Pattern Recognition 122 (2022) 108290

Contents lists available at ScienceDirect

Pattern Recognition

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

Contextual ensemble network for semantic segmentation

Quan Zhou[a,*], Xiaofu Wu[a], Suofei Zhang[b], Bin Kang[b], Zongyuan Ge[c], Longin Jan Latecki[d]

[a] National Engineering Research Center of Communications and Networking, Nanjing University of Posts and Telecommunications, Nanjing 21003, China
[b] Department of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing 21003, China
[c] Monash eResearch Centre, Monash University, Melbourne, Australia
[d] Department of Computer and Information Science, Temple University, Philadelphia, USA

ARTICLE INFO

Article history:
Received 6 November 2020
Revised 9 March 2021
Accepted 5 April 2021
Available online 20 September 2021

Keywords:
Ensemble deconvolution
Semantic segmentation
FCNs
Context aggregation
Encoder-decoder networks

ABSTRACT

Recently, exploring features from different layers in fully convolutional networks (FCNs) has gained substantial attention to capture context information for semantic segmentation. This paper presents a novel encoder-decoder architecture, called contextual ensemble network (CENet), for semantic segmentation, where the contextual cues are aggregated via densely usampling the convolutional features of deep layer to the shallow deconvolutional layers. The proposed CENet is trained in terms of end-to-end segmentation to match the resolution of input image, and allows us to fully explore contextual features through ensemble of dense deconvolutions. We evaluate our CENet on two widely-used semantic segmentation datasets: PASCAL VOC 2012 and CityScapes. The experimental results demonstrate our CENet achieves superior performance with respect to recent state-of-the-art results. Furthermore, we also evaluate CENet on MS COCO dataset and ISBI 2012 dataset for the task of instance segmentation and biological segmentation, respectively. The experimental results show that CENet obtains promising results on these two datasets.

1. Introduction

Semantic segmentation plays a significant role in computer vision, and thus is widely applied to many real-world scenarios, such as virtual/augmented reality, robotics, and self-driving. From the perspective of computer vision, the goal of semantic segmentation is to create vision systems with human-like abilities to achieve two fundamental tasks: classification and localization. As a result, a well-designed system for semantic segmentation should deal with these two issues simultaneously by assigning a unique semantic or categorical label to each image pixel.

The recent years have witnessed the substantial progress of image semantic segmentation using fully convolutional networks (FCNs) [1–3]. These models learn powerful contextual representations that lead to the successful results: a combination of feature descriptors extracted from FCNs are complementing each other to achieve remarkable improvement for semantic segmentation [4–6]. In spite of achieving promising results, previous FCNs suffer from a couple of critical limitations. Firstly, due to the consecutive pooling or convolution striding at successive layers, the spatial resolution is significantly reduced in feature maps. This invariance to local image transformation may be harmful for dense prediction tasks, where detailed spatial information is often required to delineate object shapes and boundaries [7,8]. Secondly, objects tend to be appear in multiple scales. However, the receptive field of early FCNs is not adaptive, leading to the problem that objects substantially larger or smaller than the receptive field may be fragmented or incorrectly classified [2,9].

In order to overcome these two challenges, the encoder-decoder networks and their variants (as shown in Fig. 1(a) and (b)) were proposed in recent literature [8,10,11]. In Fig. 1(a), the encoder-decoder networks consist of two parts. Like FCNs, the encoder gradually reduces the spatial dimension of feature maps. Conversely, the decoder progressively increases spatial dimension to recover object details using upsampling and deconvolution. In order to facilitate deconvolution, the indices of max-pooling are recorded in the process of encoder [12,13]. However, these architectures still suffer from the following shortcomings. Firstly, using pooling indices is storage expensive, where a large amount of memory spaces are required to save these indices. Secondly, although pooling indices are recorded to perform upsampling, the remaining elements are padded as zero [12], which may induce noise to the forthcoming deconvolution. Finally, although such structure is able to sequentially recover feature resolution, the contextual features of mid-level convolutional layers are often neglected. To improve segmentation performance, as illustrated in

[*] Corresponding author.
E-mail address: quan.zhou@njupt.edu.cn (Q. Zhou).

https://doi.org/10.1016/j.patcog.2021.108290
0031-3203/© 2021 Elsevier Ltd. All rights reserved.

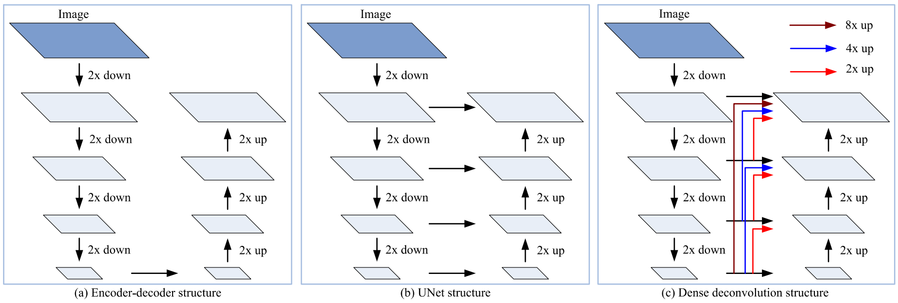


Image
2x down
2x up
8x up
4x up
(a) Encoder-decoder structure
(b) UNet structure
(c) Dense deconvolution structure

Fig. 1. Comparison of current encoder-decoder architectures (a), (b) and our dense deconvolution decoder structure (c) to capture multiple scale context. (Best viewed in color).

Fig. 1(b), encoder-decoder architecture [10,11] has been improved to capture more context information. More specifically, the feature maps of convolutional layers are always concatenated with deconvolutional ones through skip-connections. As a result, more rich features of middle layers are utilized to capture multi-scale contextual cues. This architecture, however, exhibits a pre-defined fixed structure, where the skip connections are only constructed between the convolutional layer and the corresponding deconvolution counterpart, resulting in the fact that contextual cues are not fully investigated.

To address above problems, this paper introduces a contextual ensemble network (CENet), which aggregates multi-scale context information from all convolutional layers, for semantic segmentation. CENet still adopts encoder-decoder architecture for exploring multi-level convolution features collaboratively to capture context information. Unlike previous methods that formulate contextual cues using attention schemes [4,6] or simply duplicated from the responses of encoder [10,11], the hierarchical contextual interactions of different resolutions are collaboratively investigated via *densely* upsampling multi-level pyramid features from deep convolution layers to shallow layers. More specifically, while the shallow convolutional layers of encoder always abstract low-level image statistics, the deeper layers have powerful ability to extract high-level semantics. On one hand, the high-level semantics of deeper layers is helpful to guide learning stage of low-level and medium-level convolutional features. On the other hand, the feature descriptors of shallow layers are beneficial to correctly delineate object boundaries and shapes with high-level semantics. Therefore, the features extracted from different layers are complemented each other, and their integration always leads to enhanced segmentation performance. As illustrated in Fig. 1(c), our CENet harvests and concatenates deconvolutional features with different resolutions in encoder, called *ensemble deconvolution*, to produce the feature representation in decoder, which carries both local and global context information. Specifically, for one specific deconvolution feature representation, it is first concatenated with feature maps *densely* upsampled from deeper convolutional layers in encoder (denoted as colored arrows in Fig. 1(c)), and then supplemented with its corresponding counterpart (denoted as black arrows in Fig. 1(c)). In summary, the contributions of our CENet are three-folds:

- CENet introduces a novel encoder-decoder architecture to capture multi-scale context via ensemble deconvolution. The stacked feature maps are complemented each other, allowing us to *fully* explore multiple scale contextual information embedded in images.
- Instead of using switch variables to record pooling indices [12], the feature maps of encoder are concatenated to the decoder, avoiding the extra noise introduced through padding, and without extra memory space to store pooling indices at the same time.
- The CENet is trained end-to-end and easy to execute without any postprocessing, which facilitates well for semantic segmentation. We evaluated CENet on two widely-used datasets: PASCAL VOC 2012 [14] and CityScapes [15], and the experimental results show the superior performance of CENet with respect to recent state-of-the-art networks. To further demonstrate the effectiveness of CENet, we also evaluate it on MS COCO [16] and ISBI 2012 [17]. The experiments show that CENet achieves promising results for instance segmentation and biological segmentation.

The remainder of this paper is organized as follows. After a brief discussion of related work in Section 2, the detail architecture of CENet is introduced in Section 3. In Section 4, we elaborate on the end-to-end training of CENet. The proposed network has been evaluated on PASCAL VOC 2012, CityScapes, MS COCO and ISBI 2012 datasets, respectively, and the experiments can be found in Section 5. Finally, the concluding remarks and future work are given in Section 6.

2. Related work

Due to the powerful ability to abstract image features, the recent years have witnessed vast number of convolutional neural networks (CNNs) for semantic segmentation, which are mainly divided into two categories: fully convolutional networks and encoder-decoder networks. However, multiple stages of spatial pooling greatly reduces feature resolution, and convolution with small filter size (e.g., 3×3) always leads to very limited field-of-view. Networks with insufficient field-of-view may not be able to capture enough context information and thus degrade the performance. We thus review the related work based on these two aspects, which attempt to capture contextual clues to address these issues.

2.1. Fully convolutional based architectures

Initially from CNNs, the FCNs (e.g., the VGG-16 [18] and ResNet [19]) employ convolutional layers to replace fully-connected layers. Actually, the direct estimations of FCNs are essentially of low resolution, which highly detriments segmentation accuracy. Therefore, a variety of FCN-based methods adopt a naive method of directly encoding contextual cues through skip-connections. For example, the developed version of FCN [1] utilizes skip-connections to explore mid-layer context features for high-resolution prediction. PSPNet [3] adds a global pooling branch to extract context information. In [9], image pyramid is also utilized to capture multi-scale context in FCN framework. DANet [4] and OCNet [5] encode the semantic context using non-local attention to explore global context, where spatial and channel attention schemes are adopted to construct long-ranged interactions. In [6], the authors propose criss-cross attention block to investigate global context, yet with very limited computational resources.

The alternative approaches to capture context clues emoploy conditional random fields (CRFs) as postprocess after inference from FCNs. DeepLab [2] is the typical FCN-CRF based models, where the atrous convolution is first applyed to produce unary potentials, and then the Potts model is considered as pairwise potentials to encode long-range pixel interactions. CRF-RNN [20] and deep structured network (DSN) [21] extend DeepLab family networks by implementing the mean field CRF inference as recurrent layers for end-to-end learning of the dense CRF and FCN network.

2.2. Encoder-decoder based architectures

Unlike FCN-based methods, the encoder-decoder networks [22–24] exhibit a nearly symmetrical network structure that gradually recovers image details using upsampling and deconvolution operation. For instance, [22] employs atrous separable convolution in decoder for semantic segmentation. Through adding skip connections, UNet [10] and its dense version [11] introduce elegant symmetric network architectures, which concatenate feature maps from the encoder side to the corresponding decoder activations. To capture global context, attention scheme can be also pluged into decoder. For example, peng et al. [7] design stride spatial pyramid pooling for harvesting high-level semantics, and dual attention blocks to abstract low-level statistics. In [8], context association is formulated in decoder network by learning channel contextual module and spatial contextual module, respectively. RefineNet [25] and its variants [23,24,26] carefully design cascaded deconvolution network in score map and feature map, respectively, to capture multi-scale context cues. Our CENet also utilizes the encoder-decoder architecture, and hence can be classified into this category. In contrast to previous methods that capture context information using image pyramid [9,25] or simply duplicated the convolutional features [10,11], however, our CENet considers to transfer all available encoder features in deconvolution process, yielding dense-upsampled structure to *fully* investigate context clues, which enables selective and adaptive aggregation of multi-level contextual features.

An early version of this work was first published in Yang et al. [27]. This journal version extends previous one in following aspects: (1) The previous version directly stacks deconvolution feature maps, which may produce a large number of feature channels and model parameters. We apply 1×1 convolution to reduce dimension before concatenation, which is also beneficial for training the entire network. (2) Unlike [27] that employs two stage training scheme, our CENet is trained in terms of end-to-end segmentation to match the resolution of input image. (3) In stead of using VGG-16 network as backbones, our CENet employs ResNet-101 as more powerful backbone to improve performance. (4) We have performed more exhausted experimental evaluation, and reported more comparisons and improved results.

3. CENet

This section first elaborates on the overall architecture details of our CENet. Thereafter, the details on how to design our decoder is introduced in a dense upsampling manner.

3.1. The overall architecture of CENet

Fig. 2 shows the overall architecture of the entire CENet. Similar to previous encoder-decoder networks [10,12], our CENet is also composed of two parts: encoder and decoder network. The decoder includes three fundamental components: convolution layer together with batch normalization (BN) and rectified linear unit (ReLU), concatenation layer, and softmax layer. The encoder network corresponds to feature extractor that transforms the input image to multiple scale dimensional feature representation. On the contrary, the decoder network delineates object shape boundaries that output object segmentation from the convolution features produced from encoder network. Let $\mathcal{K}$ be the total number of object categories that are required to be classified. The final output of our CENet is a $(\mathcal{K}+1)$-dimensional probability map with the same size of input image, indicating probability of each pixel belonging to one of the predefined $\mathcal{K}$ classes or to the additional background.

3.2. ResNet as backbone

In order to obtain high-quality semantic segmentation outputs and make fair comparison with recent state-of-the-art networks, we borrow the architecture widely used in ResNet [19] to construct our backbone network. As shown in Fig. 2, our encoder, pre-trained on ImageNet [28], shares the same configurations of ResNet-101 to abstract deep features. When going deeper in the backbone, however, it is very hard to recover tiny objects as their spatial information has been totally lost in the convolutional features with lowest resolution. Therefore, the final pooling layers and the following fully-connected layers are removed in our encoder network. Moreover, more layers result in a large number of additional computation, thus our backbone executes more efficiently, and at the same time ensures 2D representation that facilitates semantic segmentation. As a result, there are five stages in our backbone module, where each one has the resolution of $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ with respect to input image. One may also employ holding-resolution version of ResNet [6,29,30] using dilation convolutions, where all the feature maps in the last three stages have the same spatial size (e.g., with resolution of $\frac{1}{8}$ with respect to input image). These approaches, however, are sacrificed with expensive computation due to high feature resolutions, and suffer from the gridding artifacts that may degrade the performance.

3.3. The detail decoder architecture of CENet

On the other hand, the decoder network contains a series of concatenation layers, convolutional layers, and a Sofmax layer, which are represented by blue, brown and red boxes in Fig. 2, respectively. In order to match the resolution of input image, the outputs of encoder are upsampled three times, leading to four deconvolution stages in the architecture of decoder. Each stage contains one concatenation layer and two convolutional layers. More specifically, in addition to directly duplicate feature maps in encoder [10,25,27], our concatenation layers harvest multi-scale context clues by stacking feature representation from previous stage in decoder and a series of deconvolutional features, where the correspondingly counterparts in encoder have more deeper stages with respect to the stage of current concatenation layer.

As shown in Fig. 2, the concatenation layers are produced through three basic operations: upsampling (denoted as red, green and purple arrows), convolution (denoted as blue arrows), and concatenation. However, directly concatenating all features will lead to the vast number of feature channels in concatenation layers. Considering the case of rightmost concatenation layer in Fig. 2, it is produced by stacking convolution features from stage1 to stage4 in encoder. If features are directly concatenated, the channel number will be $64+256+512+1024=1856$ using RseNet-101 as backbone. Training such complicated network is a non-trivial work, and probably limits the generalization ability and resulting in overfitting of our model. To this end, unlike our preliminary version [27], the dimension of feature maps, which are used for aggregating, has

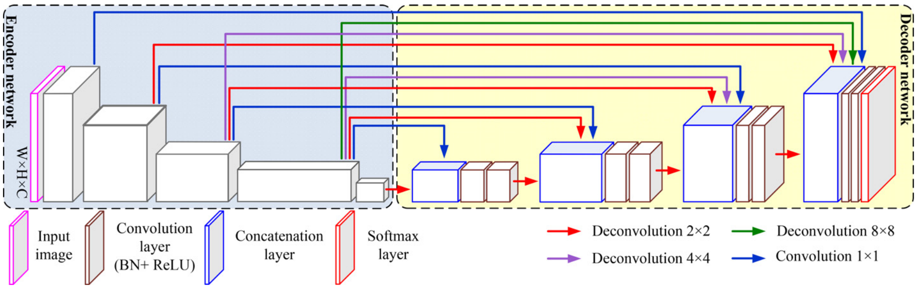


Fig. 2. Overall architecture of the proposed CENet. On top of the network based on UNet [10], we construct dense skip-connections from encoder to decoder, producing the delineated segmentation map of an input image. For input image, W, H, and C stand for width, height, and number of channels of input image, respectively. (Best viewed in color).

to be reduced. Specifically, we resort to employ an 1×1 convolution to reduce feature dimension before stacking them in each concatenated path. For convenient computing, the number of feature channels keeps the same with counterpart convolution layer in encoder. Immediately below, we introduce how to produce the concatenation layer in each stage.

As shown in Fig. 2, the resolution of feature maps $\mathbb{F}_p$ from prepositive convolutional layer is first enlarged twice, and then convoluted with a 3×3 filter kernels $\mathcal{F}_{3\times3}(\cdot)$, which halves the number of feature channels:

$$\hat{\mathbb{F}}_p = \mathcal{F}_{3\times3}(\mathcal{U}_{r=2}(\mathbb{F}_p)) \tag{1}$$

where $\mathcal{U}_{r=2}(\cdot)$ stands for two times upsampling. On the other hand, the feature maps $\mathbb{F}_e$ within deeper layers in encoder network have to be expanded with different upsampling ratio r (denoted as colored arrows in Fig. 2), resulting in feature representation of equal resolution for stacking. After that, an 1×1 convolution $\mathcal{F}_{1\times1}(\cdot)$ with stride 1 is applied into the enlarged feature maps to further extract contextual information, and reduce feature channels at the same time:

$$\hat{\mathbb{F}}_e = \mathcal{F}_{1\times1}(\mathcal{U}_r(\mathbb{F}_e)), r \in \{2, 4, 8\} \tag{2}$$

Finally, due to having the same resolution, the feature maps $\mathbb{F}_c$ of counterpart in encoder are directly fed into an 1×1 convolution $\mathcal{F}_{1\times1}(\cdot)$ for dimension reduction, without upsampling operation:

$$\hat{\mathbb{F}}_c = \mathcal{F}_{1\times1}(\mathbb{F}_c) \tag{3}$$

Thereafter, all the convolutional feature maps are stacked together to generate our concatenation layer, allowing us to fully explore multiple scale context cues:

$$\hat{\mathbb{F}}_s = \hat{\mathbb{F}}_p \bigodot \hat{\mathbb{F}}_e \bigodot \hat{\mathbb{F}}_c \tag{4}$$

where $\bigodot$ indicates concatenated operation. Still taking the rightmost concatenation layer in Fig. 2 into account, it only has 64 feature channels, since an 1×1 convolution is utilized to evenly reduce channel number to 16 in each stacking path. Therefore, although a series of stacking operations are used to aggregate densely upsampled features, the decoder of our network achieves very small model size and high implementing efficiency, demonstrated by the experimental results in Section 5.7.2.

These concatenated feature maps $\hat{\mathbb{F}}_s$, carrying both local and global context, are fed into two 3×3 convolutions, each followed by a batch normalization layer and ReLU activation layer. At the end of decoder, an 1×1 convolution is used to map each feature vector to the desired number of classes $\mathcal{K}$, received supervisions from the ground truth.

4. Training CENet

In this section, we first introduce batch normalization, which is widely used for network training. Then we will elaborate on the details of how to train CENet in terms of end-to-end manner, although it is very deep (nearly twice deeper than FCNs [1,2]).

4.1. Batch normalization (BN)

According to Ioffe and Szegedy [31], it is very hard to train a deep neural network due to the internal-covariate-shift problem. Since the parameters of previous layers have been updated, the distributions of filter responses in current layer change in the process of iterative training. This is not beneficial for optimizing our CENet since such changes may be amplified through back propagation across layers, probably leading to the vanishing or exploding gradient. In order to address this problem, there are some widely-used tricks in training process such as BN [31], Glorot initialization [32], and Adam solver [33] for solving image understanding tasks [1,2,19]. In our CENet, an extra BN layer is added to the output of every convolution layer, where the filter responses are normalized to a standard Gaussian distribution. In the experiments, we observe that the batch normalization is critical to optimize our network, which helps our training algorithm to straggle from poor local optimum.

4.2. End-to-end training

In our CENet, a soft-max function, which is the generalization of logistic function, is adopted to convert the outputs of Softmax layer to probabilities between (0,1). Let $a_k(\boldsymbol{x}, \boldsymbol{\theta})$ denotes the activation for kth category for pixel $\boldsymbol{x}$ given network parameters $\boldsymbol{\theta}$, then the soft-max function $p_k(\boldsymbol{x}, \boldsymbol{\theta})$ is defined as:

$$p_k(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\exp\{a_k(\boldsymbol{x}, \boldsymbol{\theta})\}}{\sum_{k'}^{\mathcal{K}} \exp\{a_{k'}(\boldsymbol{x}, \boldsymbol{\theta})\}} \tag{5}$$

In the inference process, the kth semantic category is assigned to pixel $\boldsymbol{x}$ if it achieves the highest predicted probability $k^* = \arg\max_k p_k(\boldsymbol{x}, \boldsymbol{\theta})$.

For the task of semantic segmentation which is always formulated as a dense pixel-wise classification problem, we use the standard cross-entropy loss [1,8] as objective function to evaluate segmentation estimation with respect to the associated ground truth. The loss is summed up over all the pixels in a mini-batch. Let N be the total number of pixels in a training batch and y_i^k is the ground-truth semantic label of kth category for pixel $\boldsymbol{x}_i$, our training target is to find an optimal model parameters $\boldsymbol{\theta}^*$ that minimizes the cross-entropy loss $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* = \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2 - \sum_{i=1}^{N} \sum_{k=1}^{\mathcal{K}} y_i^k \log p_k(\boldsymbol{x}_i, \boldsymbol{\theta}) \tag{6}$$

where a regularization is added to network parameters $\boldsymbol{\theta}$, and λ is a non-negative weight decay parameter. We can apply stochastic gradient descent (SGD) methods to optimize the above problem in the process of back propagation, where the gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{x}, \boldsymbol{\theta})$ is easily computed by applying the chain rule as in conventional CNNs [1,19]. One problem in training is category unbalancing, where there is large variation in the number of pixels in each class in the training set. An example is the category of "traffic sign" and "road" in CityScapes dataset, in which the object instants of first class occupy a very small number of image regions, while those of the second one occupy a large number of pixels. Therefore, it is required to reweight the category loss according to the true class distribution. We use median frequency balancing [34] to solve this problem, where the assigned weight to a class in Eq. (6) is the ratio of the median of class frequencies divided by the class frequency, calculated on the entire training set. This implies that the categories with larger number of training data have a weight smaller than 1, and conversely, those with smallest number of training samples achieve the highest weights.

5. Experiments

We evaluate our method on four challenging semantic segmentation datasets: CityScapes [15], PASCAL VOC 2012 [14], Microsoft COCO [16], and ISBI 2012 [17], which cover various types of applications and associated scene images, including self-driving for street scene, indoor/outdoor scene understanding, and cell segmentation for biological medicine scene etc. The purpose of experiments is to understand the underlying behavior of our network in different applications and challenges.

5.1. Datasets

Cityscapes [15] is very popular for self-driving task, where a car is treated as an autonomous robot to perceive surroundings, including recognizing and localizing objects. It provides a large-scale dataset that contains high-resolution street scene images from 50 different cities, where 5000 images with pixel-level annotations are provided for 19 object categories, such as road, car, pedestrian, bicycle, sky etc. All images are divided into three parts: 2975 training, 500 validation and 1525 testing images. We use the trainval set (3475 images) for training. Since the ground truth of the test set is not available, we evaluate our method through an online evaluation server.

PASCAL VOC 2012 [14] is widely-used for scene semantic segmentation. This dataset contains 21 object categories (20 foreground categories and one additional background class). The original dataset includes 1464 (train), 1449 (val), and 1456 (test) images for training, validation, and testing, respectively, where the images in training and validation sets have per pixel-level annotations. For training, we use the extra augmented segmentation annotations from [35], which includes 10,582 training and validation images. The remaining 1456 test images are used to evaluate the performance of our CENet.

The ISBI 2012 [17] is a biological image dataset, including a set of 30 images (with solution of 512 × 512 pixels) of the Drosophila first instar larva ventral nerve cord (VNC) from serial section transmission electron microscopy. Each image is pixel-wised annotated using a binary ground truth, where white color denotes cells and black color indicates the membranes. However, using only few training samples may lead to the variance and non-robustness of the network. Therefore, we utilize the data augmentation to expand training set. More specifically, each training image is split into 256 image patches with 32 × 32 pixels, resulting in 7680 training data. The ground truth has a corresponding split. The test set is publicly available, but the associated segmentation maps are not provided. We following [10] to evaluate CENet by sending the estimated segmentation probability maps to the organizers.

Microsoft COCO [16] is a very challenging dataset for instance segmentation. This dataset contains 115k images over 80 categories for training, 5k images for validation and 20k images for testing. Instead of using mIoU that is widely accepted to measure semantic segmentation, we evaluate our CENet over COCO datset in terms of AP and AP50.

5.2. Implementation details

To show the advantages of our approach, we selected 5 state-of-the-art models as baselines to evaluate on CityScapes [15] and PASCAL VOC 2012 [14], including FCN-8s [1], MDCNet [36], DeepLab [22], CONet [37], APCNet [38]. For ISBI 2012 [17] dataset, we employ UNet [10] and UNet++ [11] as baselines. For MS COCO [16] dataset, we use the competitive Mask R-CNN model [39] as baseline. Experimental results of some baseline models are produced using default parameter settings given by the authors, while others are directly taken from the literature.

For CityScapes and PASCAL VOC 2012 datasets, the segmentation performance is measured by the mean intersection-over-union (mIoU) score [22,37,38]. Let c_{ij} be the number of pixels with the ith category as the ground truth and the jth class as the prediction; t_i is the total number of pixels for the ith category in the ground truth, then mIoU score calculates the mean portion of the intersection between the ground truth and the prediction:

$$mIoU = \frac{1}{\mathcal{K}}\sum_i \frac{c_{ii}}{t_i + \sum_j c_{ij} - c_{ii}} \tag{7}$$

For ISBI 2012 dataset, on the other hand, we measure the segmentation outputs in terms of foreground-restricted rand scoring (RST) and foreground-restricted information theoretic scoring (IST) after border thinning, provided by the organizers [17].

The entire CENet is implemented based on Caffe framework [40]. The input images and the corresponding pixel-wised annotated ground truth are used to train the network using the stochastic gradient descent algorithm [41]. In order to make full use of the GPU memory, we favor a large batch size (set as 14) to train batch normalization parameters, where initial learning rate, momentum and weight decay are set to 10^{-3}, 0.99 and 5×10^{-4}, respectively. Following [2], we employ a "poly" learning rate policy where the initial learning rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with $power = 0.9$.

5.3. Evaluation results on CityScapes

Table 1 compares our method with baseline networks on CityScapes dataset, and reports results in terms of individual category mIoU and average mIoU over all categories. It clearly demonstrates that our CENet outperforms other state-of-the-art approaches, where 12 out of 19 categories achieve the best performance. Among all methods, our CENet achieves the best segmentation performance with 82.5% mIoU, improving 0.7% mIoU compared with the second best network CONet [37]. Among all baselines, the FCN-8s [1] is at the lowest rank, probably because of its very simple network architecture, which is not able to capture multi-scale contextual cues effectively. It is intriguing that our approach is superior to the existing methods [22] that employ CRF as post-processing to explore short-ranged and long-ranged interactions among pixels. This indicates CENet has powerful ability to capture wide scale context information to further improve the performance of segmentation outputs.

In order to show the qualitative results, we also trained our CENet using only "train" set (2,975 images), and produce segmentation results on "val" set. Some visual example of simultaneous recognition and segmentation ares shown in Fig. 3. Each example shows both the original image and the color coded output labeling.

Table 1
Individual category results and the average over all categories on the CityScapes test set in terms of mIoU scores. The best performance for each individual class is marked with a bold-face number.

Method	Roa	Sid	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter
FCN-8s [1]	97.1	76.6	88.1	32.9	38.5	48.3	56.4	62.1	90.7	66.7
MDCNet [36]	97.4	79.6	89.0	45.8	47.3	59.1	59.3	70.6	90.5	68.3
APCNet [38]	98.7	86.9	93.5	58.4	**63.8**	67.7	76.1	80.5	93.6	72.2
DeepLab [22]	98.6	86.2	93.5	55.2	63.2	70.0	77.1	81.3	93.8	72.3
CONet [37]	**98.9**	87.9	93.9	61.3	63.1	**72.1**	**79.3**	**82.4**	94.0	73.4
Ours	98.8	**89.1**	**94.6**	**62.7**	63.7	66.4	75.7	79.7	**94.7**	**73.6**
Method	Sky	Ped	Rid	Car	Tru	Bus	Tra	Mot	Bic	mIoU
FCN-8s [1]	92.7	74.3	44.3	91.5	36.9	41.3	32.8	45.7	62.7	62.1
MDCNet [36]	94.8	76.2	52.4	92.6	60.5	71.2	50.4	50.2	69.1	69.7
APCNet [38]	95.3	86.8	71.9	96.2	77.7	91.5	83.6	70.8	77.5	81.2
DeepLab [22]	95.9	87.6	73.4	96.3	75.1	90.4	85.1	72.1	78.3	81.3
CONet [37]	96.0	**88.5**	75.1	**96.5**	72.5	88.1	79.9	73.1	79.2	81.8
Ours	**96.4**	87.3	**75.4**	94.2	**79.4**	**91.9**	**86.8**	**73.3**	**79.7**	**82.5**

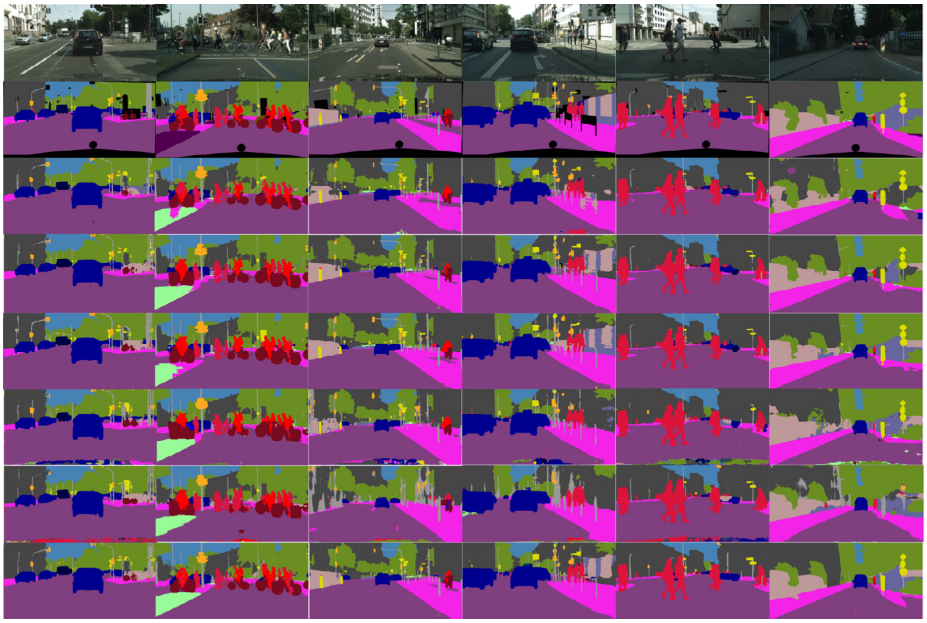


Fig. 3. The visual comparison on CityScapes val dataset. From top to bottom are original images, the corresponding ground truth, segmentation outputs from FCN-8s [1], MDCNet [36], APCNet [38], DeepLab [22], CONet [37] and our CENet. (Best viewed in color).

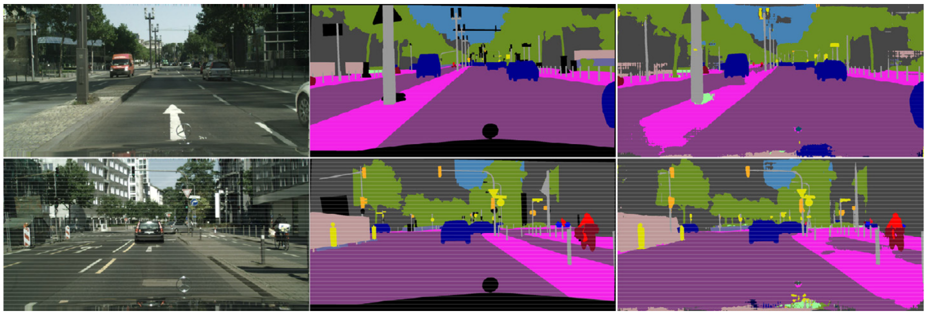


Fig. 4. Some failure visual examples on CityScapes validation dataset. From left to right are original images, corresponding ground truth, and our segmenting results. (Best viewed in color).

Except some boundary pixels that exhibit relative higher confusion, nearly all pixels are correctly classified. Our method also obtains better segmenting results for the tiny object instances, such as “bicycle”, “traffic sign” and “traffic light”, etc. As shown in Fig. 4, we also illustrate some failure visual examples. It is discovered that the area of “road” and “sidewalk” are sometimes incorrectly classified, probably due to the fact that these two categories share extremely similar visual appearance (e.g., intensity, color and texture).

5.4. Evaluation results on pascal VOC 2012

We now demonstrate that our method scales nicely in indoor/outdoor scenario when augmenting the number of images

Table 2
Individual category results and the average over all categories on the PASCAL VOC 2012 test set in terms of mIoU scores. The best performance for each individual class is marked with a bold-face number.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	mIoU
FCN-8s [1]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	62.2
DeepLab [22]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	71.6
MDCNet [36]	85.2	41.0	83.4	69.4	80.4	89.5	85.1	87.1	40.3	78.1	73.1
APCNet [38]	**95.8**	75.8	84.5	76.0	80.6	96.9	90.0	96.0	42.0	**93.7**	84.2
CONet [37]	95.7	71.9	**95.0**	**76.3**	**82.8**	94.8	90.0	95.9	37.1	92.6	84.2
Ours	95.1	**77.0**	90.8	74.2	80.9	**95.8**	**91.6**	**96.4**	**43.1**	91.5	**84.7**
Method	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mIoU
FCN-8s [1]	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLab [22]	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
MDCNet [36]	47.8	82.2	76.9	79.0	84.5	58.5	83.2	53.3	84.2	72.1	73.1
APCNet [38]	**75.4**	91.6	**95.0**	90.5	89.3	75.8	92.8	61.9	88.9	**79.6**	84.2
CONet [37]	73.0	**93.4**	94.6	89.6	88.4	74.9	**95.2**	63.2	**89.7**	78.2	84.2
Ours	73.3	91.9	94.2	**90.6**	**90.4**	**77.3**	93.8	**66.9**	89.1	78.3	**84.7**

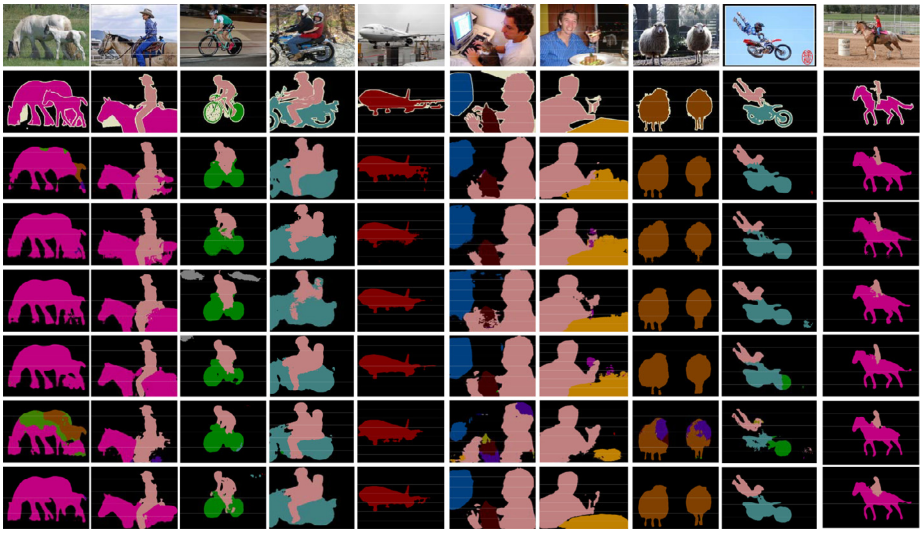


Fig. 5. The visual comparison on PASCAL VOC 2012 val dataset. From top to bottom are original images, the corresponding ground truth, segmentation outputs from FCN-8s [1], DeepLab [22], MDCNet [36], APCNet [38], CONet [37] and our CENet. (Best viewed in color).

and classes on PASCAL VOC segmentation dataset [14]. Table 2 shows the superior performance of the proposed approach. Compared with the state-of-the-art baselines, our CENet achieves highest 84.7% mIOU accuracy. It obtains the best score on 9 out of the 20 categories. This superior performance can be attributed to the architecture of CENet, which allows us to effectively capture the contextual clues within different scales. Consistent with Table 1, FCN-8s once again ranks at the bottom, only obtaining 62.2% mIoU. Compared with the images in CityScapes dataset, the images in PASCAL VOC 2012 dataset have larger variance in visual appearance and more complex and clutter background.

Fig. 5 shows some visual examples of segmentation outputs on the PASCAL VOC validation set. Each example also exhibits the input image and the corresponding color coded output. Compared with baselines, our CENet not only correctly classifies objects with different scales, but also produces better consistent qualitative results for all classes. Other networks, such as FCN-8s [1] and DeepLab [22] achieve poor segmentation output, since their receptive fields are not adaptive to encode multi-scale context, resulting in the problem that objects substantially larger or smaller than the receptive field may be fragmented or incorrectly classified. For instance, the "cat" in the sixth example and the "glass" in the seventh example in Fig. 5.

Table 3
Segmentation results on MS COCO validation set in terms of AP and AP50.

Method	AP_{mask}	AP^{50}_{mask}	AP_{bbox}	AP^{50}_{bbox}
Mask R-CNN [39]	36.1%	57.5%	40.0%	60.5%
Ours	36.7%	58.4%	40.8%	61.6%

5.5. Evaluation results on MS COCO

To further demonstrate the generality of our CENet, we conduct the instance segmentation task on MS COCO [16], where the segmentation head is replaced by our method. We use the official implementation with end-to-end joint training whose performance is almost the same as the baseline reported in Huang et al. [6], Wang et al. [42]. For fair comparison, all models are fine-tuned from pre-trained model based on ImageNet. Table 3 reports comparison results between baseline and our method. It can be seen that our method achieves 36.7% AP and 58.4% AP50 in terms of segmentation masks, and 40.8% AP and 61.6% AP50 in terms of detected bounding boxes, respectively, representing a clear improvement by a margin of 0.6%, 0.9%, 0.8%, and 1.1% over Mask R-CNN model. Fig. 6 also illustrates some qualitative visual results of our

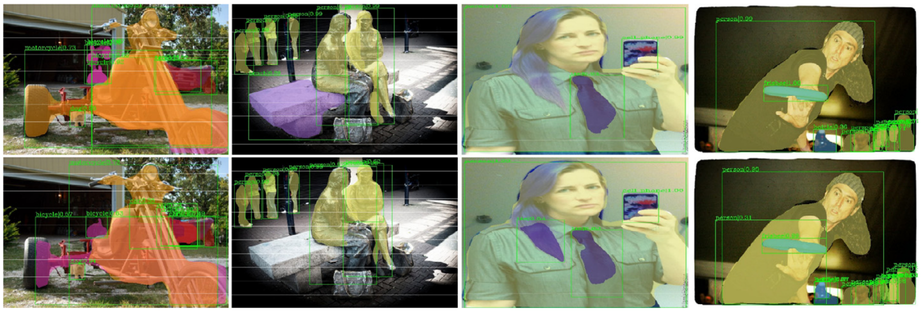


Fig. 6. The visual comparison on MS COCO val dataset. From top to bottom are instance segmentation results from our approach and baseline Mask R-CNN model [39]. (Best viewed in color).

Table 4
Segmentation results on the ISBI 2012 test set in terms of RST and IST scores.

Method	RST	IST
UNet [10]	0.9621	0.9808
UNet+ [11]	0.9653	0.9877
Ours	0.9696	0.9914

Table 5
Contributions of different scale context combinations in terms of mIoU (%).

Method	CityScapes [15]	PASCAL VOC [14]
scale1	67.1	70.3
scale1 + scale2	69.6	73.7
scale1 + scale2 + scale3	70.2	74.1
scale1 + scale2 + scale3 + scale4	70.5	74.4

method in terms of detected bounding boxes and segmentation masks. It clearly shows that our method achieves better visual results with respect to baseline model. For instance, in first example, our method correctly segments and identifies the area of “motorcycle”, while the left tire is misclassified as “bicycle” using Mask R-CNN model [39]. This can be also observed in second and third examples, where the “bench” is missing and woman hair are incorrectly classified as “tie”. Although the goal of our method is to perform semantic segmentation, it is interesting to point out that our method achieves less false positive of detective bounding boxes, such as “person” in second and forth examples. This is probably because that object detection task is benefit from our segmentation results.

5.6. Evaluation results on ISBI 2012

In this section, we evaluate our CENet on the task of bio-image segmentation over ISBI 2012 dataset [17], where the segmentation results are binary outputs. The quantitative results are reported on Table 4. Our CENet obtains 0.9696 and 0.9914 score in terms of RST and IST. Compared with second-rank model UNet++ [11], our approach improves RST and IST by 0.0043 and 0.0037, respectively. Some qualitative results compared with UNet++ [11] model are shown in Fig. 7, in which our method produces more flat segmentation area of cell inside regions (marked with red rectangles), and more smooth segmentation boundary of membranes (marked with blue rectangles). This is probably because the integration of multiple deconvolutional features has more powerful representation than the context encoding scheme adopted in UNet++ [11], where individual deconvolutional feature is considered.

5.7. Ablative studies

To understand the underlying behavior of our system, this section reports the results of a series of ablation studies. Note all the experiments are evaluated on validation set.

5.7.1. Ablative study on sequential context introduction

To investigate the effectiveness of the different scale context of our proposed CENet, we conduct ablative studies on CityScapes and PASCAL VOC 2012 dataset, where multiple scale context cues (denoted as different color arrows in Fig. 2) are sequentially added to our systerm, using the same training scheme and loss functions. More specifically, the baseline, denoted as *scale1*, is constructed by concatenating the feature maps with same resolution of encoder and decoder (blue arrows shown in Fig. 2). Then different scales of context features, such as red, purple, and green arrows shown in Fig. 2 (denoted as *scale2, scale3*, and *scale4*, respectively), are sequentially introduced. Table 5 reports the contributions of their combinations in terms of mIoU.

It is observed that the performance increases as more scale context cues are investigated. Specifically, using the full scale context leads to 70.5% and 74.4% for the two datasets. This is due to the fact that the semantic features of deeper layers are helpful to rectify classification error from shallow layers, while the shallow layers provide more spatial details to delineate object shapes and boundaries. Another interesting observation is that, with the introduction of different scale context information, the performance gain is gradually reduced, i.e. 2.5%, 0.6%, and 0.3% on CityScapes dataset, and 3.4%, 0.4%, and 0.3% on PASCAL VOC 2012 dataset, respectively. This indicates that integrating 4 scales of middle-level convolutional features provides enough context for these segmentation tasks. Some segmentation outputs of visual examples from two datasets are illustrated in Fig. 8. It is evident that when more contextual information is captured in our CENet, the segmented objects have more accurate boundaries, i.e., “building”, “tree”, and “person”, even for tiny object instances such as “bicycle”, “traffic sign” and “traffic light”.

5.7.2. Ablative study on implementing efficiency

To analyze running efficiency of our CENet, we carry on ablative studies on the Cityscapes dataset by adopting different scales of contextual features. We have aslo compare with some recent state-of-the-art networks including UNet [10], UNet++ [11], and CCNet [6] in terms of model size and FLOPs. Note all experiments are conducted using ResNet-101 as backbone, thus we only compare model size and FLOPs of decoder. Besides, the resolution of input images keeps 769 × 769 for fair comparison. The results are reported in Table 6. We observe that in spite of adopting concatenating features, our system has similar model size and FLOPs with respect to UNet++ [11], but achieves smaller model size and lower FLOPs than CCNet [6]. The main reason is that an 1 × 1 convolution is always adopted before feature concatenation in each upsampling

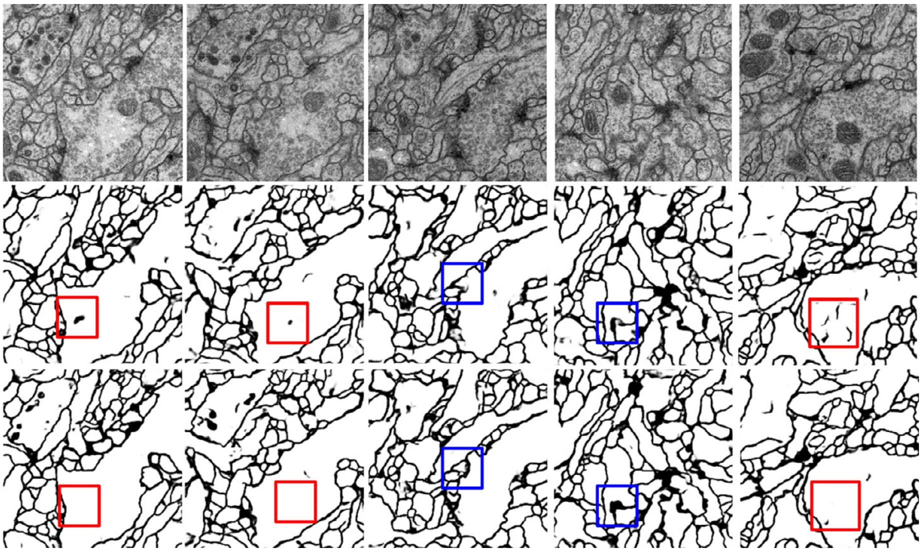
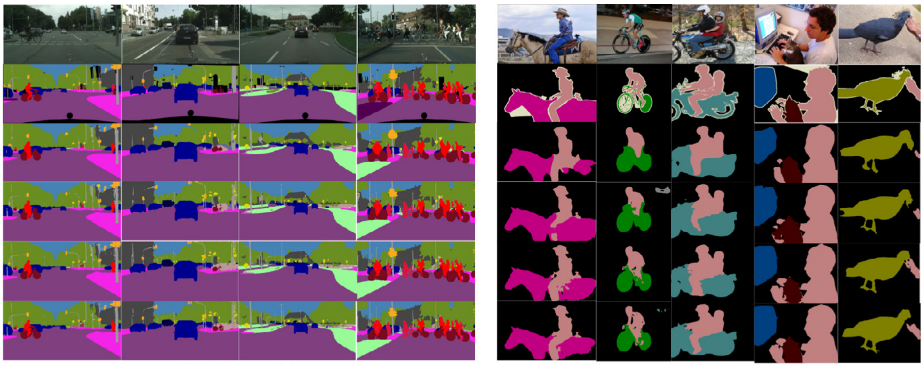


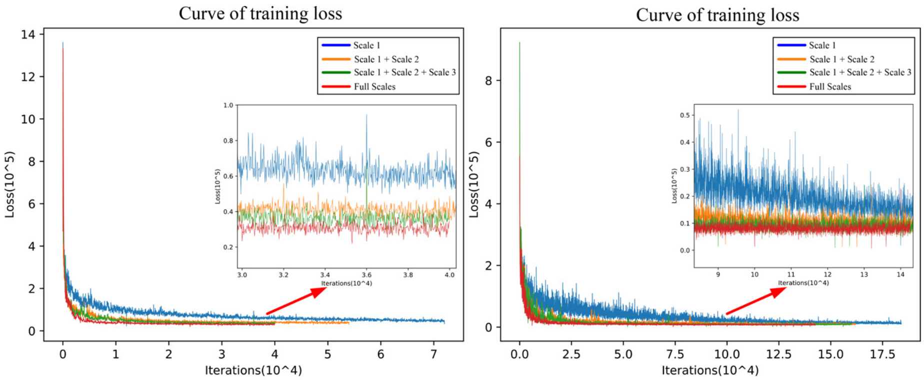
Fig. 7. The visual comparison on ISBI 2012 test dataset. The first row depicts original images, the second and third rows are binary segmentation results produced by UNet++ [11] and our CENet. (Best viewed in color).



(a) CityScapes dataset

(b) PASCAL VOC 2012 dataset

Fig. 8. Some visual segmentation outputs by sequentially adding multiple scale context on (a) CityScapes and (b) PASCAL VOC 2012 dataset. From top to bottom are input images, the ground truth, and segmentation results from *scale1, scale1 + scale2, scale1 + scale2 + scale3*, and full scales. (Best viewed in color).



(a) CityScapes dataset

(b) PASCAL VOC 2012 dataset

Fig. 9. The loss vs iteration on (a) CityScapes and (b) PASCAL VOC 2012 dataset, respectively, where multi-scale context information are sequentially considered in our network. (Best viewed in color).

step, resulting in great reduction of model parameters and FLOPs to accelerate running speed. In spite of having nearly symmetrical structure as well as UNet [10], our system is still implemented slower than UNet [10] since our CENet involves more complicated skipped connections. Note when only first scale context is added to baseline, our network degenerates to UNet model, thus achieving nearly the same model size and FLOPs with respect to UNet.

5.7.3. *Ablative study on convergence of training process*

To further demonstrate the effectiveness of our method, Fig. 9 also plots the curves of loss function as the iteration number increases on CityScapes and PASCAL VOC 2012 datasets. Once again, one can observe that the more context information is utilized, the faster convergence speed can be achieved, which is consistent with the conclusion of Table 5. We also discover that, com-

Table 6
Comparison of implementing efficiency on Cityscapes validation dataset. FLOPs and model size are estimated for an input of resolution 769 × 769.

Method	FLOPs	Parameters (M)
baseline	0	0
baseline + scale1	6.7	40.6
baseline + scale1 + scale2	8.1	46.6
baseline + scale1 + scale2 + scale3	8.8	50.5
baseline + full scales	9.3	53.3
CCNet [6]	24.7	208
UNet [10]	6.7	40.5
UNet+ [11]	8.9	51.8

pared with PASCAL VOC 2012 dataset, more flatten curves are obtained on CityScapes dataset. This is probably because PASCAL VOC 2012 dataset involves more training data, greater visual variance, and larger number of object categories.

6. Conclusion remarks and future work

This paper has proposed a novel encoder-decoder network, named CENet, to explore hierarchy convolution features collaboratively for accurate pixel-wised semantic segmentation. Compared with recent encoder-decoder networks, our CENet provides a more powerful representation to capture multi-scale context information through constructing ensemble deconvolution from encoder to decoder. Dense upsampling enables CENet to combine feature maps with different receptive fields, thus allowing us to fully investigate local and global context cues. To evaluate our method, the experiments are conducted on CityScapes and PASCAL VOC 2012 datasets. The experimental results show that our CENet outperforms recent state-of-the-art networks, and demonstrate that our approach can produce more accurate predictions and delineated segmentation outputs. We also validate the scalability of CENet on MS COCO dataset for instance segmentation with augmented training images and semantic categories. Our method still achieves outstanding performance. Finally, we demonstrate our approach for the task of biological segmentation, where the experimental results show the effectiveness of our approach on ISBI 2012 dataset.

In spite of obtaining impressive results on segmentation accuracy, our method sacrifices implementing efficiency. The experimental results show that our CENet performs slower than UNet [10] and UNet++ [11]. As a result, one future direction will involve in-depth model design, regarding the lightweight architecture to reduce the number of model parameters and computational burden, without significant performance drop simultaneously. In addition, we are interested in extending our model in spatio-temporal domain (e.g., video sequence) to perform video segmentation.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Acknowledgments

The authors would like to thank the Editor in Chief and all the anonymous reviewers for their insightful suggestions and valuable comments. This work was jointly supported in part by the National Natural Science Foundation of China under Grants 61876093, 61701252, 61801242, the National Natural Science Foundation of Jiangsu Province under Grant BK20181393, the National Science Foundation under Grant IIS-1302164.

References

[1] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 640–651.
[2] C. Liang-Chieh, P. George, K. Iasonas, M. Kevin, Y. Alan L., DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.
[3] H. Zhao, J. Shi, X. Qi, X. Wang, J.Y. Jia, Pyramid scene parsing network, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 6230–6239.
[4] J. Fu, J. Liu, H. Tian, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2019, pp. 3141–3149.
[5] Y.H. Yuan, J.D. Wang, Ocnet: object context network for scene parsing, arXiv preprint arXiv:1809.009162018.
[6] Z.L. Huang, X.G. Wang, L.C. Huang, C. Huang, Y.C. Wei, W.Y. Liu, CcNet: criss-cross attention for semantic segmentation, in: IEEE International Conference on Computer Vision, 2019, pp. 603–612.
[7] C.L. Peng, J.Y. Ma, Semantic segmentation using stride spatial pyramid pooling and dual attention decoder, Pattern Recognit. 107 (6) (2020) 107498–107513.
[8] J. Fu, J. Liu, Y. Li, Y.J. Bao, W.P. Yan, Z.W. Fang, H.Q. Lu, Contextual deconvolution network for semantic segmentation, Pattern Recognit. 101 (1) (2020) 107152–107163.
[9] M. Orši, S. Šegvic, Efficient semantic segmentation with pyramidal fusion, Pattern Recognit. 110 (8) (2021) 107611–107624.
[10] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer Assisted Intervention, 2015, pp. 225–233.
[11] Z.W. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J.M. Liang, Unet++: redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Image 39 (6) (2020) 1856–1867.
[12] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[13] B. Vijay, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[14] M. Everingham, S.M.A. Eslami, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: aretrospective, Int. J. Comput. Vis. 111 (1) (2015) 98–136.
[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.
[16] T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. Zitnick, Microsoft coco: common objects in context, in: European Conference on Computer Vision, 2014, pp. 740–755.
[17] Www: Web page of the em segmentation challenge, (http://brainiac2.mit.edu/isbi_challenge/), Accessed May 2, 2012.
[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations, 2015, pp. 248–255.
[19] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[20] S. Zheng, S. Jayasumana, B.R. Paredes, V. Vineet, Z.Z. Su, D.L. Du, C. Huang, P.H. Torr, Conditional random fields as recurrent neural networks, in: IEEE International Conference on Computer Vision, 2015, pp. 1529–1537.
[21] G.S. Lin, C.H. Shen, D.H. Van, I. Reid, Exploring context with deep structured models for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1352–1366.
[22] L.C. Chen, Y.K. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: European Conference on Computer Vision, 2018, pp. 1–18.
[23] H.H. Ding, X.D. Jiang, B. Shuai, A.Q. Liu, G. Wang, Semantic segmentation with context encoding and multi-path decoding, IEEE Trans. Image Process. 29 (1) (2020) 3520–3533.
[24] P. Bilinski, V. Prisacariu, Dense decoder shortcut connections for single-pass semantic segmentation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2018, pp. 6596–6605.
[25] G.S. Lin, F.Y. Liu, A. Milan, C.H. Shen, I. Reid, Refinenet: multi-path refinement networks for dense prediction, IEEE Trans. Pattern Anal. Mach. Intell. 42 (5) (2020) 1228–1242.
[26] T. Pohlen, A. Hermans, M. Mathias, B. Leibe, Full-resolution residual networks for semantic segmentation in street scenes, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2017, pp. 3309–3318.
[27] W.B. Yang, Q. Zhou, J.N. Lu, X.F. Wu, S.F. Zhang, L.J. Latecki, Dense deconvolutional network for semantic segmentation, in: IEEE International Conference on Image Processing, 2018, pp. 1573–1577.
[28] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, F.F. Li, ImageNet: a large-scale hierarchical image database, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[29] Z. Zhu, M.D. Xu, S. Bai, T.T. Huang, X. Bai, Asymmetric non-local neural networks for semantic segmentation, in: IEEE International Conference on Computer Vision, 2019, pp. 593–602.
[30] F. Zhang, Y.Q. Chen, Z.H. Li, Z.B. Hong, J.T. Liu, F.F. Ma, J.Y. Han, E. Ding, AcfNet: attentional class feature network for semantic segmentation, in: IEEE International Conference on Computer Vision, 2019, pp. 6797–6806.

[31] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, 2015, pp. 448–456.
[32] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: International Conference on Artificial Intelligence and Statistics, 2012, pp. 315–323.
[33] D. Kingma, J. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, 2015, pp. 1–11.
[34] D. Eigen, R. Fergus, Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, in: IEEE International Conference on Computer Vision, 2015, pp. 2650–2658.
[35] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, J. Malik, Semantic contours from inverse detectors, in: IEEE International Conference on Computer Vision, 2011, pp. 991–998.
[36] Q. Zhou, W. Yang, G. Gao, W. Ou, H. Lu, J. Chen, L.J. Latecki, Multi-scale deep context convolutional neural networks for semantic segmentation, World Wide Web 22 (3) (2019) 555–570.
[37] H. Zhang, H. Zhang, C.G. Wang, J.Y. Xie, Co-occurrent features in semantic segmentation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2019, pp. 548–557.
[38] J.J. He, Z.Y. Deng, L. Zhou, Y.L. Wang, Y. Qiao, Adaptive pyramid context network for semantic segmentation, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2019, pp. 7511–7520.
[39] K.M. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.
[40] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, in: ACM International Conference on Multimedia, 2014, pp. 675–678.
[41] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: International Conference on Computational Statistics, 2010, pp. 177–186.
[42] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

Quan Zhou received Ph.D. degree in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China in 2013. Now he is an associated professor in the college of Telecommunications and Information engineering at Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition.

Xiaofu Wu received the B.S. and M.S. degrees in electrical engineering from the Nanjing Institute of Communications Engineering, Nanjing, China, in 1996 and 1999, respectively, and the Ph.D. degree in electrical engineering from the Peking University, Beijing, China, in 2005. He is now a full Professor of Nanjing University of Posts and Telecommunications. His research interests include, machine learning, and computer vision.

Suofei Zhang received the Ph.D. degree in School of Information Science and Engineering from Southeast University in 2013. In 2013, he joined the School of Internet of Things at the Nanjing University of Posts and Telecommunications. His research interests include computer vision, video surveillance, real-time object tracking and deep learning based image processing.

Bin Kang received the M.S. degree in Circuits and Systems from Lanzhou University, and the Ph.D. degree in Electrical Engineering from Nanjing University of Posts and Telecommunications, in 2011 and 2016, respectively. He is currently a lecturer at College of Internet of Things, Nanjing University of Posts and Telecommunications. His research interests include computer vision and pattern recognition.

Zongyuan Ge received the bachelor's degree in electrical engineering from The Australian National University, Australia, in 2012, and the Ph.D. degree in engineering from the Queensland University of Technology, Australia, in 2016. He is currently a Research Scientist at Monash University, Australia. His research interests are face verification, and medical image processing.

Longin Jan Latecki received the Ph.D. degree in computer science from Hamburg University, Germany, in 1992. He is a professor of computer science at Temple University, Philadelphia. His main research interests include shape representation and similarity, object detection and recognition in images, robot perception, data mining, and digital geometry. He is now the Associate Editors-in-Chief of Pattern Recognition and an editorial board member of the International Journal of Mathematical Imaging.