmentation. IEEE TPAMI (2017)
2. Carreira, J., Sminchisescu, C.: Cpmc: Automatic object segmentation using constrained parametric mincuts. IEEE TPAMI. **34**(7), 1312–1328 (2012)

3. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: scale-aware semantic image segmentation. In: Proceedings of CVPR, pp. 3640–3649 (2016)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE TPAMI (2017)
5. Erhan, D., Szegedy, C., Toshev, A., Anguelov, D.: Scalable object detection using deep neural networks. In: Proceedings of CVPR, pp. 2147–2154 (2014)
6. Everingham, M., Eslami, S.A., Van, G.L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. IJCV **11**(1), 98–136 (2015)
7. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE TPAMI. **35**(8), 1915–1929 (2013)
8. Fulkerson, B., Vedaldi, A., Soatto, S.: Class Segmentation and Object Localization with Superpixel Neighborhoods. In: Proceedings of ICCV, pp. 670-677 (2009)
9. Gao, L.L., Song, J.K., Nie, F.P., Zhou, F.H., Sebe, N., Shen, H.T.: Graph-Without-Cut: an ideal graph learning for image segmentation. In: Proceedings of AAAI, pp. 1188–1194 (2016)
10. Gao, L.L., Guo, Z., Zhang, H.W., Xu, X., Shen, H.T.: Video captioning with Attention-Based LSTM and semantic consistency. IEEE TMM. **19**(9), 2045–2055 (2017)
11. Girshick, R.: Fast R-Cnn. In: Proceedings of ICCV, pp. 1440–1448 (2015)
12. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of CVPR, pp. 580–587 (2014)
13. Hariharan, B., ArbelAez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Proceedings of ICCV, pp. 991–998 (2011)
14. He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE TPAMI. **37**(9), 1904–1916 (2015)
15. He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J.: Deep residual learning for image recognition. In: Proceedings of CVPR, pp. 770–778 (2016)
16. Jia, Y.Q., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of ACMMM, pp. 675–678 (2014)
17. Kamran, S.A., Sabbir, A.S.: Efficient yet deep convolutional neural networks for semantic segmentation. In: Arxiv (2017)
18. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of NIPS, pp. 1097–1105 (2012)
19. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: Proceedings of CVPR, pp. 2169–2178 (2006)
20. Lin, G.S., Shen, C.H., Van, D.H., Reid, I.: Exploring context with deep structured models for semantic segmentation. IEEE TPAMI (2017)
21. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE TPAMI. **33**(5), 978–994 (2011)
22. Liu, Z.W., Li, X.X., Luo, P., Loy, C.C., Tang, X.O.: Semantic image segmentation via deep parsing network. In: Proceedings of ICCV, pp. 1377–1385 (2015)
23. Liu, Y., Chen, M.M., Hu, X.W., Wang, K., Bai, X.: Richer convolutional features for edge detection. In: Proceedings of CVPR, pp. 5872–5881 (2017)
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE TPAMI. **39**(4), 640–651 (2017)
25. Long, J.L., Zhang, N., Darrell, T.: Do convnets learn correspondence? In: Proceedings of NIPS, pp. 1601–1609 (2014)
26. Mostajabi, M., Yadollahpour, P., Shakhnarovich, G.: Feedforward semantic segmentation with zoom-out features. In: Proceedings of CVPR, pp. 3376–3385 (2015)
27. Nguyen, K., Fookes, C., Sridharan, S.: Deep context modeling for semantic segmentation. In: Proceedings of WACV, pp. 56–63 (2017)
28. Noh, H., Hong, S., Han, B.Y.: Learning deconvolution network for semantic segmentation. In: Proceedings of ICCV, pp. 1520–1528 (2015)
29. Pinherio, R.C., Pedro, H.: Recurrent convolutional neural networks for scene parsing. In: Proceedings of ICML (2014)
30. Ren, S.Q., He, K.M., Girshick, R., Sun, J.: Faster R-Cnn: towards real-time object detection with region proposal networks. In: Proceedings of NIPS, pp. 91–99 (2015)
31. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of MICCAI, pp. 234–241 (2015)

32. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: Proceedings of CVPR, pp. 1–8 (2008)
33. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. IJCV **81**(1), 2–23 (2009)
34. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
35. Song, J.K., Gao, L.L., Nie, F.P., Shen, H.T., Yan, Y., Sebe, N.: Optimized graph learning using partial tags and multiple features for image and video annotation. IEEE TIP. **25**(11), 4999–5011 (2016)
36. Song, J.K., Gao, L.L., Puscas, M.M., Nie, F.P., Shen, F.M., Sebe, N.: Joint graph learning and video segmentation via multiple cues and topology calibration. In: Proceedings of ACM MM, pp. 831–840 (2016)
37. Song, J.K., Gao, L., Liu, L., Zhu, X., Sebe, N.: Quantization-based hashing: a general framework for scalable image and video retrieval. PR (2017)
38. Song, J.K., Zhang, H.W., Li, X.P., Gao, L.L., Wang, M., Hong, R.C.: Self-supervised video hashing with hierarchical binary auto-encoder. IEEE TIP (2018)
39. Szegedy, C., Liu, W., Jia, Y.Q., Sermanet, P., Reed, S., Anguelo, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of CVPR, pp. 1–9 (2015)
40. Tighe, J., Lazebnik, S.: Finding things: image parsing with regions and per-exemplar detectors. In: Proceedings of CVPR, pp. 3001–3008 (2013)
41. Tu, Z.W., Bai, X.: Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE TPAMI. **32**(10), 1744–1757 (2010)
42. Uijlings, J.R., Van, D.S., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV. **104**(2), 154–171 (2013)
43. Vladlen, K.: Efficient Inference in Fully Connected Crfs with Gaussian Edge Potentials. In: Proceedings of NIPS, pp. 4–10 (2011)
44. Wang, X., Gao, L., Wang, P., Sun, X., Liu, X.: Two-stream 3D convNet fusion for action recognition in videos with arbitrary size and length. IEEE Transactions on Multimedia (2017)
45. Xu, X., He, L., Shimada, A., Taniguchi, R.I., Lu, H.: Self-supervised video hashing with hierarchical binary auto-encoder. Neurocomputing **21**(3), 191–203 (2016)
46. Xu, X., Shen, F., Yang, Y., Shen, H.T., Li, X.L.: Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE TIP. **26**(5), 2494–2507 (2017)
47. Yang, W.B., Zhou, Q., Fan, Y.W., Gao, G.W., Wu, S.S., Ou, W.H., Lu, H.M., Cheng, J., Longin, J.L.: Deep context convolutional neural networks for semantic segmentation. In: Proceedings of CCCV (2017)
48. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv:1511.07122 (2015)
49. Zhao, H.H., Shi, J.P., Qi, X.J., Wang, X.G., Jia, J.Y.: Pyramid scene parsing network. arXiv:1612.01105 (2017)
50. Zheng, S., Jayasumana, S., Paredes, B.R., Vineet, V., Su, Z.Z., Du, D.L., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of ICCV, pp. 1529–1537 (2015)
51. Zhou, Q., Zhu, J., Liu, W.Y.: Learning dynamic hybrid Markov random field for image labeling. IEEE TIP. **22**(6), 2219–2232 (2013)
52. Zhou, Q., Zheng, B.Y., Zhu, W.P., Latecki, L.J.: Multi-scale context for scene labeling via flexible segmentation graph. PR **2016**(59), 312–324 (2016)