

Dual context prior and refined prediction for semantic segmentation

Long Chen, Jiajie Liu, Han Li, Wujing Zhan, Baoding Zhou & Qingquan Li

To cite this article: Long Chen, Jiajie Liu, Han Li, Wujing Zhan, Baoding Zhou & Qingquan Li (2021) Dual context prior and refined prediction for semantic segmentation, Geo-spatial Information Science, 24:2, 228-240, DOI: [10.1080/10095020.2020.1785957](https://doi.org/10.1080/10095020.2020.1785957)

To link to this article: <https://doi.org/10.1080/10095020.2020.1785957>



© 2020 Wuhan University. Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 14 Jul 2020.



Submit your article to this journal [↗](#)



Article views: 1775



View related articles [↗](#)







View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

Dual context prior and refined prediction for semantic segmentation

Long Chen ^a, Jiajie Liu ^a, Han Li ^a, Wujing Zhan ^a, Baoding Zhou ^{b,c,d} and Qingquan Li ^{b,c}

^aSchool of Data and Computer Science, Sun Yat-sen University, Guangzhou, China; ^bGuangdong Key Laboratory of Urban Informatics, Shenzhen University, Shenzhen, China; ^cGuangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen, China; ^dCivil and Transportation Engineering, Shenzhen University, Shenzhen, China

ABSTRACT

Recently, the focus of semantic segmentation research has shifted to the aggregation of context prior and refined boundary. A typical network adopts context aggregation modules to extract rich semantic features. It also utilizes top-down connection and skips connections for refining boundary details. But it still remains disadvantage, an obvious fact is that the problem of false segmentation occurs as the object has very different textures. The fusion of weak semantic and low-level features leads to context prior degradation. To tackle the issue, we propose a simple yet effective network, which integrates dual context prior and spatial propagation—dubbed DSPNet. It extends two mainstreams of current segmentation researches: (1) Designing a dual context prior module, which pays attention to context prior again with a shortcut connection. (2) The network can inherently learn semantic aware affinity values for each pixel and refine the segmentation. We will present detailed comparisons, which perform on PASCAL VOC 2012 and Cityscapes. The result demonstrates the validation of our approach.

KEYWORDS

Deep learning; semantic segmentation; linear spatial propagation; context information

1. Introduction

Segmentation is a fundamental task among many computer vision tasks, such as scene parse (Chen et al. 2019), autonomous driving (Chen et al. 2020), objects detection (Chen et al. 2017a, 2017b), to name a few. Its mission is to assign each pixel with a category, which is crucial to subsequent task. As the thriving of Deep Convolution Neural Networks (DCNNs), particularly with the development of FCN (Long, Shelhamer, and Darrell 2015), many breakthroughs of semantic segmentation have been achieved based on many prior works. These improvements of segmentation should give credits to the adoption of taking advanced networks as feature extractor, such as ResNet (He et al. 2016), ResNeXt (Xie et al. 2017), XceptionNet (Chollet 2017). Dilated convolution is also a powerful tool since it can effectively enlarge receptive fields while remains high resolution feature map. It relieves the issue of intra-class inconsistent segmentation via extracting rich context information. Intra-class inconsistent segmentation means parts of the object (which belong to the same category) are falsely classified into other classes. Context information is so crucial for segmentation mainly due to it can highlight the co-occurrent visual patterns. Nevertheless, as a result of using large windows in both convolution and pooling operation, the segmentation of many prior researches may lack of local location information and precise boundary, like PSPNet (Zhao et al. 2017), Deeplabv3 (Chen et al. 2017e).

A very recent work, Deeplabv3+ (Chen et al. 2018) improves the segmentation through better reconstruction of location information. It performs deconvolution (Long, Shelhamer, and Darrell 2015) and bilinear interpolation over the coarse prediction. After that, low-level features are introduced for fusion process. Other similar works are also focusing on prediction refinement. However, though these decoding networks are effective to some extent, redundant boundary is introduced due to the absent of rich semantic awareness. It highlights the problem of intra-class inconsistent segmentation, which is emphasized by prior works (Zhao et al. 2017). As shown in Figure 1, parts of sheep and cow are falsely classified into cats and horses respectively.

With above discussion, we revalue how to possess both refined boundary and intra-class consistent segmentation. We bring in current neural network named UPerNet (Xiao et al. 2018) as our basic network. It includes top-down connection with inline context aggregation module followed by down-top and skip connections. As shown in Figure 2, each lateral branch gradually brings in features with object scales, local location, boundary details, which are crucial premise to the generation of dense prediction. However, these low-level features, which have disadvantages of weak semantic representation and redundant boundary details, result in being deficient in ability of learning the most distinctive features. For instance, a bus is likely to be classified as a car if the network responses too much to irrelevant features like windows or wheels. To this end, we introduce a graphical model-based method to inherently learn

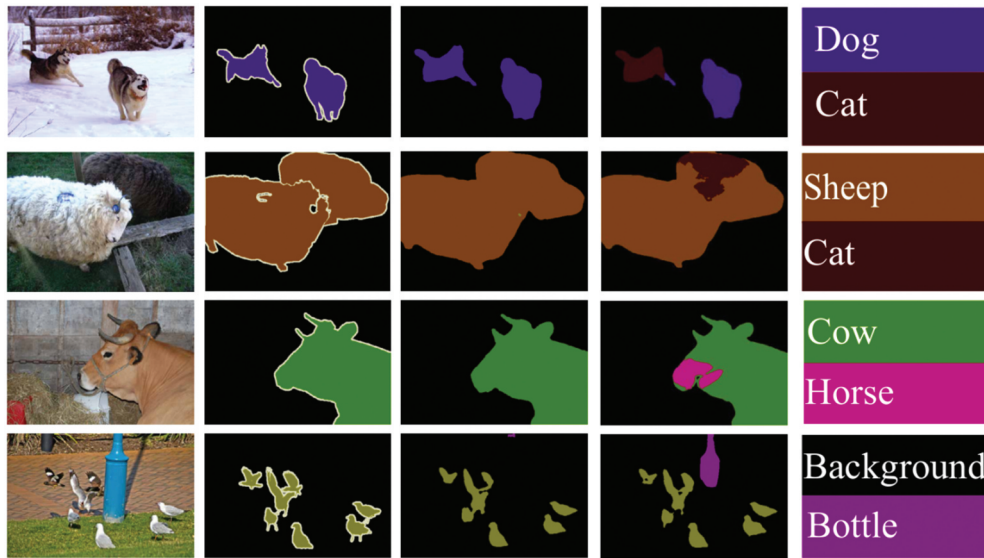


Figure 1. Visual results on PASCAL VOC 2012. From 1st to 4th column are images, ground truth, results from PSPNet and Deeplabv3+, respectively. Comparing to Deeplabv3+, PSPNet has better results that can suppress the intra-class inconsistent segmentation.

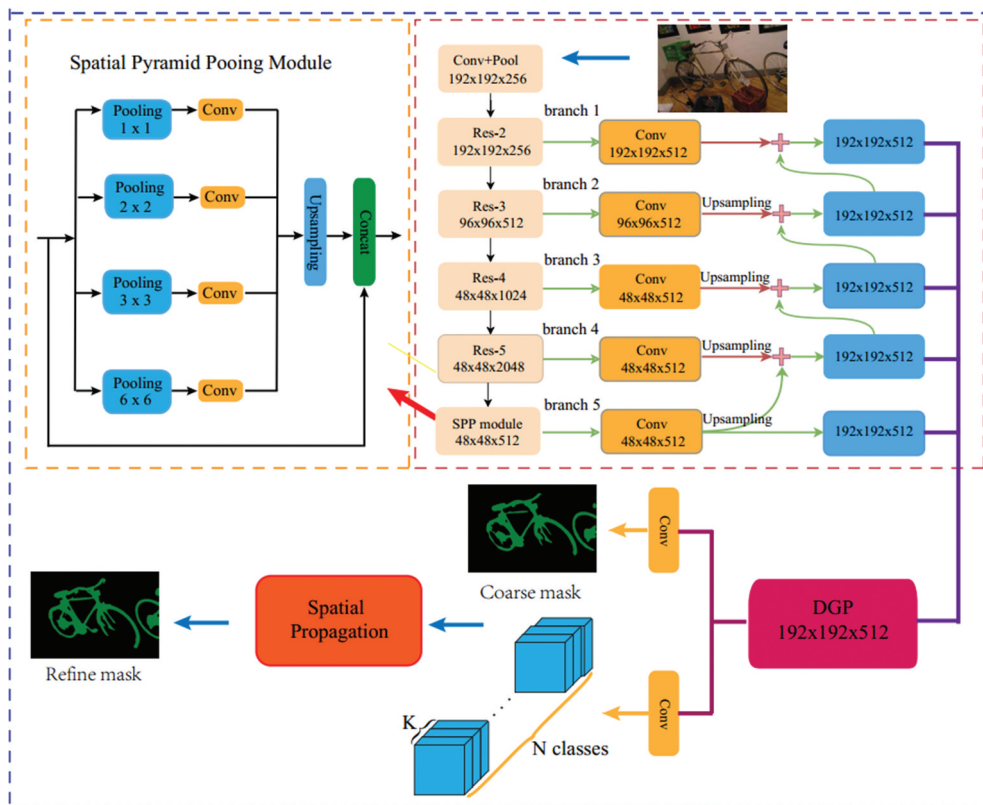


Figure 2. Overview of our proposed DSPNet.

semantic-aware global pairwise relationships of an image. A recent work, Liu et al. utilize an auxiliary network (Liu et al. 2017a) to learn semantic aware affinity values for high-level vision task and achieve promising result. We extend it to online training via a simple yet effective approach. After the processing of attention module, once more, the context information is introduced. Hence, the network obtains the capability of selectively combining category region and pixel segmentation to

suppress the problem of intra-class inconsistent segmentation. Meanwhile, with a bold yet reasonable assumption, lateral branches with features of object scales, pose, viewpoints can replace the aforementioned auxiliary network to learn affinity matrix with rich semantic-awareness. During training, our network refines the prediction via online linear propagation, which can enable the network to learn pairwise relationships in local-to-global feature space. Unlike SPN (Liu et al. 2017a)

performs spatial diffusion over the last hidden layer, we directly perform it over the prediction result. An intuition is that the inter-class diffusion should be reciprocal inhibition after the softmax layer.

In summary, there are three contributions in our paper:

- (1) We review the problem of intra-class inconsistent segmentation, which occurs in the procedure of refining prediction when employing down-top and skip connections.
- (2) We propose a simple yet effective architecture which incorporates Dual Context Prior (DCP) information module and refined prediction module. The DCP module can selectively combine category region and pixel segmentation to suppress the intra-class inconsistent segmentation.
- (3) We also develop an online spatial propagation network, which can perform local-to-global diffusion over prediction result by learning pairwise affinity value and yield precise segmentation.

2. Related work

Our work is built upon prior works of dilated convolution and context aggregation, prediction refinement, attention module and linear spatial propagation.

2.1. Dilated convolution and context aggregation

DCNNs achieve many astonishing accomplishments in the domain of image classification. Kai et al. propose residual convolution module along with a much deeper network (He et al. 2016). Benefiting from this, also for much dense segmentation, Deeplab (Chen et al. 2014) utilizes dilated convolution, which can effectively enlarge the receptive field while remains high resolution feature map. Further, to extract context information, GCN (Peng et al. 2017) constructs a large convolution kernel via a series of small ones and PSPNet (Zhao et al. 2017) employs parallel pooling module for context aggregation. Deeplabv3 (Chen et al. 2017e) adopts parallel dilated convolution with different rates. More recently, Zhang et al. propose Encnet (Zhang et al. 2018), which can extract context information much more effectively and set a new baseline on benchmarks.

2.2. Prediction refinement

Structures with top-down, down-top and skip connections are widely used among many computer vision tasks, like object detection (Ren et al. 2017), boundary detection (Xie and Tu 2015; Liu et al. 2017b; Xie and Tu 2015; Yang et al. 2016; Yu et al. 2017) and semantic segmentation (Long, Shelhamer, and Darrell 2015). To integrate different level features, FCN (Long,

Shelhamer, and Darrell 2015) adopts fully convolutional network while UNet (Ronneberger, Fischer, and Brox 2015) introduces the U-shape (Xie and Tu 2015; Peng et al. 2017; Ghiasi and Fowlkes 2016; Lin et al. 2016) structures with side connected. Kai et al. propose a variant of pyramid structure named FPN (Lin et al. 2017) to obtain more precise prediction.

2.3. Attention module

The attention module (Mnih et al. 2014; Wang et al. 2017; Chen et al. 2017c), which can make the model more responsive to what we need, becomes a powerful tool for deep neural networks (Chen et al. 2016; Hu, Shen, and Sun 2018; Zhang et al. 2018; Yu et al. 2018). The method (Chen et al. 2016) enables the network to pay attention to different scales information for semantic segmentation, PAD-Net (Xu et al. 2018) uses the attention module to control the features from other tasks into the target task. A very recent work, SE-Net (Hu, Shen, and Sun 2018) explores the cross-channel information to learn a channel-wise attention and achieves state-of-the-art performance in image classification task. In two lately researches on semantic segmentation, both EncNet (Zhang et al. 2018) and DFN (Yu et al. 2018) utilize attention module to obtain assumption factors, including scale attention factors and global attention factors.

2.4. Linear spatial propagation

The affinity matrices, which define pairwise relationships, are widely used for image filtering (Tomasi and Manduchi 1998) and image segmentation (Krahenbuhl and Koltun 2011). It improves performance among above tasks and propagating information over feature map. It also can retain the information of the edge for prediction refinement. For an effective learning strategy, Bertasius et al. (Bertasius, Shi, and Torresani 2016) take the level-features from DCNNs to extract the global pairwise relationships and takes a random walk network to share weights between nodes, which lead to high quality semantic segmentation. Since the computation of the random walk network is so expensive that the algorithm cannot converge stably, Liu et al. propose a spatial propagation network (Liu et al. 2017a) for learning the affinity matrix for visual tasks. By constructing a row/column linear propagation model, the spatial sales transformation matrix accurately and constitutes an affinity matrix, simulating the dense global pairwise relationship of the image. Taking the spatial propagation network as a post-processing strategy, Cheng et al. (Cheng et al. 2017) refine the coarse mask of instance-level object segmentation into a refined mask.

3. Proposed method

In this section, we will elaborate our methods, including Dual Context Prior (DCP) module and refined prediction module. Firstly, we will review the linear spatial propagation network and to extend it based on current architecture. Then, we will describe how to embed DCP information into prediction-end. Finally, we will give the overall architecture of our network.

3.1. Dual context prior (DCP)

As mentioned above, multi-scale context pooling module can effectively solve the problem of inconsistent segmentation, which may occur when an object has very different textures. As shown in Figure 2, in the decoding part with down-top and skip connections, the intra-class inconsistent segmentation occurs again. For this propose, we introduce the context prior attention again after the fusion of multiscale feature. The attention mechanism has been successfully employed in various tasks for filtering useful information. For instance, DFN (Yu et al. 2018), which extracts channel-wise attention factor, achieves state-of-the-art performance in semantic segmentation. Hence, we utilize attention mechanism to determine the context information whether introduce into the feature map after fusion of low-level and high-level features. In other words, in some case, the context

information may be redundant. The attention module can be regarded as a control gate to determine the usage of context information. This strategy allows the network inherently to pay or not to pay attention to context information. As shown Figure 3, our proposed DCP module learns attention factor G from multi-scale features,

$$G = \sigma(\text{conv}(F_s; w)) \quad (1)$$

where w denotes the weights of convolution kernel while σ denotes sigmoid activation function, and F_s denotes the fused features from both low- and high-level features. The final output of this module can be written as:

$$F_{dc} = F_s + G * \text{conv}(F_g; w) \quad (2)$$

where F_g is the multi-scale context information which from SPP module, as the left module shown in Figure 2. With respect to the role of DCP module, an intuition is that the network can adaptively select the useful information based on the reintroduced context prior information. For instance, if it is a bus, comparing to a car, the network should ignore the common features like windows or region, which with the same textures under illumination while remains the most distinctive features like boundary and color texture to ensure intra-class consistent segmentation. The inspiration for this work comes from (Xu et al. 2018), which uses the attention method to associate the features of other tasks.

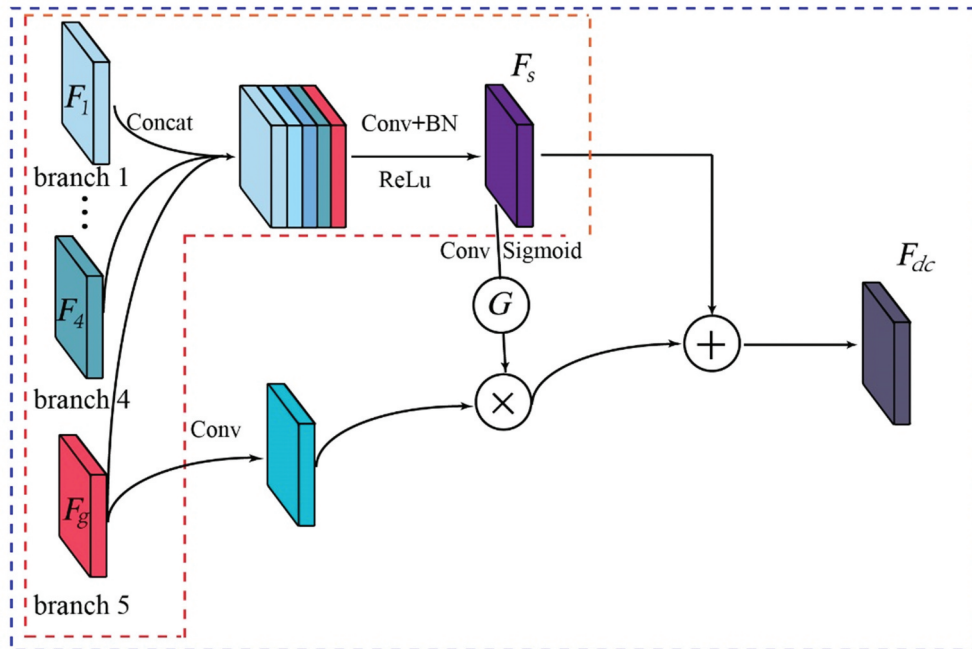


Figure 3. The components of the Dual Context Prior module (DCP). After the fusion between the high-level features from branch 5 (i.e. F_g) and the low-levels features (i.e. branch 1, 2, 3, 4), channel reduction operation (Conv+BN+ReLU) is performed on the fused features to make sure the dimension of channels from the fused features and F_g are equal. Then the Sigmoid operation is performed to the fused features F_s to generate an attention map G . The values G ranges from 0 to 1, which indicates the per-pixel importance of the original context feature F_g . Through the multiplication operation, the network can adaptively filter out the redundant context features from F_g via the attention map G . Furthermore, the addition between F_s and the filtered F_g can be deemed as residual learning.

3.2. Online linear spatial propagation

An intuitive understanding of SPN (Liu et al. 2017a) is to learn a semantic perceptual affinity value for each pixel pair through an auxiliary network (Long, Shelhamer, and Darrell 2015). In other words, it reveals the similarity of a pixel pair. Thus, for a pixel to be classified, a linear weighting operation is performed depending on the affinity value between the pixel and its adjacent pixels. Noted that the linear weighting operation is anisotropic which can retain the information of the edge. The computation of local to global diffusion is expensive. So, the SPN develops the linear spatial propagation direction with four directions, such as from left to right, and propagates once in one direction only associates with three adjacent pixels.

We modify this module to the point that it can be learned from online training. As shown in Figure 2, after the fusion of low-level and high-level features, we introduce features of object scales, poses and viewpoints to make sure local to global diffusion under limited propagation times. A feature map F_{dc} with size $m * n$, which is the output of DCP module serve as the input of the affinity matrix. Here, we have:

$$M^K = \text{conv}(F_{dc}; w) \quad (3)$$

Where M^K denotes the affinity matrix that need to be learned. Where K belongs to $N(i; j)$, which indicates the affinity values of a pixel with position of $(i; j)$ which response to its adjacent pixels, e.g. top-to-down, the column j is from $j - n + 1$ to j . It defines the similarity between pixels based on high-level vision features. Meanwhile, the network output a coarse segmentation mask based on the feature map F_{dc} , which is:

$$X = \text{conv}(F_{dc}; w) \quad (4)$$

We define as a propagation-hidden layer above feature map X , h_{ij} and are pixels with position $(i; j)$ for the hidden layer and the coarse prediction map, respectively. The 2D linear propagation from one direction can be described as:

$$h_{ij} = \left(1 - \sum_{K \in N(i,j)} M_{i,j}^K \right) x_{i,j} + \sum_{K \in N(i,j)} M_{i,j}^K h_k \quad (5)$$

where k is an adjacent pixel of $(i; j)$ in the hidden layer. Therefore, each direction of propagation ensures each pixel to obtain information from its adjacent pixels in a same direction. So, taking the node-wise max-pooling for merge four different directions, each pixel can obtain information from all over the prediction map. As mentioned by (Liu et al. 2017a), this diffusion operation will be stable under the condition as:

$$\sum_{K \in N(i,j)} |M_{i,j}^K| \leq 1 \quad (6)$$

The propagation in Eq. (5) is performed as column-wise transitions, which can be expressed by the following linear operation:

$$H_i = (1 - M_{i-1,i} X_i + M_{i-1,i} H_{i-1}) \quad (7)$$

Here, H_i , X_i denotes the i th column from linear propagation layer and coarse prediction map respectively. Where $h_0 = x_0$, and is a linear transition matrix. We define that this propagation repeating T times, (Liu et al. 2017a) proves that the two hidden matrices in adjacent states domain have:

$$H^T = -LH^{T-1} \quad (8)$$

Where L denotes a Laplacian matrix. Therefore, this linear propagation process is equivalent to spatial anisotropic diffusion process, which can smooth the non-boundary region and response to boundary details, which is the premise to high quality segmentation. As for the final result H^T , the softmax function will be employed to perform channel-wise quantized operation followed by cross-entropy loss function for final prediction.

$$P_{i,j,c} = \frac{\exp^{H_{i,j,c}^T}}{\sum_k \exp^{H_{i,j,c}^T}}$$

$$\text{loss} = - \sum_{i,j} \arg \max_c P \quad (9)$$

In the overall processing of refined prediction, except for using a serious of convolution operation for affinity matrix learning, all other layers directly perform on the coarse prediction and learned under Eq. (9). A motivation of these is that we consider the inter-class diffusion should be reciprocal inhibition.

3.3. Overall network architecture

Based on the DCP module and online linear spatial propagation module, we propose a simple yet effective deep semantic neural network (DSPNet), which integrates cascaded DCP attention module and prediction refinement module.

As shown in Figure 2, we adopt deep residual networks, which followed by a SPP module as the backbone of our propose network. The SPP module extracts multi-scale context information, which is rich in semantic-awareness. Skip connections are utilized to extract multi-scale feature of object scales, local location information. After channel dimension reduction, each lateral branch has a 512-dimensional feature map followed by bilinear interpolation up-sample operation to restore the resolution to a quarter of the input size. This strategy is inspired by two prior works, object detection network FPN (Lin et al. 2017) and scene understanding network UPerNet (Xiao et al. 2018). The output of aforesaid

modules will be concatenated to a feature map, which serves as the input of DCP attention module, which determines the fate of context information. Then, it outputs a coarse mask which contains robust intra-class consistent segmentation. Meanwhile, an affinity matrix, which contains the pixel affinity, learned via the information from the previous feature map. The affinity matrix performs spatial propagation over the coarse segmentation mask to further sharpen the boundary and smooth the intra-class region. Then, it outputs the final semantic segmentation prediction. Meanwhile, the network also supervises the coarse segmentation to enable the network for obtaining stable prediction result quickly and learns affinity values of high confidence.

4. Experimental results

To evaluate our proposed approach, experiments are conducted on the PASCAL VOC 2012 (Everingham et al. 2015) and the Cityscapes benchmark (Cordts et al. 2016). In this section, we firstly introduce the datasets and illustrate the implementation details. Thereafter, we evaluate each module of the proposed network by ablation study. Finally, we present the performance comparison with other state-of-the-art methods.

4.1. Datasets and metrics

4.1.1. PASCAL VOC 2012

The PASCAL VOC 2012 (Everingham et al. 2015) is a well known semantic segmentation dataset, which contains 20 object classes and one background, involving 1464 images for training, 1449 images for validation and 1456 images for testing. The original dataset is augmented by the Semantic Boundaries Dataset (Hariharan et al. 2011), resulting in 10,582 images for training.

4.1.2. Cityscapes datasets

The Cityscapes datasets consists of images collected from 50 different cities in Europe. 5000 images are with fine annotations, and 20,000 additional images are only with coarse annotations. These images are captured with urban street scenes, and the pixels are categorized into 19 testing classes.

4.1.3. Metrics

To evaluate the segmentation performance of our proposed architecture, we resort to the standard Jaccard Index, known as the mean intersection-over-union (mIOU) metric.

4.2. Network implementation and training

Our approach is based on the ResNeXt network (Xie et al. 2017). In regarding to ablation study, all parameters of batch normalization layers are fixed.

4.2.1. Training

We train the network using mini-batch stochastic gradient descent optimizer. The momentum is set to 0.9, and weight decay is set to 0.0001. Similar to (Chen et al. 2018; Zhao et al. 2017), we take patches with a size of 512 and 720 as input for PASCAL VOC and Cityscapes separately. We also use the “poly” learning rate policy where the learning rate is multiplied by $(1 - \frac{iter}{maxiter})^{0.9}$ and the initial learning rate is set to 0.007 and 0.0035 with or without SPN.

4.2.2. Data augmentation

We operate data augmentation as recommended in training process of (Zhao et al. 2017). Scale factor is sampled from the range (0.5, 2) and a rotation is from $(-10^\circ, 10^\circ)$.

4.3. PASCAL VOC 2012

In this section, we will discuss the influence of the DCP attention module on segmentation quality. Also, we will compare the segmentation result of DCP attention after incorporating different levels of low-level features. The result will prove that more low-level features lead to severe intra-class inconsistent segmentation. Finally, we will discuss the enhancement, which the online linear spatial propagation brings in semantic segmentation and influence of DCP attention module on the quality of linear spatial propagation.

4.3.1. Ablation study for dual context prior

We define decoder with stride of 4 and 2 as decoder A and B, respectively. Each lateral branch has a 512-dimensional channel feature map. On the issue of intra-class inconsistent segmentation during the procedure of gradually decoding, we adopt DCP module. The comparison result can be seen in Table 1. Firstly, in order to prove the importance of context prior to segmentation, the ResNeXt-101 with decoder B is similar to FCN-4 s (Long, Shelhamer, and Darrell 2015), which without context module has 71.8% mIOU. Furthermore, it achieves 74.9% mIOU with a growth of 3.1% after brings the SPP module for context prior, which is

Table 1. Ablation study for Dual Context Prior Attention. **decoder A:** A down-top network with stride 4. **decoder B:** A down-top network with stride 2. **SPP:** Multi-scale pooling module. **DCP:** Dual context prior attention module.

Model	mIOU
ResNeXt101 + decoder B	71.8
ResNeXt101 + SPP	74.9
ResNeXt101 + SPP + decoder A	76.1
ResNeXt101 + DCP + decoder A	77.3
ResNeXt101 + SPP + decoder B	76.5
ResNeXt101 + DCP + decoder B	77.9
ResNeXt152 + SPP + decoder B	80.5
ResNeXt152 + DCP + decoder B	81.2

same as PSPNet (Zhao et al. 2017). This result indicates that context aggregation is important to segmentation. After integrating decoder A, it has a growth of 1.2% mIOU and while with decoder B, which has more lateral branches has a growth of 1.6% mIOU, the PSPNet with the decoder B is from UPerNet (Xiao et al. 2018). The result indicates that decoder B is better than decoder A, which means that these low-level features are disturbing but useful. After introducing the dual context prior module with decoder A and B, it has a growth of 1.2 and 1.4% comparing to the SPP with decoder A and B, respectively. It also shows that network integrates DCP with decoder B has better result, which demonstrates our argument that integrates too much low-level features resulting in severe intra-class inconsistent segmentation, which is needed to be suppressed. When using the ResNeXt-152, which has deeper layers, the SPP with decoder B achieves 80.48%, the improvement in our methodology is still substantial, which has a growth of 0.7%. As the examples shown in Figure 4, The 1st and 2nd column are images and ground truth respectively. After introduce DCP (5th column), comparing to the SPP with decoder B (4th column) with integration of low-level features, it has a much smooth within intra-class region, which is close to PSPNet (3th column) but has stronger refined boundary.

4.3.2. Ablation study for refining prediction

As shown in Table 1, it yields better segmentation result when adopts DCP module before low-level features aggregation. To this end, to verify the validation

of linear spatial propagation network, we adopt the DCP module with decoder B as basic network. As shown in Table 2, the network with SPN has better performance, which has a growth of 1.6% in the ResNeXt-101 while a growth of 0.9% in the ResNeXt-152. Our DSPNet network performance grows with deeper networks, which mainly due to bias to the features of object scales, which also help learning the affinity values between pixels.

4.3.3. Ablation study for offline linear spatial propagation

In order to justify the impact of online or offline linear spatial propagation learning on segmentation, we perform a series of 32 channels convolution operation on the preliminary prediction to output a 32-dimensional feature map. Then, we use FCN-4 S (Long, Shelhamer, and Darrell 2015) as an auxiliary network to learn an affinity matrix. After the affinity matrix propagates over the 32-dimensional feature map, a series of 64 channels

Table 2. Ablation study for Refining with Linear Spatial Propagation. **xt**: The ResNext network. **SPN**: Online learning for Linear Spatial Propagation. **SPN↓**: Offline learning for linear spatial propagation.

Model	mIou
DCP-Xt101 + decoder B	77.9
DCP-Xt101 + decoder B + SPN	79.5
DCP-Xt101 + decoder B + SPN ↓	79.8
DCP-Xt152 + decoder B	81.2
DCP-Xt152 + decoder B + SPN	82.1
DCP-Xt152 + decoder B + SPN ↓	82.3



Figure 4. Visual improvements on PASCAL VOC 2012. From left to right are images, ground truth, predictions of PSPNet, UPerNet and DSPNet, respectively. It shows that DSPNet can both suppress the intra-class inconsistent segmentation and improve the quality of boundary.

Table 3. Per-class results on PASCAL VOC 2012 testing set. Methods pre-trained on MS-COCO are marked with “+”.

Method	CRF-RNN +	BoxSup +	Dilation8 +	DPN +	Piecewise +	FCRNs +	LRR +	DeepLabv2 +	PSPNet +	DeepLabv3 +	DeepLabv3+ +
Aero	90.4	89.8	91.7	89	94.1	92	92	93	95.8	96	97.5
Bike	55.3	38	39.6	61.6	40.7	48	45	60	72.7	77	77.9
Bird	88.7	89.2	87.8	87.7	84.1	93	95	92	95	93	96.2
Boat	68.4	68.9	63.1	66.8	67.8	69	65	63	78.9	78	80.4
Bottle	69.8	68	71.8	74.7	75.9	76	76	76	84.4	88	90.8
Bus	88.3	89.6	89.7	91.2	93.4	94	95	95	94.7	97	98.3
Car	82.4	83	82.9	84.3	84.3	88	89	88	92	90	95.5
Cat	85.1	87.7	89.8	87.6	88.4	93	92	93	95.7	95	97.6
Chair	32.6	34.4	37.2	36.5	42.5	37	39	33	43.1	48	58.8
Cow	78.5	83.6	84	86.3	86.4	87	86	89	91	93	96.1
Table	64.4	67.1	63	66.1	64.7	65	70	68	80.3	76	79.2
Dog	79.6	81.5	83.3	84.4	85.4	89	89	90	91.3	91	95
Horse	81.9	83.7	89	87.8	89	90	89	92	96.3	97	97.3
Mobike	86.4	85.2	83.8	85.6	85.8	87	89	87	92.3	91	94.1
Person	81.8	83.5	85.1	85.4	86	87	87	87	90.1	92	93.8
Plant	58.6	58.6	56.8	63.6	67.5	65	66	63	71.5	71	78.5
Sheep	82.4	84.9	87.6	87.3	90.2	90	86	88	94.4	91	95.5
Sofa	53.5	55.8	56	61.3	63.8	60	57	60	66.9	69	74.4
Train	77.4	81.2	80.2	79.4	80.9	86	86	87	88.8	91	93.8
Tv	70.1	70.7	64.7	66.4	73	73	77	75	82	79	81.6
mIOU	74.7	75.2	75.3	77.5	78	79	79	80	85.4	86	89

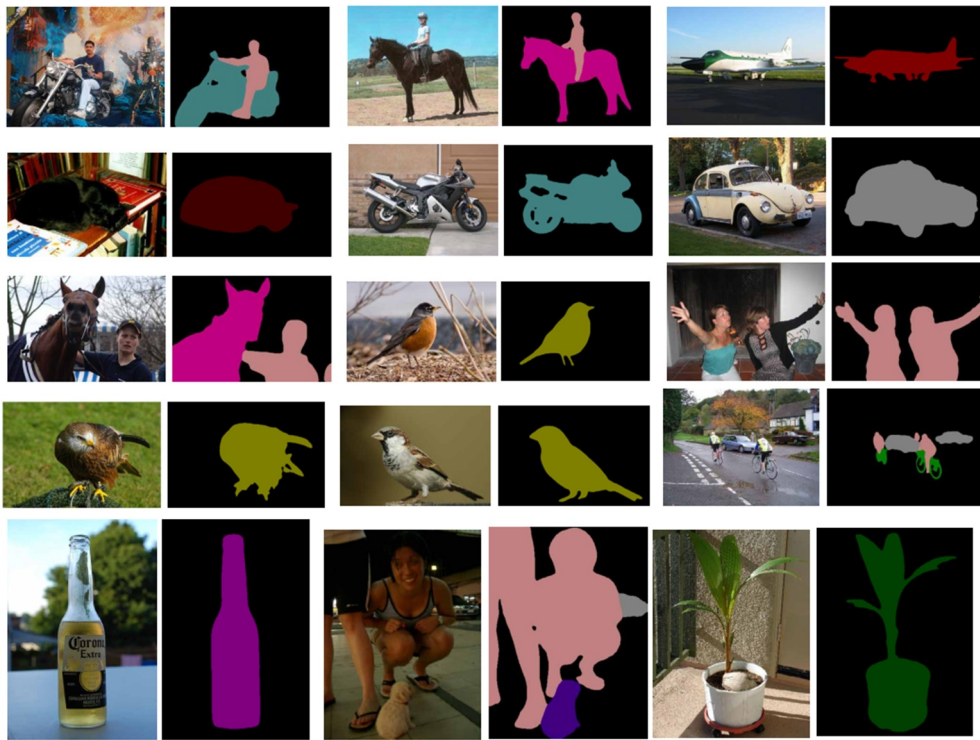


Figure 5. Visualization results on test set of PASCAL VOC 2012.

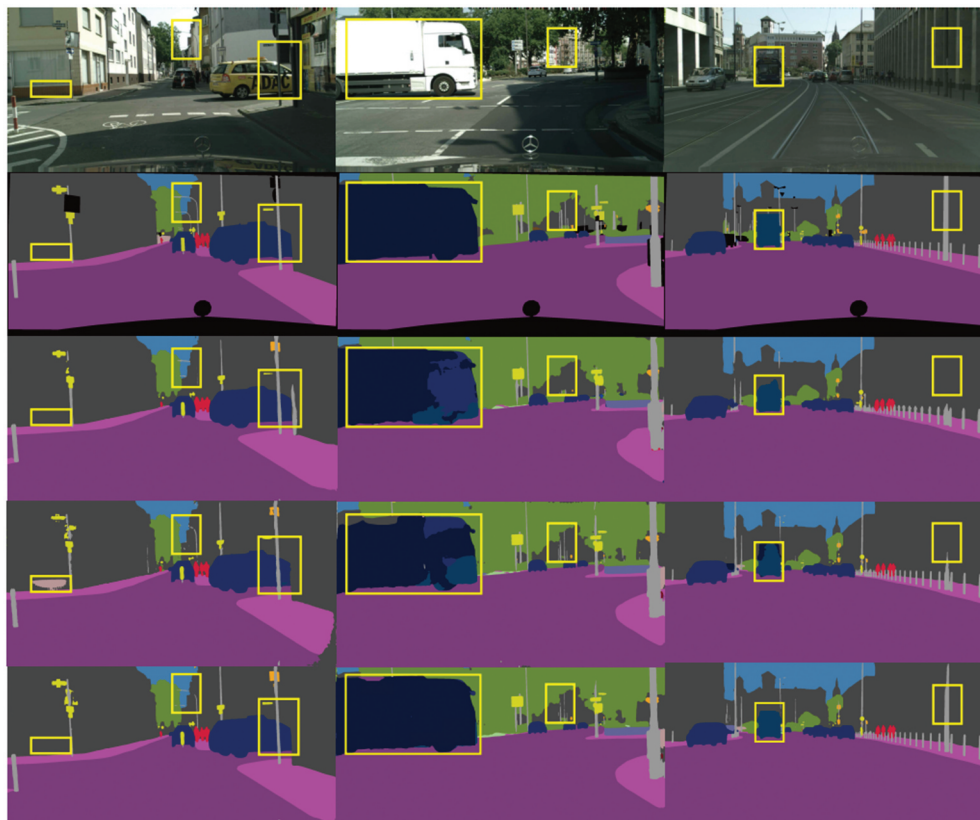


Figure 6. Visual improvements on Cityscapes datasets. From top to bottom are images, ground truth, predictions of PSPNet, UPerNet and DSPNet, respectively. In the attention frame (denoted with yellow box), it shows that the DSPNet can both suppress the intra-class segmentation inconsistency and improve the quality of boundary.

Table 4. Ablation study for the Cityscapes datasets.

Model	mIoU
ResNeXt152 + SPP + decoder B	80.2
DCP-xt152 + decoder B	80.9
DSPNet	81.2

Table 5. Results on Cityscapes testing set, iIoU is instance-level intersection-over-union metrics. The model in the top half of the Table only are trained with the fine data, and the models of the bottom half are trained with both fine and coarse data which are mark with ‡.

Model	IoU class	iIoU class	IoU category	iIoU category
CRFasRNN (Zheng et al. 2015)	62.5	34.4	82.7	66.0
FCN (Long, Shelhamer, and Darrell 2015)	65.3	41.7	85.7	70.1
DPN (Liu et al. 2015)	66.8	39.1	86.1	69.1
DeepLabv2 (Chen et al. 2017d)	70.4	42.6	86.4	67.7
PSPNet (Zhao et al. 2017)	78.4	56.7	90.6	78.6
DUC (Wang et al. 2018)	80.1	56.9	90.7	77.8
Ours	80.4	59.4	91.5	80.5
SegModel (Shen et al. 2017)‡	79.2	56.4	90.4	77.0
DFN (Yu et al. 2018)‡	80.3	58.3	90.8	79.6
ResNet-38 (Wu, Shen, and Hengel 2019)‡	80.6	57.8	91.0	79.1
PSPNet (Zhao et al. 2017)‡	81.2	59.6	91.2	79.2
DeepLabv3 (Chen et al. 2017e)	81.3	57.7	91.5	80.7
Ours‡	82.2	59.7	91.4	79.7

convolution operation are employed for the final refined prediction. This method comes from SPN (Liu et al. 2017a). As shown in Table 2, # denotes offline processing. It has similar results either online learning or offline learning. When going with deeper layers, the offline learning just has a growth of 0.2%, which is proved that the learning of semantic-aware affinity values can be shared from the same network. Though the offline SPN has a slightly higher performance than the online version, the disadvantages of the original offline SPN is that it takes much more time to train and needs extra memories to store the features. Thus, we modify the offline SPN so that it can train with the whole network, which can significantly reduce the demand of computation resource and memories.

4.3.4. PASCAL VOC 2012

In evaluation, we apply the multiscale scheme on inputs and also horizontally flip the inputs to further improve the performance. We further fine-tune our model on PASCAL VOC 2012 train and val set for evaluation on test set. More performance details are listed in Table 3, our model achieves 82.5%, which is competitive result. Example can be seen in Figure 5.

4.4. Cityscapes datasets

In previous sections, we elaborately discuss how the decoder module impacts the performance on SPP. For the process of progressive decoding, the problem of intra-class inconsistent segmentation shows up again, which also exists in the road scene dataset. Examples can be seen in Figure 6, the headstocks of the truck are falsely classified into bus due to similar textures between these two classes. Therefore, we introduce the DCP module and SPN module in the Cityscapes ablation study. We use the ResNeXt-152 (Xie et al. 2017) as base network. We take the patch with a size of 720 pixels as input. We also use fixed BN operation for training with the batch size smaller than 16. We apply the multi-scale inputs with scales range from (0.5, 2.0) and also horizontally flip images to further improve the performance. Without employing coarse data, when using the DCP module which can bring 0.6%, with the linear space propagate, our final model is evaluated on the Cityscapes val set and achieves an mIOU of 81.2%, as seen in Table 4. With the fine set which contains 3475 images, our final model is evaluated on the Cityscapes test set and achieves an mIOU of 80.4%, which outperforms state-of-the-art methods, as seen in Table 5. To compare with other state-of-the-art methods, we train our network with the train-val and coarse set, which has the extra 20,000 coarsely annotated images. We pretrained our model on the coarse data of Cityscapes, and then fine tune it on fine data. The final mIOU is 82.2%, which is still competitive to other approaches. Examples can be seen in Figure 7.

5. Conclusions

We review the most crucial problem, which is ignored by most of the researches: Context prior information does effectively solve the problem of intra-class inconsistent segmentation. But it starts to degrade when the network incorporates multi-scale features for refining prediction. For this purpose, this work proposes a simple yet effective network (DSPNet), which can pay attention to context information again by attention mechanism. Our network can perform robust intra-class consistent segmentation while inherently extract rich semantic affinity feature, which is utilized within the linear propagation network to sharpen boundary and smooth intra-class region for refined prediction. The results of experiment demonstrate the validation of our method.

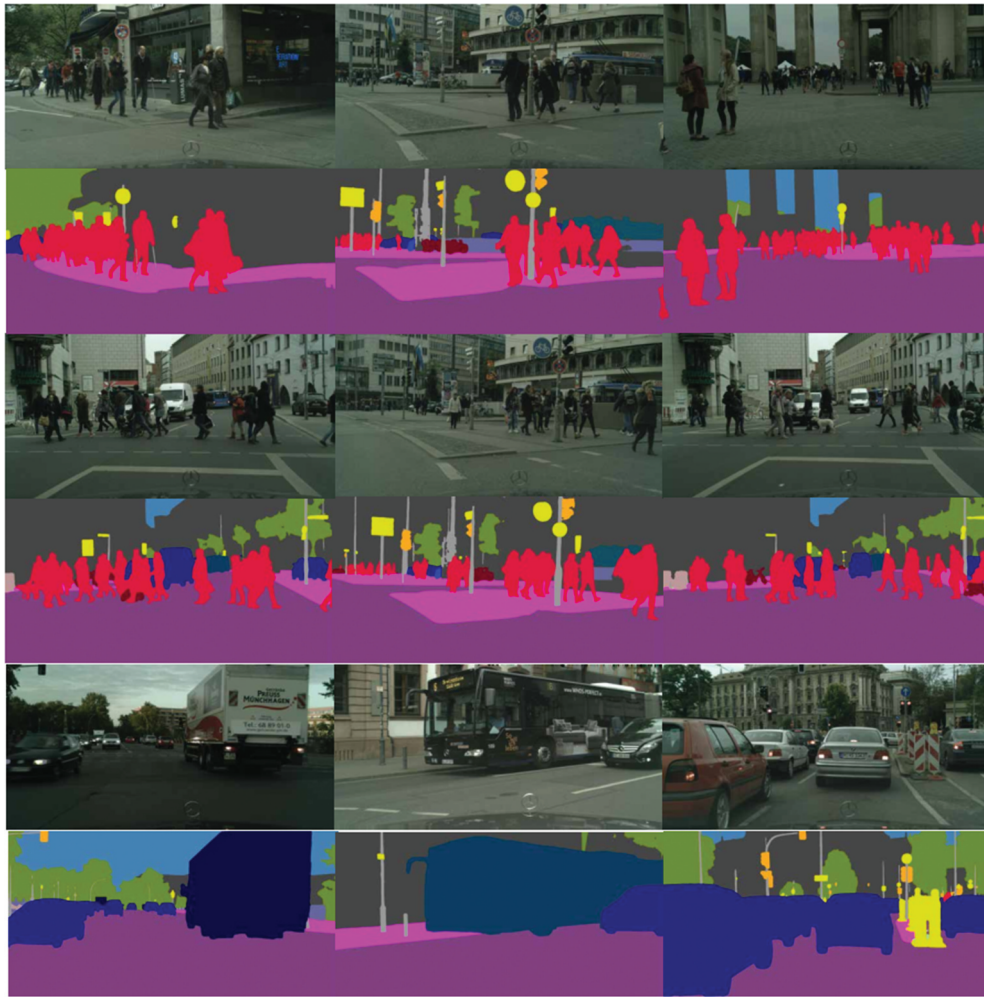


Figure 7. Visual results on test set of Cityscapes.

Notes on contributors

Long Chen received the BSc degree in communication engineering and the PhD degree in signal and information processing from Wuhan University in 2007 and in 2013, respectively. From October 2010 to November 2012, he was co-trained PhD Student at National University of Singapore. He is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China. He received the IEEE Vehicular Technology Society 2018 Best Land Transportation Paper Award, IEEE Intelligent Vehicle Symposium 2018 Best Student Paper Award. His areas of interest include autonomous driving, robotics, artificial intelligence where he has contributed more than 50 publications.

Jiajie Liu received BS in Civil Engineering from Guangzhou University in 2017. He is a postgraduate student and studies machine learning and deep learning in applications of both pattern recognition and computer vision currently. His areas of interest include Scene Flow Estimation, Semantic Segmentation.

Han Li received a BS in Software Engineering from Shenzhen University in 2018. He is currently pursuing his MS degree in software engineering at the School of Data Science and Computer Science, Sun Yat-Sen University. He is interested in autonomous driving and image processing.

Wujing Zhan is with the School of Data and Computer Science, Sun Yat-Sen University, Guangzhou. He is a master's degree student and studies machine learning and deep learning in applications of scene recognition, scene segmentation, and scene flow.




Baoding Zhou received the PhD degree in photogrammetry and remote sensing from Wuhan University in 2015. He is currently an Assistant Professor with the College of Civil and Transportation Engineering, Shenzhen University. His research interests include indoor localization and mapping, mobile computing, and intelligent transportation.

Qingquan Li received the PhD degree in geographic information system and photogrammetry from Wuhan Technical University of Surveying and Mapping in 1998. He is currently a Professor with Shenzhen University, and Wuhan University. His research areas include 3-D and dynamic data modeling in GIS, location-based service, surveying engineering, integration of GIS, global positioning system and remote sensing, intelligent transportation system, and road surface checking.

ORCID

Long Chen  <http://orcid.org/0000-0003-4925-0572>

Jiajie Liu  <http://orcid.org/0000-0002-7651-5540>

Han Li  <http://orcid.org/0000-0002-3575-2355>
 Wujing Zhan  <http://orcid.org/0000-0001-9683-7996>
 Baoding Zhou  <http://orcid.org/0000-0003-1607-2626>
 Qingquan Li  <http://orcid.org/0000-0002-2438-6046>

References

- Bertasius, G., J. Shi, and T. Torresani 2016. "Semantic Segmentation with Boundary Neural Fields." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26–July 1.
- Chen, L., H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua. 2017c. "Sca-cnn: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26.
- Chen, L., L. Fan, K. Huang, G. Xie, and A. Nuetcher. 2017a. "Moving-Object Detection from Consecutive Stereo Pairs Using Slanted Plane Smoothing." *IEEE Transactions on Intelligent Transportation Systems* 18 (11): 3093–3102. doi:10.1109/TITS.2017.2680538.
- Chen, L., Q. Wang, X. Lu, D. Cao, and F. Wang. 2020. "Learning Driving Models from Parallel End-to-End Driving Data Set." *Proceedings of the IEEE* 18 (12): 3303–3314. doi:10.1109/TITS.2017.2683641.
- Chen, L., W. Zhan, W. Tian, Y. He, and Q. Zou. 2019. "Deep Integration: Multi-label Architecture for Road Scene Recognition." *IEEE Transactions on Image Processing* 28 (10): 4883–4898. doi:10.1109/TIP.2019.2913079.
- Chen, L., X. Hu, T. Xu, H. Kuang, and Q. Li. 2017b. "Turn Signal Detection during Nighttime by CNN Detector and Perceptual Hashing Tracking." *IEEE Transactions on Intelligent Transportation Systems* 18 (12): 3303–3314. doi:10.1109/TITS.2017.2683641.
- Chen, L.-C., G. Papandreou, F. Schroff, and H. Adam. 2017e. *Rethinking Atrous Convolution for Semantic Image Segmentation*. In arXiv preprint arXiv:1706.05587. Ithaca, NY: Cornell University. <https://arxiv.org/abs/1706.05587>
- Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2014. *Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected Crfs*. arXiv preprint arXiv:1412.7062. Ithaca, NY: Cornell University. <https://arxiv.org/abs/1412.7062>
- Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2017d. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected Crfs." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (4): 834–848. doi:10.1109/TPAMI.2017.2699184.
- Chen, L.-C., Y. Yang, J. Wang, W. Xu, and A. L. Yuille. 2016. "Attention to Scale: Scale-aware Semantic Image Segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26–July 1.
- Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018. "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation." In Proceedings of the European conference on computer vision, Munich, Germany, September 8–14.
- Cheng, J., S. Liu, Y. Tsai, W. Hung, S. D. Mello, J. Gu, J. Kautz, S. Wang, and M. Yang. 2017. *Learning to Segment Instances in Videos with Spatial Propagation Network*. In arXiv preprint arXiv:1709.04609. Ithaca, NY: Cornell University. <https://arxiv.org/abs/1709.04609>
- Chollet, F. 2017. "Xception: Deep Learning with Depthwise Separable Convolutions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26.
- Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. 2016. "The Cityscapes Dataset for Semantic Urban Scene Understanding." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26 – July 1.
- Everingham, M., S. M. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2015. "The Pascal Visual Object Classes Challenge: A Retrospective." *International Journal of Computer Vision* 111 (1): 98–136. doi:10.1007/s11263-014-0733-5.
- Ghiasi, G., and C. C. Fowlkes 2016. "Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation." In European Conference on Computer Vision (ECCV), Amsterdam, Netherlands, October 8–16.
- Hariharan, B., P. Arbelaez, L. D. Bourdev, S. Maji, and J. Malik. 2011. "Semantic Contours from Inverse Detectors." In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Barcelona, Spain, November 8–11.
- He, K., X. Zhang, S. Ren, and J. Sun. 2016. "Deep Residual Learning for Image Recognition." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26–July 1.
- Hu, J., L. Shen, and G. Sun 2018. "Squeeze-and-excitation Networks." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, June 19–22.
- Krahenbuhl, P., and V. Koltun. 2011. "Efficient Inference in Fully Connected Crfs with Gaussian Edge Potentials." In Advances in Neural Information Processing Systems (NIPS), Sierra Nevada, Spain, December 16–17.
- Lin, G., C. Shen, A. V. Den Hengel, and I. D. Reid. 2016. "Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, June 26–July 1.
- Lin, T., P. Dollar, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. 2017. "Feature Pyramid Networks for Object Detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26.
- Liu, S., S. D. Mello, J. G. G. Zhong, M. Yang, and J. Kautz. 2017a. "Learning Affinity via Spatial Propagation Networks." In Advances in Neural Information Processing Systems (NIPS), Long Beach, California, December 3–9.
- Liu, Y., M. Cheng, X. Hu, K. Wang, and X. Bai. 2017b. "Rich Convolutional Features for Edge Detection." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26.
- Liu, Z., X. Li, P. Luo, C. C. Loy, and X. Tang. 2015. "Semantic Image Segmentation via Deep Parsing Network." In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Las Condes, Santiago, December 13–16.
- Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, June 8–12.

- Mnih, V., N. Heess, A. Graves, and K. Kavukcuoglu. 2014. "Recurrent Models of Visual Attention." In *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Quebec, December 8-11.
- Peng, C., X. Zhang, G. Yu, G. Luo, and J. Sun. 2017. "Large Kernel Matters — Improve Semantic Segmentation by Global Convolutional Network." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 21-26.
- Ren, S., K. He, R. B. Girshick, and J. Sun. 2017. "Faster R-cnn: Towards Real-time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (6): 1137-1149. doi:10.1109/TPAMI.2016.2577031.
- Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-net: Convolutional Networks for Biomedical Image Segmentation." In *International Conference on Medical Image Computing and Computer Assisted Intervention*, 234-241. Munich, Germany, October 5-9.
- Shen, F., R. Gan, S. Yan, and G. Zeng. 2017. "Semantic Segmentation via Structured Patch Prediction, Context Crf and Guidance Crf." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 21-26.
- Tomasi, C., and R. Manduchi. 1998. "Bilateral Filtering for Gray and Color Images." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Bombay, January 4-7.
- Wang, F., M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Tang. 2017. "Conditional Random Fields as Recurrent Neural Networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 21-26.
- Wang, P., P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. W. Cottrell. 2018. "Understanding Convolution for Semantic Segmentation." In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Lake Tahoe, March 12-15.
- Wu, Z., C. Shen, and A. V. D. Hengel. 2019. "Wider or Deeper: Revisiting the Resnet Model for Visual Recognition." *Pattern Recognition* 90: 119-133. doi:10.1016/j.patcog.2019.01.006.
- Xiao, T., Y. Liu, B. Zhou, Y. Jiang, and J. Sun. 2018. "Unified Perceptual Parsing for Scene Understanding." In *European Conference on Computer Vision (ECCV)*, Munich, Germany, September 8-14.
- Xie, S., R. B. Girshick, P. Dollar, Z. Tu, and K. He. 2017. "Aggregated Residual Transformations for Deep Neural Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 21-26.
- Xie, S., and Z. Tu. 2015. "Holistically-nested Edge Detection." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Las Condes, Santiago, Chile, December 13-16.
- Xu, D., W. Ouyang, X. Wang, and N. Sebe. 2018. "Pad-net: Multitasks Guided Prediction-and-distillation Network for Simultaneous Depth Estimation and Scene Parsing." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, June 19-22.
- Yang, J., B. L. Price, S. D. Cohen, H. Lee, and M. Yang. 2016. "Object Contour Detection with a Fully Convolutional Encoderdecoder Network." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, June 26-July 1.
- Yu, C., J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. 2018. "Learning a Discriminative Feature Network for Semantic Segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, June 19-22.
- Yu, Z., C. Feng, M. Liu, and S. Ramalingam. 2017. "Casenet: Deep Category-aware Semantic Edge Detection." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 21-26.
- Zhang, H., K. J. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. K. Agrawal. 2018. "Context Encoding for Semantic Segmentation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, June 19-22.
- Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia. 2017. "Pyramid Scene Parsing Network." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 21-26.
- Zheng, S., S. Jayasumana, B. Romeraparedes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. 2015. "Conditional Random Fields as Recurrent Neural Networks." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Las Condes, Santiago, Chile, December 13-16.