SCIENCE CHINA
Information Sciences

February 2020, Vol. 63 120106:1–120106:14
https://doi.org/10.1007/s11432-019-2738-y

• RESEARCH PAPER •
Special Focus on Deep Learning for Computer Vision

Preserving details in semantics-aware context for scene parsing

Shuai MA[1], Yanwei PANG[1*], Jing PAN[2] & Ling SHAO[1,3]

[1]*Tianjin Key Laboratory of Brain-Inspired Intelligence Technology, School of Electrical and Information Engineering, Tianjin University, Tianjin* 300072*, China;*
[2]*School of Electronic Engineering, Tianjin University of Technology and Education, Tianjin* 300222*, China;*
[3]*Inception Institute of Artificial Intelligence, Abu Dhabi* 999041*, UAE*

Received 13 November 2019/Revised 6 December 2019/Accepted 24 December 2019/Published online 15 January 2020

Abstract Great success of scene parsing (also known as, semantic segmentation) has been achieved with the pipeline of fully convolutional networks (FCNs). Nevertheless, there are a lot of segmentation failures caused by large similarities between local appearances. To alleviate the problem, most of existing methods attempt to improve the global view of FCNs by introducing different contextual modules. Though the reconstructed high resolution output of these methods is of rich semantics, it cannot faithfully recover the fine image details owing to lack of desired precise low-level information. To overcome the problem, we propose to improve the spatial decoding process through embedding possibly lost low-level information in a principled way. To this end, we make the following three contributions. First, we propose a semantics conformity module to make low-level features variations agnostic. Second, we introduce semantics into the conformed low level features through guidance from semantically aware features. Finally, we institute the availability of various possible contextual features at feature fusion to enrich context information. The proposed approach demonstrates competitive performance on challenging PASCAL VOC 2012, Cityscapes, and ADE20K benchmarks in comparison to the state-of-the-art methods.

Keywords fully convolutional networks, semantic segmentation, cityscapes, semantic-aware context

Citation Ma S, Pang Y W, Pan J, et al. Preserving details in semantics-aware context for scene parsing. Sci China Inf Sci, 2020, 63(2): 120106, https://doi.org/10.1007/s11432-019-2738-y

1 Introduction

The scene parsing task (or semantic segmentation) assigns a unique category label to each pixel in an image. It has enjoyed a central place in computer vision for decades owing to its numerous important applications, e.g., autonomous driving [1] and domestic robots. The pioneering fully convolutional networks (FCNs) [2] for semantic segmentation outperformed prior art relying on hand-crafted methods. Contrary to classification networks, FCNs replaces fully connected layers with convolutional layers, allowing pixel-level category prediction. This prediction is, however, made utilizing only the local view of the image. Under unconstrained semantic segmentation, local appearance similarity is often unsolvable with the limited context.

Several solutions have been proposed to circumvent this problem through capturing multi-scale context. One way is taking image pyramid as input to deep convolutional neural networks (DCNNs) [3,4]. Besides, encoder-decoder networks [5–8] downsample gradually to widen network context. Further, for the same purpose, there are modules either as post-processing [9,10] after original network or trainable [3,11–13]

* Corresponding author (email: pyw@tju.edu.cn)

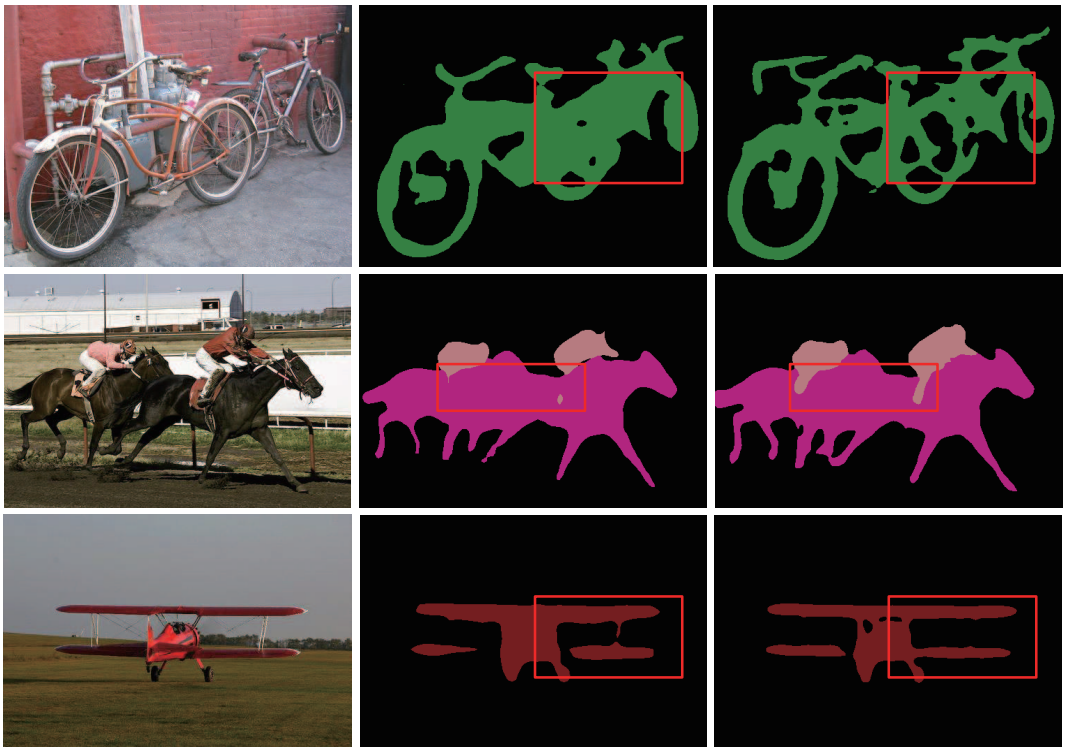

Figure 1 (Color online) First column: original image; second column: baseline predictions; and last column: predictions from the proposed approach. Baseline fails to recover many details, inside object or around object boundaries (marked in red boxes). Whereas the proposed approach convincingly segments them, for instance, fine spokes of wheel in the first image, leg of the rider in the second image, right wing boundaries of the plane in the last image.

with the original network. Among other alternatives, global pooling operations [14], spatial pyramid pooling [15], and spatial attention blocks [16] have proved beneficial for context aggregation.

Similarly, techniques such as Atrous convolution (also known as, dilated convolution) [10,17] have been proved to be effective in gathering context. It fills zero between the neighbor elements in a convolutional kernel for increasing receptive fields while keeping the number of parameters same. Atrous convolution has been exploited in parallel [18] or cascaded arrangement [19] with variable dilation rates for achieving multiple fields of view. While these techniques address the scene context integration, they are limited in their capability to retain local features that relate to fine image details, e.g., around the object boundaries as shown in Figure 1 where atrous spatial pyramid pooling (ASPP) [18] fails to accurately segment the fine spokes in the round wheels of bikes.

We hypothesize that the existing methods sacrifice fine image details in an effort to incorporate a broader context. In this paper, we intend to complement contextually richer, coarser representation with spatially richer details through leveraging local (high resolution) features after conforming and compensating the latter according to the former.

Contributions. We intend to improve the spatial decoding process through fusing missing local information in a simple, yet principled approach. Firstly, we propose semantics conformity module to make low level features invariant to local deformations. This is a so-called registration of local features to a reference frame. Secondly, we induce semantics into the conformed low level features in guidance from semantically-aware features. This is achieved by attending to the joint feature via gated attention mechanism. Finally, we institute the availability of various possible contextual features at feature fusion to make available context information while spatial decoding process. The proposed approach displays favourable performance across challenging PASCAL VOC 2012, Cityscapes and ADE20K benchmarks in comparison to the state-of-the-art methods.

2 Related work

FCNs [2] displayed major improvements over prior art in semantic segmentation. Since then considerable effort went into enhancing its global field-of-view through external modules-integrating multi-scale context, encoder-decoder architectures, and spatial pyramid pooling, to further boost its performance.

Although FCNs entirely features convolutions intertwined with downsampling steps for deep representation, they lack the long-range context for resolving ambiguities often caused by subtle similarity of local appearance. A few methods incorporate CRF (conditional random fields) based module after FCN to encode contextual information [9, 10]. These approaches, however, employ CRF based module as a post-processing step and are not end-to-end trainable. To overcome the drawback, some methods targeted joint training of CRF with DCNN [3, 11, 12]. Some recent studies have proposed a duo of spatial and channel attentions modules [16] or graph convolution over latent feature space [20] to introduce long-range contextual information. In this study, we also use FCNs based method deploying an extra module to capture wider scene context.

Encoder design in encoder-decoder type methods is typically a popular bottom-up trained classification network such as ResNet [21] or VGG [22]. Encoder captures global view through successive downsampling at low resolution whereas the decoder recovers the spatial resolution required in pixel-level predictions tasks. Skip connections towards decoder from encoder help it in better recovering the spatial details [5, 6, 23]. Stacked transposed convolutional layers were utilized in [24, 25] to gradually upsample the low resolution feature maps. Ghiasi et al. [26] proposed a Laplacian pyramid reconstruction network to refine particular low resolution feature maps. Finally, such models have shown success in other vision tasks such as human pose estimation [27], saliency detection [28], and object detection [29, 30]. These methods can recover spatial details, albeit may perform poorly in segmenting fine object boundaries. It maybe owing to under and/or naive utilization of low-level features while reconstructing high resolution output.

Attention mechanism is a crucial module for semantic segmentation to improve the performance. Because the SE-Net [31] proposes a channel attention module, some methods have used or revised this module as a part of their networks. EncNet [32] uses the channel attention module at the tail of whole network for catching more context information. DANet [16] proposes two non-local attention modules to enhance global information. For reducing the calculation, ANNN [33] changes the attention generation method with asymmetric structure and maintains accuracy. The attention become more and more complicated but the performance improves gradually. Our method provides a simple but effective attention module which makes full use of different semantic level information.

Spatial pyramid pooling (SPP) is another way to embody multiple feature scales. It pools input features from multiple field of views. PSPNet [15] and DeepLab [10] use SPP at several grid scales, while Parsenet [14] further exploits image-level features. FSG [34] and MDCCNet [35] also use multi-scale features to capture more context information. Lately, ASPP [18] approach probes features from multiple field of views. Dense ASPP [19] accomplishes the same however in a denser way and at a much larger scale. Such approaches encode multi-scale context in the final output, but they lack the (required) spatial details possibly lost after the successive downsampling steps in the encoder. Proposed approach aims to induce these details via readily available low level backbone features, however, after adjusting them first and then compensating them with the semantics through the proposed modules.

3 Proposed approach

In this section, we describe the details of our proposed method for semantic segmentation, by first reviewing the standard encoder-based method employing multi-scale contextual block below, and then explaining the novel mechanisms introduced for boosting their segmentation capabilities from Subsection 3.1 onwards.

Baseline settings. We follow the encoder-based architecture integrating multi-scale features for crafting overall design as in [18] (see Figure 2(a)). The encoder is a ResNet-101 model pretrained on

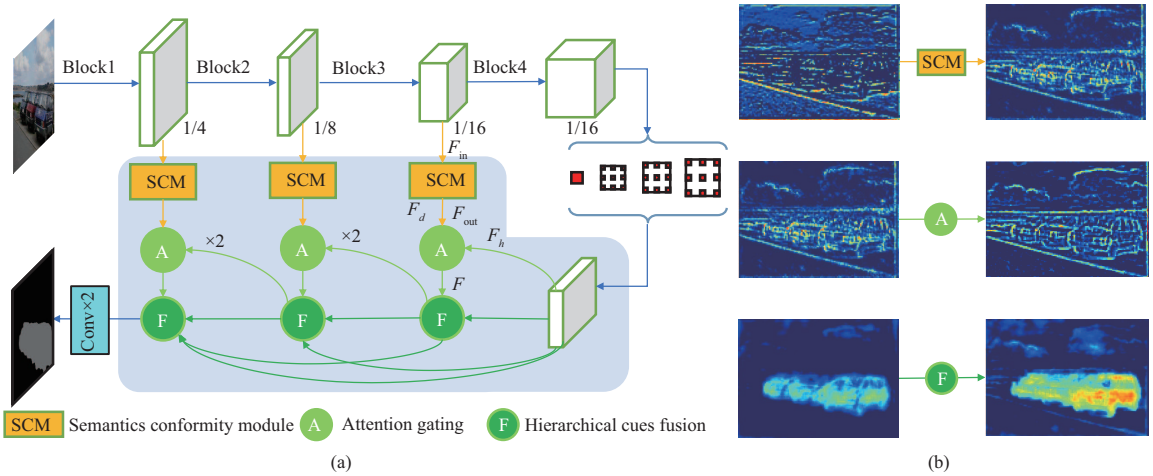

Figure 2 (Color online) The network architecture of our overall framework. We propose semantic conformity to adjust to local deformations in possibly messy high resolution representation and design attention gating to compensate semantics in high resolution features through guidance for enriching spatial details in context-aware, semantically richer, low resolution feature maps. Finally, we introduce hierarchical cues fusion along the proposed spatial decoding to enrich contextual information after the fusion of (adjusted and compensated) low level features. We display the impact of SCM, SCM+AG, and SCM+AG+HCF on feature maps in Figure 2(b). SCM reduces noise through adjusting local features, while AG on top of it improves the semantics such as enhancing cars boundaries, and finally HCF further enhances foreground (cars) and simultaneously suppresses background with the fusion of complementary contextual features. (a) Overview of our method; (b) the effect of different module.

ImageNet unless otherwise specified. For image classification tasks, the output stride for this model is always 32. In contrast, for semantic segmentation, it is typically modified to 16 or even 8 to improve spatial resolution. It is achieved by replacing the convolution layers with stride = 2 by stride = 1 and introducing dilated convolutions with rate = 2 and rate = 4. Further, towards the tail of encoder, dilated convolutions of various rates are organized in parallel scheme to capture multi-scale scene context.

Though the aforementioned encoder-based frameworks are capable of modelling higher order dependencies for improving discriminative ability, they often fail to preserve fine image details typically found around the objects boundaries. The second column in Figure 1 illustrates representative failure cases of such methods. The culprits are successive downsampling operations present in the encoder (ResNet-101) backbone, which destroy spatial information in the wake of increasing network field-of-view. Skip connections [5, 6] give rise to loss of spatial details which can be solved by simple spatial reconstruction. The direct skip based connection(s), however, are sub-optimal owing to the complementary nature of two feature representations. We believe a step forward is to merge only after alignment while conserving context. That is, the spatial resolution recovery from coarse semantics-aware low resolution signal can be improved through inducing low level representation after the alignment of latter with the former while preserving various types of contextual information.

3.1 Semantics conformity module

Encoders based on pretrained ImageNet models such as ResNet generate features at various semantic levels. Initial features, from the shallow part of the network, are of high resolution and spatially rich in details such as edges and corners. Deep features, from the tail of the network, are of low resolution — relatively coarser — such as concepts and objects. While reconstructing high resolution segmentation output from the output of such encoders, FCNs [2] and related methods [5, 6] leverage low-level features in different manifestations of direct skip connections for filling spatial details. As discussed, this might reduce, instead improve, the segmentation ability of the coarser output because the low level information is often noisy and therefore messy.

We identify that the local variations owing to changes in pose, scale and object deformations act as a bottleneck while mapping the low-level image details to their true semantics. For this purpose, we

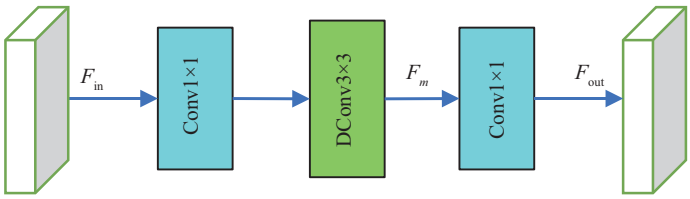

Figure 3 (Color online) Semantics conformity module. It adjusts local variations caused by various geometric deformations via bottleneck block featuring deformable convolution, thereby preparing them before semantics introduction.

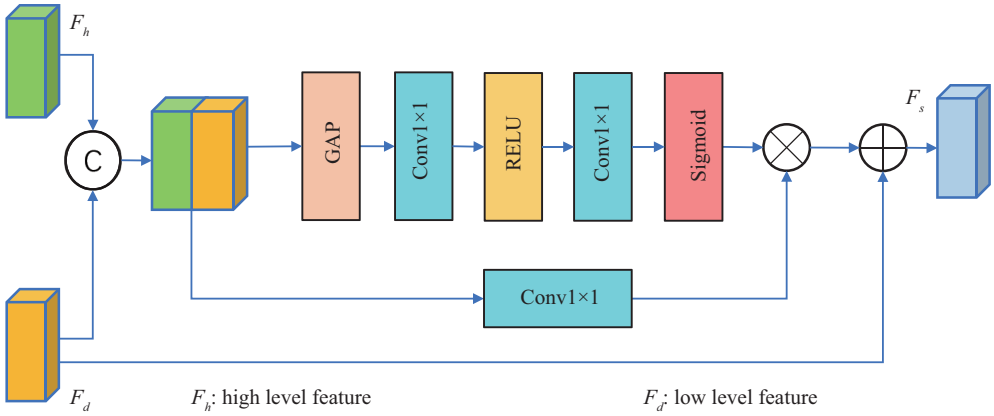

Figure 4 (Color online) Semantics enhancement via attention gating. We compensate semantics in high resolution features through guidance for enriching spatial details in context-aware, semantically richer, low resolution feature maps.

introduce the semantics conformity module (as shown in Figure 3) featuring deformable convolution (DConv) [36] that is robust to local deformations. It can be imagined as a so-called canonical registration to a common frame of reference, performed to prepare low level features for semantics introduction.

Specifically, let us denote the input (local) features as $F_{\text{in}} \in \mathbb{R}^{C\times H\times W}$. First, a 1×1 convolution is employed to decrease the number of channel dimensions from C to C', where $C' \ll C$, for the sake of reducing computations. Next, a 3×3 deformable convolution generates new features $F_m \in \mathbb{R}^{C'\times H\times W}$ in a local, dense, and adaptive manner via learnable 2D offsets. Finally, another 1×1 convolution projects the new features back to input channel dimensions C to get $F_{\text{out}} \in \mathbb{R}^{C\times H\times W}$ as the output of this module.

3.2 Semantics enhancement via attention gating

We intend to compensate the semantic vent in low level feature representation through guidance. To actualize this, we propose attention gating mechanism that attends to joint feature representation in order to produce compensation maps. Because joint feature combines low as well as high level features, the resulting compensation maps are extracted through so-called implicit guidance from contextually richer high level features. This compensation map is then used to further boost the semantic capability of low level, spatially richer features, thereby further aligning the two representations.

The process for semantics compensation through attending joint feature is described below and displayed in Figure 4. The first input to the module is $F_d \in \mathbb{R}^{C'\times H\times W}$, which is an adjusted low level feature representation. The second input is $F_h \in \mathbb{R}^{C'\times H\times W}$, which corresponds to (relatively) high level feature representation. We design joint feature representation after concatenating both F_d and F_h. Note, we simply upsample F_h or reduce its channels if the resolution or the channel number does not match, respectively, with F_d. This joint feature is then subjected to global average pooling (GAP) to achieve context aggregation from full spatial extent. We then reduce number of channels from $2C$ back to C before applying sigmoid operation to produce a probability map. Joint features are modulated through this probability map after doing 1×1 convolution for making channels equivalent to C to obtain compensation maps. Finally, these compensation maps are added element-wise to F_d for improving semantics. We denote the output of this block as F_s.

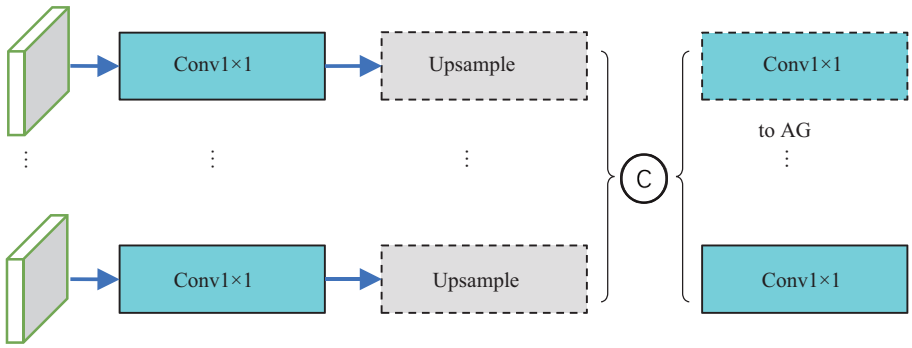

Figure 5 (Color online) Hierarchical cue fusion. We tend to enrich the context information while fusing the contradictory (adjusted and compensated) low level features through exploiting all possible previous contextual features.

3.3 Hierarchical cues fusion

During spatial decoding, consistent fusion of local features with their higher counterparts, they may be adjusted and compensated accordingly, could potentially weaken the impact of global information in the resulting features leading to final output. This can result in reducing the network ability in discerning, possibly local, appearance similarities in a complex scene.

We try to balance the contribution of two contradictory, yet equally important features, i.e., low and high level along reconstructing the segmentation output. Dense feature connectivity has been successfully exploited in the realm of classification networks for improved gradient back propagation and feature reuse [37]. Further, dense connections have also been exploited for semantic segmentation, however, to capture features over a larger scale in a dense manner using dilated convolutions in [19]. To this end, we introduce hierarchical cues fusion in the decoder as illustrated in Figure 5. In HCF building blocks, the inputs are from AG and more higher semantic nodes. We use 1×1 convolution to reduce the dimensions of channels. Here we use dotted boxes to wrap some unsample layers and convolution layers. It indicates that these layers may not exist because some features for concatenating are in same size. After concatenating, we send one branch features to next AG, and others are sent to each more shallow semantic nodes.

4 Experimental results

In this section, we evaluate the proposed approach through reporting ablation studies and comparison (both quantitative and qualitative) with existing art after briefing about used benchmarks and adopted training protocol.

Datasets. PASCAL VOC 2012 semantic segmentation benchmark [38] features 20 foreground object classes and a background class; it consists 1464 training images, 1449 validation images, and 1456 testing images. Ref. [39] increased the number of training images to 10582 after augmenting extra annotations.

Cityscapes [40] is a large, urban street scene dataset captured from car perspective. It comprises 5K high quality images photographed from 50 different cities and is annotated at pixel-level with labels from 19 different semantic classes. Train, Val, and test split is 2979, 500, and 1525 images, respectively. Note, we do not employ coarse data in our experiments.

ADE20K [41] is a recently launched dataset with 20210 images for training and 2K images for validation. It features 150 semantic categories and many different complex scenes.

Overall training protocol. We perform data enhancement: mean subtraction, horizontal flipping, scale noise (0.75–2.0) on the training images. We then crop them to fixed size. We use poly learning rate policy, as in [18], where the initial learning rate is multiplied by $(1 - \frac{\text{iter}}{\text{max\_iter}})^{\text{power}}$ after each learning iteration with power = 0.9. Momentum and weight decay coefficients are set to 0.9 and 0.0001, respectively. To ensure appropriate batch normalization statistics while training, we restrict the output stride = 16 and thus retain a minimum batch size to 16. We fix batch normalization parameters with

Table 1 Performance of SCM with different convolutional variants. We see the deformable convolution adjusts to local variations better amongst others.

SCM	mIoU (%)
Standard 3×3 convolution	77.40
Dilated 3×3 convolution (rate = 2)	77.61
Deformable 3×3 convolution	**77.76**

Table 2 Ablation study on PASCAL VOC 2012 validation set[a)]. We observe that AG requires SCM to adapt to local feature variations and thus shows improvement. Further, HCF boosts performance both in isolation as well as along with other components.

Method	SCM	AG	HCF	mIoU (%)
Deeplabv3 [18]	–	–	–	77.21
Ours	✓	–	–	77.76
	–	✓	–	77.55
	–	–	✓	77.91
	✓	✓	–	78.00
	✓	–	✓	78.11
	✓	✓	✓	**78.46**

a) Abbreviations used stand for the following: SCM is the semantics conformity module, AG is the semantics enhancement via attention gating component, and HCF is the hierarchical cues fusion process.

decay = 0.9997.

4.1 PASCAL VOC 2012 dataset results

To evaluate on PASCAL VOC 2012 validation set, we set the image crop size to be 513 and train with an initial learning rate of 0.007 for 40K iterations on the train-aug set (containing 10582 images) keeping an output stride = 16. Further, we freeze our batch normalization layers and fine tune with the reduced learning rate 0.0001 for 40K iterations on trainval set employing output stride = 8.

4.1.1 *Ablation study*

We conduct ablation study to highlight the contribution of proposed components in the overall framework. We begin by investigating the importance of SCM as it is the first step towards semantics introduction in low level features. Table 1 reports the performance improvement over baseline after introducing SCM with different possible convolutional variants. SCM realizing deformable convolution outperforms all other choices and achieves 77.76% mIoU. It seems obvious because free-form sampling of kernel offsets allows adaptive adjustment of receptive field to various local deformations.

Table 2 [18] displays the performance contribution of AG under two different settings. AG complementing SCM boosts the performance from 77.76% to 78.00% mIoU. This supports our intuition behind making local features variation agnostic before inducing semantics.

Finally, we reveal the performance contribution of HCF in retaining global information in the recovered spatial resolution under two settings: (1) in isolation, i.e., without SCM and AG; (2) after consistent fusion of adjusted and compensated local features, i.e., with SCM and AG. From Table 2, as expected, we see HCF improving performance over baseline in isolation as well as in combination with other components by almost same percentage. HCF exploits (intrinsic) synergy between various semantics-aware complementary features.

Figure 6 shows the impact of adding each proposed module qualitatively. We see that each of the modules progressively improves the segmentation quality of baseline. For instance, the baseline segmentation result contains many obvious holes and unsmooth edges. SCM module restrains small holes because it broadens receptive field and eliminates changes in object deformations. AG module fills large holes by using high level feature guidance to compensate the semantic vent. HCF module makes a dense connection between different semantic level features. It effectively enhances the tail feature of encoder which

Figure 6 (Color online) Contribution of each proposed module towards final segmentation. Two different examples are shown. (a) Image; (b) baseline; (c) SCM; (d) SCM+AG; (e) all.

Table 3 Performance comparisons upon employing different inference strategies on the PASCAL VOC 2012 validation set similar to [18, 42][a)]. We see each inference strategy boosting performance by noticeable margins, however, the best performance of 80.46% is obtained after using MS and Flip inputs with output stride = 8.

Method	OS = 16	OS = 8	MS	Flip	mIoU (%)
Ours	✓	–	–	–	78.46
	✓	–	✓	–	79.65
	✓	–	✓	✓	80.09
	–	✓	–	–	78.76
	–	✓	✓	–	79.97
	–	✓	✓	✓	**80.46**

a) OS: output stride. MS: multi-scale inputs. Flip: adding left-right flipped inputs.

contains abundant context information in order to make objects more complete (such as cheetah mark on the bus).

Improvement with different inference strategies. Similar to [18, 42], we experiment with different inference strategies on PASCAL VOC 2012 validation set and the results are reported in Table 3. Multi-scale (MS) inputs strategy with output stride = 16 boosts performance from 78.46% to 79.65% and inclusion of image flips (Flip) on top further peaks performance to 80.09%. We then increase the segmentation output resolution by 2 times after reducing the output stride to 8 and observe 0.3% performance gain compared to stride = 16 counterpart. Employing MS inputs and then including image flipping on top at output stride = 8 produces the best performance of 80.46% amongst all.

4.1.2 *Comparison with the state-of-the-art*

Figure 7 draws qualitative comparison between the baseline and our proposed approach on PASCAL VOC 2012 validation set. Three different example images are shown in the first column. The second and the third columns display segmentation outputs from baseline and our approach, respectively. The ground truth segmentations are also shown for comparison in the last column. We see that the proposed approach is distinctly superior to baseline in reducing both false positives and false negatives in all scenarios. For instance, consider the second image comprising a bird surrounded in leaves. The baseline mistakes leave as part of the bird more than our approach. Further, it almost misses bird legs whereas our approach finely recovers them.

We compare category-wise results with the existing state-of-the-art on PASCAL VOC 2012 test set in Table 4 [2, 3, 10, 12, 15, 24, 34, 35, 43, 44]. Proposed approach shows improved performance over the previous best methods in 6 out of 20 different categories.

4.2 Results on Cityscapes dataset

In Cityscapes dataset experiments, following [18], we crop images to 769×769 resolution while training to reduce memory footprint. First, we set the initial learning rate to 0.007 and train for 100K iterations

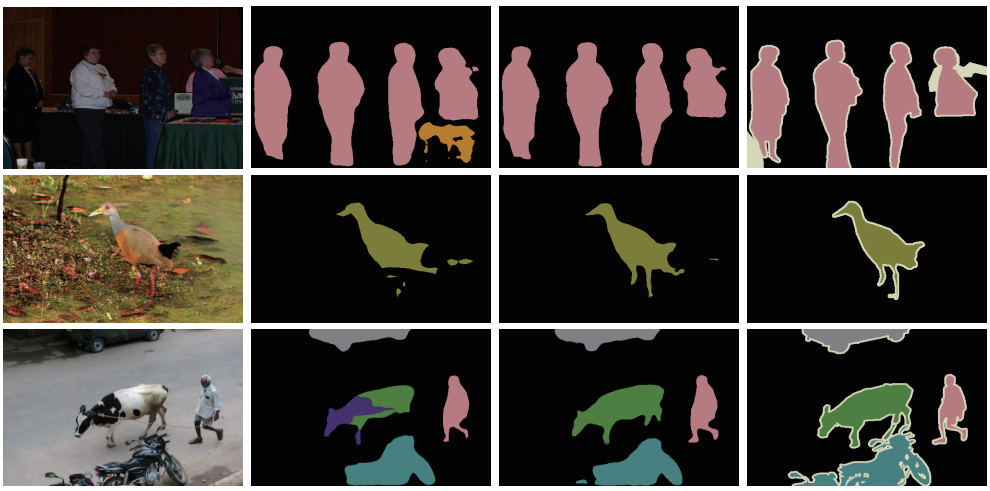

Figure 7 (Color online) Qualitative comparison between baseline and the proposed approach on PASCAL VOC 2012 validation set. Further, it almost misses bird legs whereas our approach finely recovers them. (a) Image; (b) baseline; (c) ours; (d) groundtruth.

Table 4 Per-class comparison of the proposed approach with the state-of-the-art the PASCAL VOC 2012 test set

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	
FCN [2]	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	
FSG [34]	–	–	–	–	–	–	–	–	–	–	
DeepLab [10]	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	
CRF-RNN [12]	87.5	39.0	79.7	64.2	68.3	87.6	80.8	84.4	30.4	78.2	
DeconvNet [24]	89.9	39.3	79.7	63.9	68.2	87.4	81.2	86.1	28.5	77.0	
DPN [43]	87.7	59.4	78.4	64.9	70.3	89.3	83.5	86.1	31.7	79.9	
Piecewise [3]	90.6	37.6	80.0	67.8	74.4	92.0	85.2	86.2	39.1	81.2	
MDCCNet [35]	87.6	43.7	85.3	72.3	83.0	91.7	86.5	89.9	43.8	80.5	
ResNet38 [44]	94.4	**72.9**	**94.9**	68.8	**78.4**	90.6	90.0	92.1	40.1	90.4	
PSPNet [15]	91.8	71.9	94.7	**71.2**	75.8	95.2	89.9	**95.9**	39.3	**90.7**	
Ours	**95.4**	72.7	93.2	69.6	77.1	**95.3**	**91.5**	94.9	**40.2**	87.6	
Method	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU (%)
FCN [2]	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
FSG [34]	–	–	–	–	–	–	–	–	–	–	64.4
DeepLab [10]	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
CRF-RNN [12]	60.4	80.5	77.8	83.1	80.6	59.5	82.8	47.8	78.3	67.1	72.0
DeconvNet [24]	62.0	79.0	80.3	83.6	80.2	58.8	83.4	54.3	80.7	65.0	72.5
DPN [43]	62.6	81.9	80.0	83.5	82.3	60.5	83.2	53.4	77.9	65.0	74.1
Piecewise [3]	58.9	83.8	83.9	84.3	84.8	62.1	83.2	58.2	80.8	72.3	75.3
MDCCNet [35]	50.6	84.2	79.7	81.0	86.6	61.5	85.7	55.6	86.3	74.8	75.5
ResNet38 [44]	**71.7**	89.9	93.7	91.0	89.1	71.3	**90.7**	61.3	**87.7**	**78.1**	82.5
PSPNet [15]	**71.7**	90.5	**94.5**	88.8	**89.6**	**72.8**	89.6	**64.0**	85.1	76.3	**82.6**
Ours	69.4	**91.0**	91.0	**92.3**	88.7	64.8	89.4	61.1	84.7	74.2	81.9

with a batch size of 16 on train set. Then, we freeze batch normalization layers, set the learning rate to 0.001 and finetune the model with a batch size of 16 for 90K iterations on trainval set.

Improvement with different inference strategies. We also experiment with different inference strategies on Cityscapes validation set to see performance improvement. The results are reported in Table 5. All different inference strategies show performance gains. Multi-scale input strategy provides the highest absolute gain of 1.17% and 1.14% at output stride = 16 and output stride = 8, respectively, amongst others reported.

Table 5 Experiments with different inference strategies on Cityscapes validation set. All different inference strategies show performance gains. Noticeably, MS input strategy provides the highest absolute gain of 1.17% and 1.14% at output stride = 16 and output stride = 8, respectively, amongst others[a)].

Method	OS = 16	OS = 8	MS	Flip	mIoU (%)
Ours	✓	–	–	–	77.90
	✓	–	✓	–	79.07
	✓	–	✓	✓	79.20
	–	✓	–	–	78.14
	–	✓	✓	–	79.28
	–	✓	✓	✓	**79.40**

a) OS: output stride. MS: multi-scale inputs. Flip: adding left-right flipped inputs.

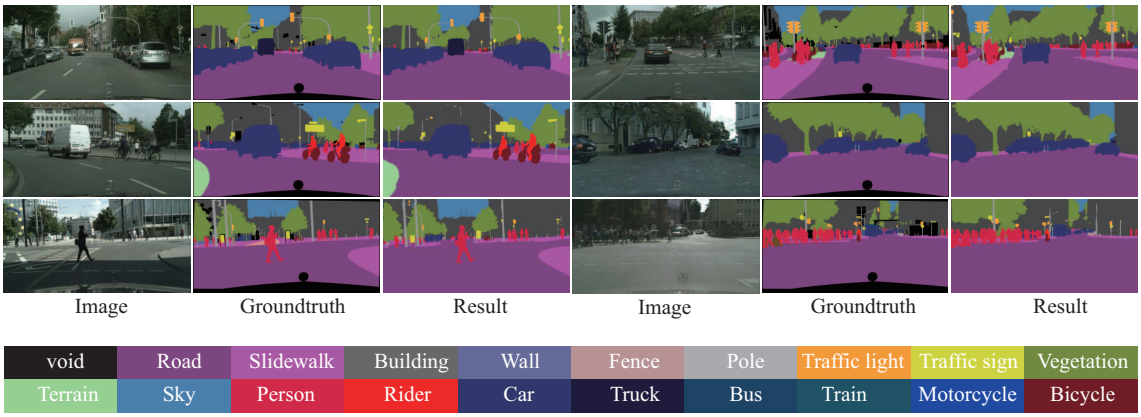

Figure 8 (Color online) Some example segmentation maps from our approach on Cityscapes dataset. Proposed approach well distinguishes objects appearing at various scales while preserving details across their boundaries. Example objects are people and cars.

Table 6 Category-wise comparison of proposed approach with the existing state-of-the-art on Cityscapes test set. Note, our result is obtained without using coarse annotations.

Method	road	swalk	build	wall	fence	pole	tlight	sign	veg	terrain
CRF-RNN [12]	96.3	73.9	88.2	47.6	41.3	35.2	49.5	59.7	90.6	66.1
FCN [2]	97.4	78.4	89.2	34.9	44.2	47.4	60.1	65.0	91.4	69.3
Dilation10 [17]	97.6	79.2	89.9	37.3	47.6	53.2	58.6	65.2	91.8	69.4
DeepLab [10]	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4
RefineNet [45]	98.2	83.3	91.3	47.8	50.4	56.1	66.9	71.3	92.3	70.3
GCN [46]	–	–	–	–	–	–	–	–	–	–
DUC [47]	98.5	85.5	92.8	**58.6**	55.5	65.0	73.5	77.9	93.3	72.0
PSPNet [15]	**98.6**	**86.2**	92.9	50.8	58.8	64.0	**75.6**	79.0	93.4	72.3
AAF [48]	98.5	85.6	93.0	53.8	59.0	65.9	75.0	78.4	**93.7**	**72.4**
Ours	**98.6**	**86.2**	**93.1**	54.2	**60.5**	**67.4**	74.9	**79.1**	93.6	71.3
Method	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU (%)
CRF-RNN [12]	93.5	70.4	34.7	90.1	39.2	57.5	55.4	43.9	54.6	62.5
FCN [2]	93.9	77.1	51.4	92.6	35.3	48.6	46.5	51.6	66.8	65.3
Dilation10 [17]	93.7	78.9	55.0	93.3	45.5	53.4	47.7	52.2	66.0	67.1
DeepLab [10]	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8	70.4
RefineNet [45]	94.8	80.9	63.3	94.5	64.6	76.1	64.3	62.2	70.0	73.6
GCN [46]	–	–	–	–	–	–	–	–	–	76.9
DUC [47]	95.2	84.8	68.5	95.4	70.9	78.8	68.7	65.9	73.8	77.6
PSPNet [15]	95.4	86.5	**71.3**	95.9	68.2	79.5	73.8	**69.5**	**77.2**	78.4
AAF [48]	95.6	86.4	70.5	95.9	**73.9**	82.7	76.9	68.7	76.4	79.1
Ours	**95.8**	**86.7**	71.0	**96.0**	73.0	**87.8**	**85.9**	68.8	76.4	80.0

Table 7 Performance comparison of our method with the existing state-of-the-art approaches on ADE20K validation set. Employing the same backbone network, it surpasses PSPNet [15].

Method	Backbone	mIoU (%)
FCN [2]	–	29.39
SegNet [49]	–	21.64
DilatedNet [17]	–	32.31
CascadeNet [50]	–	34.90
RefineNet [45]	ResNet152	40.70
PSPNet [15]	ResNet101	43.29
Ours	ResNet101	**43.76**

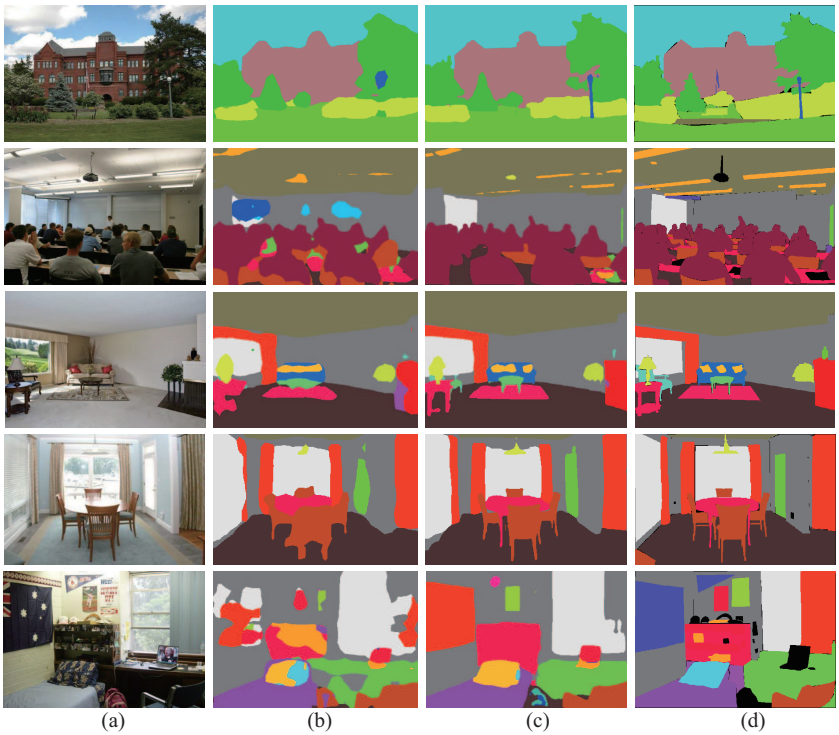

Figure 9 (Color online) Example segmentation maps from proposed approach on the ADE20K validation set. Our approach accurately preserves quite delicate boundary details around variably sized objects, for instance, segmentation output for the intermingled leaves of the tree in first example image. (a) Image; (b) baseline; (c) ours; (d) groundtruth.

Comparison with the state-of-the-art. Figure 8 reveals some example segmentation maps from our approach on Cityscapes dataset. Proposed approach can distinguish well objects appearing at various scales while preserving details across their boundaries. Example objects are people and cars.

Table 6 [2, 10, 12, 15, 17, 45–48] shows category-wise comparison of the proposed approach with the existing state-of-the-art on Cityscapes test set. Our method displays improved performance over above methods in 11 out of 19 classes.

4.3 Results on ADE20K dataset

We set the initial learning rate to 0.007, use a relatively larger batch size of 24 and train for 15K iterations. We crop original image to a fixed 513×513 resolution before training. We compare the proposed approach with the existing state-of-the-art segmentation methods on ADE20K dataset in Table 7 [2,15,17,45,49,50]. Using the same backbone network, it outperforms PSPNet [15] with an absolute gain of 0.47% and shows 43.76% mIoU.

Figure 9 displays some example segmentation maps from the proposed approach on the ADE20K validation set. Our approach accurately preserves quite delicate boundary details around variably sized

objects, for instance, the segmentation output for the intermingled leaves of the tree in the first example image.

5 Conclusion

In this paper, we proposed to improve the spatial decoding process for semantic segmentation through inducing possibly lost local information in a simple way. Semantics conformity module has been proposed to make low level features variation agnostic. Then, in our method, semantics are compensated into these conformed low level features via guidance from semantically-aware features through an attention gating mechanism. Finally, the synergy is exploited between various contextual features to enrich context information in spatial decoding process. It was found that the proposed approach achieves competitive performance on challenging PASCAL VOC 2012, Cityscapes, and ADE20K benchmarks in comparison to the state-of-the-art methods.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 61632018) and Science and Technology Innovation 2030: the Key Project of Next Generation of Artificial Intelligence (Grant No. 2018AAA01028).

References

1 Chen S T, Jian Z Q, Huang Y H, et al. Autonomous driving: cognitive construction and situation understanding. Sci China Inf Sci, 2019, 62: 081101

2 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 3431–3440

3 Lin G, Shen C, van Den Hengel A, et al. Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 3194–3203

4 Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 2650–2658

5 Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, 2015. 234–241

6 Shah S, Ghosh P, Davis L-S, et al. Stacked U-Nets: a no-frills approach to natural image segmentation. 2018. ArXiv: 1804.10343

7 Zhou Q, Wang Y, Liu J, et al. An open-source project for real-time image semantic segmentation. Sci China Inf Sci, 2019, 62: 227101

8 Huang T T, Xu Y C, Bai S, et al. Feature context learning for human parsing. Sci China Inf Sci, 2019, 62: 220101

9 Chen L-C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. 2014. ArXiv: 1412.7062

10 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 834–848

11 Schwing A-G, Urtasun R. Fully connected deep structured networks. 2015. ArXiv: 1503.02351

12 Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1529–1537

13 Sun H Q, Pang Y W. GlanceNets—efficient convolutional neural networks with adaptive hard example mining. Sci China Inf Sci, 2018, 61: 109101

14 Liu W, Rabinovich A, Berg A-C. Parsenet: looking wider to see better. 2015. ArXiv: 1506.04579

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

16 Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 3146–3154

17 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. 2015. ArXiv: 1511.07122

18 Chen L-C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation. 2017. ArXiv: 1706.05587

19 Yang M, Yu K, Zhang C, et al. Denseaspp for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 3684–3692

20 Chen Y, Rohrbach M, Yan Z, et al. Graph-based global reasoning networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, 2019. 433–442

21 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 770–778

22 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv: 1409.1556

23 Chen J, Lian Z H, Wang Y Z, et al. Irregular scene text detection via attention guided border labeling. Sci China Inf Sci, 2019, 62: 220103

24 Noh H, Hong S, Han B. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1520–1528

25 Jgou S, Drozdzal M, Vazquez D, et al. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 11–19

26 Ghiasi G, Fowlkes C-C. Laplacian pyramid reconstruction and refinement for semantic segmentation. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 519–534

27 Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: Proceedings of European Conference on Computer Vision, Amsterdam, 2016. 483–499

28 Liu N, Han J, Yang M-H. PiCANet: learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 3089–3098

29 Shrivastava A, Sukthankar R, Malik J, et al. Beyond skip connections: top-down modulation for object detection. 2016. ArXiv: 1612.06851

30 Lin T-Y, Dollr P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 2117–2125

31 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 7132–7141

32 Zhang H, Dana K, Shi J, et al. Context encoding for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 7151–7160

33 Zhu Z, Xu M, Bai S, et al. Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, Seoul, 2019. 593–602

34 Zhou Q, Zheng B, Zhu W, et al. Multi-scale context for scene labeling via flexible segmentation graph. Pattern Recogn, 2016, 59: 312–324

35 Zhou Q, Yang W, Gao G, et al. Multi-scale deep context convolutional neural networks for semantic segmentation. World Wide Web, 2019, 22: 555–570

36 Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, Venice, 2017. 764–773

37 Huang G, Liu Z, van Der Maaten L, et al. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 4700–4708

38 Everingham M, Eslami S M A, van Gool L, et al. The pascal visual object classes challenge: a retrospective. Int J Comput Vis, 2015, 111: 98–136

39 Hariharan B, Arbelez P, Bourdev L, et al. Semantic contours from inverse detectors. In: Proceedings of the IEEE International Conference on Computer Vision, Barcelona, 2011. 991–998

40 Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016. 3213–3223

41 Zhou B, Zhao H, Puig X, et al. Scene parsing through ADE20K dataset. In: Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, Honolulu, 2017. 633–641

42 Chen L-C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 801–818

43 Liu Z, Li X, Luo P, et al. Semantic image segmentation via deep parsing network. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 1377–1385

44 Wu Z, Shen C, van den Hengel A. Wider or deeper: revisiting the resnet model for visual recognition. Pattern Recogn, 2019, 90: 119–133

45 Lin G, Milan A, Shen C, et al. Refinenet: multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 1925–1934

46 Peng C, Zhang X, Yu G, et al. Large Kernel matters improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017. 4353–4361

47 Wang P, Chen P, Yuan Y, et al. Understanding convolution for semantic segmentation. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, 2018. 1451–1460

48 Ke T-W, Hwang J-J, Liu Z, et al. Adaptive affinity fields for semantic segmentation. In: Proceedings of European Conference on Computer Vision, Munich, 2018. 587–602

49 Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 2481–2495

50 Zhou B, Zhao H, Puig X, et al. Semantic understanding of scenes through the ADE20K dataset. Int J Comput Vis, 2019, 127: 302–321