Learning a Discriminative Feature Network for Semantic Segmentation

Changqian Yu[1] Jingbo Wang[2] Chao Peng[3] Changxin Gao[1*] Gang Yu[3] Nong Sang[1]
[1]Key Laboratory of Ministry of Education for Image Processing and Intelligent Control,
School of Automation, Huazhong University of Science and Technology
[2]Key Laboratory of Machine Perception, Peking University
[3]Megvii Inc. (Face++)
{changqian_yu,cgao,nsang}@hust.edu.cn,wangjingbo1219@pku.edu.cn,{pengchao,yugang}@megvii.com

Abstract

Most existing methods of semantic segmentation still suffer from two aspects of challenges: intra-class inconsistency and inter-class indistinction. To tackle these two problems, we propose a Discriminative Feature Network (DFN), which contains two sub-networks: Smooth Network and Border Network. Specifically, to handle the intra-class inconsistency problem, we specially design a Smooth Network with Channel Attention Block and global average pooling to select the more discriminative features. Furthermore, we propose a Border Network to make the bilateral features of boundary distinguishable with deep semantic boundary supervision. Based on our proposed DFN, we achieve state-of-the-art performance ***86.2%*** *mean IOU on PASCAL VOC 2012 and* ***80.3%*** *mean IOU on Cityscapes dataset.*

1. Introduction

Semantic segmentation is a fundamental technique for numerous computer vision applications like scene understanding, human parsing and autonomous driving. With the recent development of the convolutional neural network, especially the Fully Convolutional Network (FCN) [27], a lot of great work such as [40, 6, 19, 30] have obtained promising results on the benchmarks. However, the features learned by these methods are usually not discriminative to differentiate 1) the patches which share the same semantic label but different appearances, named intra-class inconsistency as shown in the first row of Figure 1; 2) the two adjacent patches which have different semantic labels but with similar appearances, named inter-class indistinction as shown in the second row of Figure 1.

To address these two challenges, we rethink the semantic segmentation task from a more macroscopic point of view. In this way, we regard the semantic segmentation as a task to assign a consistent semantic label to a category of things, rather than to each single pixel. From a macroscopic perspective, regarding each category of pixels as a whole, inherently considers both intra-class consistency and inter-class variation. It means that the task demands discriminative features. To this end, we present a novel Discriminative Feature Network (DFN) to learn the feature representation which considers both the "intra-class consistency" and the "inter-class distinction".

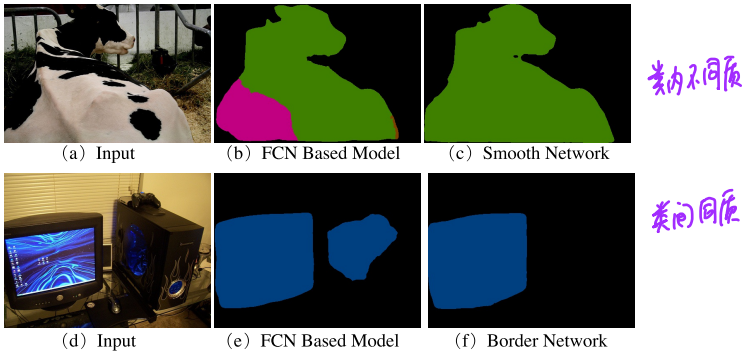


Figure 1. Hard examples in semantic segmentation. The second column is the output of FCN based model. The third column is the output of our proposed approach. In the first row, the left bottom corner of the cow is recognized as a horse. This is the **Intra-class Inconsistency** problem. In the second row, the computer case has the similar blue light and black shell with the computer screen, which is hard to distinguish. This is the **Inter-class Indistinction** problem.

Our DFN involves two components: Smooth Network and Border Network, as Figure 2 illustrates. The Smooth Network is designed to address the intra-class inconsistency issue. To learn a robust feature representation for intra-class consistency, we usually consider two crucial factors. On the one hand, we need multi-scale and global context features to encode the local and global information. For example, the small white patch only in Figure 1(a) usually

*Corresponding author.

cannot predict the correct category due to the lack of sufficient context information. On the other hand, as multi-scale context is introduced, for a certain scale of thing, the features have different extent of discrimination, some of which may predict a false label. Therefore, it is necessary to select the discriminative and effective features. Motivated by these two aspects, our Smooth Network is presented based on the U-shape [30, 19, 31, 11, 36] structure to capture the multi-scale context information, with the global average pooling [21, 24, 40, 6] to capture the global context. Also, we propose a Channel Attention Block (CAB), which utilizes the high-level features to guide the selection of low-level features stage-by-stage.

Border Network, on the other hand, tries to differentiate the adjacent patches with similar appearances but different semantic labels. Most of the existing approaches [24, 40, 6, 30] consider the semantic segmentation task as a dense recognition problem, which usually ignores explicitly modeling the inter-class relationship. Consider the example in Figure 1(d), if more and more global context is integrated into the classificiation process, the computer case next to the monitor can be easily misclassified as a monitor due to the similar appearance. Thus, it is significant to explicitly involve the semantic boundary to guide the learning of the features. It can amplify the variation of features on both sides. In our Border Network, we integrate semantic boundary loss during the training process to learn the discriminative features to enlarge the "inter-class distinction".

In summary, there are four contributions in our paper:

- We rethink the semantic segmentation task from a new macroscopic point of view. We regard the semantic segmentation as a task to assign a consistent semantic label to one category of things, not just at the pixel level.
- We propose a Discriminative Feature Network to simultaneously address the "intra-class consistency" and "inter-class variation" issues. Experiments on PASCAL VOC 2012 and Cityscapes datasets validate the effectiveness of our proposed algorithm.
- We present a Smooth Network to enhance the intra-class consistency with the global context and the Channel Attention Block.
- We design a bottom-up Border Network with deep supervision to enlarge the variation of features on both sides of the semantic boundary. This can also refine the semantic boundary of prediction.

2. Related Work

Recently, lots of approaches based on FCN have achieved high performance on different benchmarks [42, 9, 8]. Most of them are still constrained by intra-class inconsistency and inter-class indistinction issues.

Encoder-Decoder: The FCN model has inherently encoded different levels of feature. Naturally, some methods integrate them to refine the final prediction. This branch of methods mainly consider how to recover the reduced spatial information caused by consecutive pooling operator or convolution with stride. For example, SegNet [1] utilizes the saved pool indices to recover the reduced spatial information. U-net [31] uses the skip connection, while the Global Convolutional Network [30] adapts the large kernel size. Besides, LRR [11] adds the Laplacian Pyramid Reconstruction network, while RefineNet [19] utilizes multi-path refinement network. However, this type of architecture ignores the global context. In addition, most methods of this type are just summed up the features of adjacent stages without consideration of their diverse representation. This leads to some inconsistent results.

Global Context: Some modern methods have proven the effectiveness of global average pooling. ParseNet [24] firstly applies global average pooling in the semantic segmentation task. Then PSPNet [40] and Deeplab v3 [6] respectively extend it to the Spatial Pyramid Pooling [13] and Atrous Spatial Pyramid Pooling [5], resulting in great performance in different benchmarks. However, to take advantage of the pyramid pooling module sufficiently, these two methods adopt the base feature network to 8 times downsample with atrous convolution [5, 38], which is time-consuming and memory intensive.

Attention Module: Attention is helpful to focus on what we want. Recently, the attention module becomes increasingly a powerful tool for deep neural networks [28, 33, 16, 3]. The method in [7] pays attention to different scale information. In this work, we utilize channel attention to select the features similar to SENet [16].

Semantic Boundary Detection: Boundary detection is a fundamental challenge in computer vision. There are lots of specific methods proposed for the task of boundary detection [39, 36, 37, 25]. Most of these methods straightly concatenate the different level of features to extract the boundary. However, in this work, our goal is to obtain the features with inter-class distinction as much as possible with accurate boundary supervision. Therefore, we design a bottom-up structure to optimize the features on each stage.

3. Method

In this section, we first detailedly introduce our proposed Discriminative Feature Network containing Smooth Network and Border Network. Then, we elaborate how these two networks specifically handle the intra-class consistency issue and the inter-class distinction issue. Finally, we describe the complete encoder-decoder network architecture, Discriminative Feature Network.

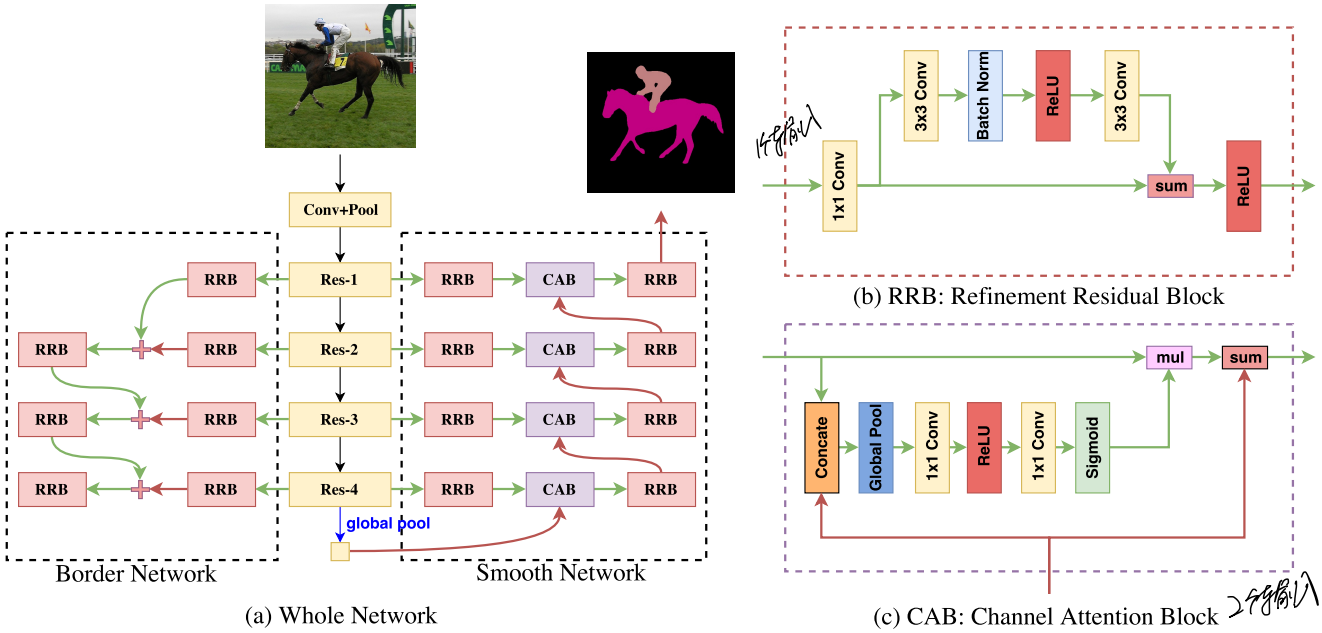


Figure 2. An overview of the Discriminative Feature Network. (a) Network Architecture. (b) Components of the Refinement Residual Block (RRB). (c) Components of the Channel Attention Block (CAB). The red and blue lines represent the upsample and downsample operators, respectively. The green line can not change the size of feature maps, just a path of information passing.

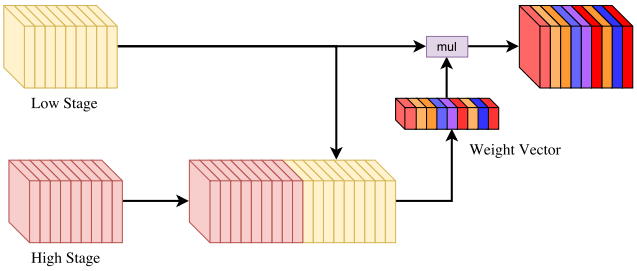
3.1. Smooth network

In the task of semantic segmentation, most of modern methods consider it as a dense prediction issue. However, the prediction sometimes has incorrect results in some parts, especially the parts of large regions and complex scenes, which is named intra-class inconsistency issue.

The intra-class inconsistency problem is mainly due to the lack of context. Therefore, we introduce the global context with global average pooling [24, 21, 40, 6]. However, global context just has the high semantic information, which is not helpful for recovering the spatial information. Consequently, we further need the multi-scale receptive view and context to refine the spatial information, as most modern approaches [40, 6, 30] do. Nevertheless, there exists a problem that the different scales of receptive views produce the features with different extents of discrimination, leading to inconsistent results. Therefore, we need to select more discriminative features to predict the unified semantic label of one certain category.

In our proposed network, we use ResNet [14] as a base recognition model. This model can be divided into five stages according to the size of the feature maps. According to our observation, the different stages have different recognition abilities resulting in diverse consistency manifestation. In the lower stage, the network encodes finer spatial information, however, it has poor semantic consistency because of its small receptive view and without the guidance of spatial context. While in the high stage, it has strong semantic consistency due to large receptive view, however, the prediction is spatially coarse. Overall, the lower stage makes more accurate spatial predictions, while the higher stage gives more accurate semantic predictions. Based on this observation, to combine their advantages, we propose a Smooth Network to utilize the high stage's consistency to guide the low stage for the optimal prediction.

We observe that in the current prevalent semantic segmentation architecture, there are mainly two styles. The first one is "Backbone-Style", such as PSPNet [40], Deeplab v3 [6]. It embeds different scale context information to improve the consistency of network with the Pyramid Spatial Pooling module [13] or Atrous Spatial Pyramid Pooling module [5]. The other one is "Encoder-Decoder-Style", like RefineNet [19], Global Convolutional Network [30]. This style of network utilizes the inherent multi-scale context of different stage, but it lacks the global context which has the strongest consistency. In addition, when the network combines the features of adjacent stages, it just sums up these features by channel. This operation ignores the diverse consistency in different stages. To remedy the defect, we first embed a global average pooling layer [24] to extend the U-shape architecture [27, 36] to a V-shape architecture. With the global average pooling layer, we introduce the strongest consistency constraint into the network as a guidance. Furthermore, to enhance consistency, we design a Channel Attention Block, as shown in Figure 2(c). This design combines the features of adjacent stages to compute a channel attention vector 3(b). The fea-



(a) Channel Attention Block

(b) Attention Vector

Figure 3. Schematic diagram of Channel Attention Block. In (a), the yellow block represents the feature of low stage, while the red one represents high stage. We concatenate the features of adjacent stages to compute a weight vector, which re-weights the feature maps of low stage. The hotter color represents the high weight value. In (b), it is the real attention value vector from the stage-4 channel attention block. The deeper blue represents the higher weight value.

tures of high stage provide a strong consistency guidance, while the features of low stage give the different discrimination information of features. In this way, the channel attention vector can select the discriminative features.

Channel attention block: Our Channel Attention Block (CAB) is designed to change the weights of the features on each stage to enhance the consistency, as illustrated in Figure 3. In the FCN architecture, the convolution operator outputs a score map, which gives the probability of each class at each pixel. In Equation 1, the final score at score map is just summed over all channels of feature maps.

$$y_k = F(x; w) = \sum_{i=1,j=1}^{D} w_{i,j} x_{i,j} \quad (1)$$

where x is the output feature of network. w represents the convolution kernel. And $k \in \{1, 2, \ldots, K\}$. K is the number of channels. D is the set of pixel positions.

$$\delta_i(y_k) = \frac{exp(y_k)}{\sum_{j=1}^{K} exp(y_j)} \quad (2)$$

where δ is the prediction probability. y is the output of network.

As shown in Equation 1 and Equation 2, the final predicted label is the category with highest probability. Therefore, we assume that the prediction result is y_0 of a certain patch, while its true label is y_1. Consequently, we can introduce a parameter α to change the highest probability value from y_0 to y_1, as Equation 3 shows.

$$\bar{y} = \alpha y = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_K \end{bmatrix} = \begin{bmatrix} \alpha_1 w_1 \\ \vdots \\ \alpha_K w_K \end{bmatrix} \times \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix} \quad (3)$$

where $\bar{y}$ is the new prediction of network and $\alpha = Sigmoid(x; w)$

Based on the above formulation of the Channel Attention Block (CAB), we can explore its practical significance. In Equation 1, it implicitly indicates that the weights of different channels are equal. However, as mentioned in Section 1, the features in different stages have different degrees of discrimination, which results in different consistency of prediction. In order to obtain the intra-class consistent prediction, we should extract the discriminative features and inhibit the indiscriminative features. Therefore, in Equation 3, the α value applies on the feature maps x, which represents the feature selection with CAB. With this design, we can make the network to obtain discriminative features stage-wise to make the prediction intra-class consistent.

Refinement residual block: The feature maps of each stage in feature network all go through the Refinement Residual Block, schematically depicted in Figure 2(b). The first component of the block is a 1×1 convolution layer. We use it to unify the number of channels to 512. Meanwhile, it can combine the information across all channels. Then the following is a basic residual block, which can refine the feature map. Furthermore, this block can strengthen the recognition ability of each stage, inspired from the architecture of ResNet [14, 15].

3.2. Border network

In the semantic segmentation task, the prediction is confused with the different categories with similar appearances, especially when they are adjacent spatially. Therefore, we need to amplify the distinction of features. With this motivation, we adopt a semantic boundary to guide the feature learning. To extract the accurate semantic boundary, we apply the explicit supervision of semantic boundary, which makes the network learn a feature with strong inter-class distinctive ability. Therefore, we propose a Border Network to enlarge the inter-class distinction of features. It directly learns a semantic boundary with an explicit semantic boundary supervision, similar to a semantic boundary detection task. This makes the features on both sides of semantic boundary distinguishable.

As stated in Section 3.1, the feature network has different stages. The low stage features have more detailed information, while the high stage features have higher semantic information. In our work, we need semantic boundary with more semantic meanings. Therefore, we design a bottom-up Border Network. This network can simultaneously get

accurate edge information from low stage and obtain semantic information from high stage, which eliminates some original edges lack of semantic information. In this way, the semantic information of high stage can refine the detailed edge information from low stage stage-wise. The supervisory signal of the network is obtained from the semantic segmentation's groundtruth with a traditional image processing method, such as Canny [2].

To remedy the imbalance of the positive and negative samples, we use focal loss [22] to supervise the output of the Border Network, as shown in Equation 4. We adjust the parameters α and γ of focal loss for better performance.

$$FL(p_k) = -(1-p_k)^{\gamma}\log p_k \tag{4}$$

where p_k is the estimated probability for class k, $k \in \{1, 2, \ldots, K\}$. And K is the maximum value of class label.

The Border Network mainly focuses on the semantic boundary which separates the classes on two sides of the boundary. For extracting accurate semantic boundary, the features on both sides will become more distinguishable. This exactly reaches our goal to make the features with inter-class distinction as much as possible.

3.3. Network Architecture

With Smooth Network and Border Network, we propose our Discriminative Feature Network for semantic segmentation as illustrated in Figure 2 (a).

We use pre-trained ResNet [14] as a base network. In the Smooth Network, we add the global average pooling layer on the top of the network to get the strongest consistency. Then we utilize the channel attention block to change the weights of channels to further enhance the consistency. Meanwhile, in the Border Network, with the explicit semantic boundary supervision, the network obtains accurate semantic boundary and makes the bilateral features more distinct. With the support of both sub-networks, the intra-class features become more consistent, while the inter-class ones grow more distinct.

For explicit feature refinement, we use deep supervision to get better performance and make the network easier to optimize. In the Smooth Network, we use the softmax loss to supervise the each stage's upsampled output excluding the global average pooling layer, while we use the focal loss to supervise the outputs of Border Network. Finally, we use a parameter λ to balance the segmentation loss ℓ_s and the boundary loss ℓ_b, as Equation 7 shows.

$$\ell_s = SoftmaxLoss(y; w) \tag{5}$$

$$\ell_b = FocalLoss(y; w) \tag{6}$$

$$L = \ell_s + \lambda\,\ell_b \tag{7}$$

4. Experimental Results

We evaluate our approach on two public datasets: PASCAL VOC 2012 [9] and Cityscapes [8]. We first introduce the datasets and report the implementation details. Then we evaluate each component of the proposed method, and analyze the results in detail. Finally, we present the comparison results with other state-of-the-art methods.

PASCAL VOC 2012: The PASCAL VOC 2012 is a well-known semantic segmentation benchmark which contains 20 object classes and one background, involving 1,464 images for training, 14,449 images for validation and 1,456 images for testing. The original dataset is augmented by the Semantic Boundaries Dataset [12], resulting in 10,582 images for training.

Cityscapes: The Cityscapes is a large semantic segmentation dataset of urban street scene in car perspective. The dataset contains 30 classes, of which 19 classes are considered for training and evaluation. There are 2,979 images for training, 500 images for validation and 1,525 images for testing, which are all fine annotated. And there are another 19,998 images with coarse annotation. The images all have a high resolution of 2,048×1,024.

4.1. Implementation details

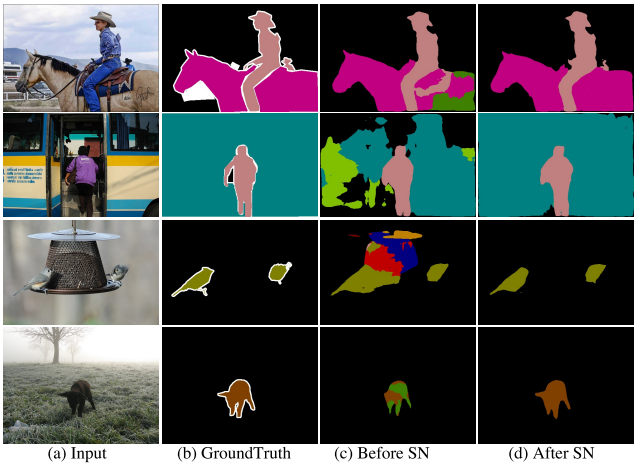
Our proposed network is based on the ResNet-101 pre-trained on ImageNet [32]. And we use the FCN4 [27, 36] as our base segmentation framework.

Training: We train the network using mini-batch stochastic gradient descent (SGD) [17] with batch size 32, momentum 0.9 and weight decay 0.0001. Inspired by [5, 24], we use the "poly" learning rate policy where the learning rate is multiplied by $\left(1 - \frac{iter}{max\_iter}\right)^{power}$ with power 0.9 and initial learning rate $4e^{-3}$. As for the λ, we finally use the value of 0.1 after a series of comparison experiments. For measuring the performance of our proposed network, we use the mean pixel intersection-over-union (mean IOU) as the metric.

Data augmentation: We use mean subtraction and random horizontal flip in training for both PASCAL VOC 2012 and Cityscapes. In addition, we find it is crucial to randomly scale the input images, which improves the performance obviously. We use 5 scales $\{0.5, 0.75, 1, 1.5, 1.75\}$ on both datasets.

4.2. Ablation study

In this subsection, we will step-wise decompose our approach to reveal the effect of each component. In the following experiments, we evaluate all comparisons on PASCAL VOC 2012 dataset [9]. And we report the comparison results in PASCAL VOC 2012 dataset [9] and Cityscapes dataset [8].



(a) Input (b) GroundTruth (c) Before SN (d) After SN

Figure 4. Results of Smooth Network on PASCAL VOC 2012 dataset.

Table 1. The performance of ResNet-101 with and without random scale augmentation.

Method	Random_Scale	Mean IOU(%)
Res-101		69.26
Res-101	√	72.86

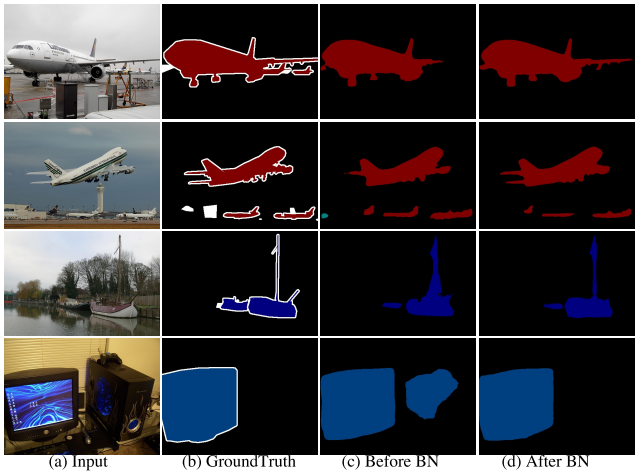
4.2.1 Smooth network

We use the ResNet-101 as our base feature network, and directly upsample the ouput. First, we evaluate the performance of the base ResNet-101, as shown in Table 1. Then we extend the base network to FCN4 structure [27, 36] with our proposed Refinement Residual Block (RRB), which improves the performance from 72.86% to 76.65%, as Table 2 shows. We visualize the effect of the Smooth Network. Figure 4 presents some examples of semantic segmentation results. Obviously, our Smooth Network can effectively make the prediction more consistent.

Ablation for global pooling: We need the features with strong consistency. Thus based our observation in Section 3, we add the global average pooling on the top of the network. As shown in Table 2, the global average pooling introduces the strongest consistency to guide other stages. This improves the performance from 76.65% to 78.20%, which is an obvious improvement.

Ablation for deep supervision: To refine the hierarchical features, we use deep supervision. We add the softmax loss on each stage excluding the global average pooling layer. As shown in Table 2, this further improves the performance by almost 0.4%.

Ablation for channel attention block: Based on the aforementioned architecture, we add the Channel Attention Block (CAB). It utilizes the high stage to guide the low stage with a channel attention vector to enhance consistency, which improves the performance from 78.51% to 79.54% over evaluation, as Table 2 shows.



(a) Input (b) GroundTruth (c) Before BN (d) After BN

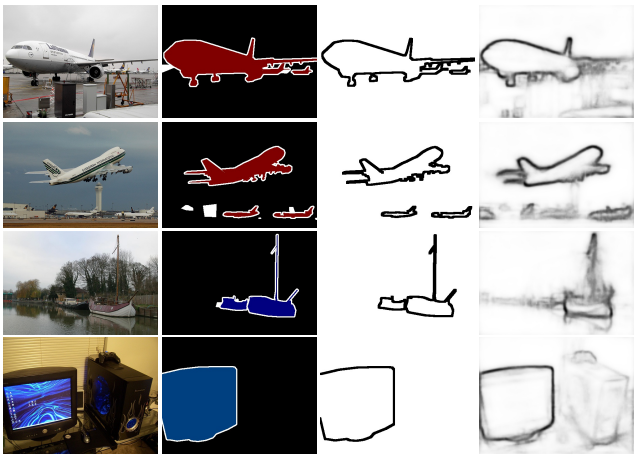
Figure 5. Results of Border Network on PASCAL VOC 2012 dataset. The boundary on prediction is refined by the Border Network.

Table 2. Detailed performance comparison of our proposed Smooth Network. **RRB:** refinement residual block. **GP:** global pooling branch. **CAB:** channel attention block. **DS:** deep supervision.

Method	Mean IOU(%)
Res-101	72.86
Res-101+RRB	76.65
Res-101+RRB+GP	78.20
Res-101+RRB+GP+CAB	79.31
Res-101+RRB+DS	77.08
Res-101+RRB+GP+DS	78.51
Res-101+RRB+GP+CAB+DS	79.54

4.2.2 Border network

While the Smooth Network pays attention to the intra-class consistency, the Border Network focuses on the inter-class indistinction. Due to the accurate boundary supervisory signal, the network amplifies the distinction of bilateral feature to extract the semantic boundary. Then we integrate the Border Network into the Smooth Network. This improves the performance from 79.54% to 79.67%, as shown in Table 3. The Border Network optimizes the semantic boundary, which is a comparably small part of the whole image, so this design makes a minor improvement. We visualize the effect of Border Network, as shown in Figure 5. In addition, Figure 6 shows the predicted semantic boundary of Border Network. We can obviously observe that the Border Network can focus on the semantic boundary preferably.



(a) Input (b) GroundTruth (c) Boundary (d) Prediction

Figure 6. The boundary prediction of Border Network on PASCAL VOC 2012 dataset. The third column is the semantic boundary extracted from GroundTruth by Canny operator. The last column is the prediction results of Border Network.

Table 3. Combining the Border Network and Smooth Network as Discriminative Feature Network. **SN**: Smooth Network. **BN:** Border Network. **MS_Flip:** Adding multi-scale inputs and left-right flipped inputs.

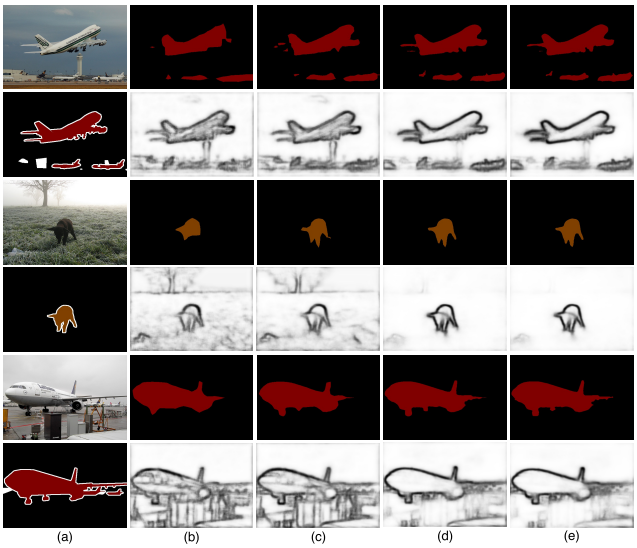
Method	Mean IOU(%)
Res-101+SN	79.54
Res-101+SN+BN	79.67
Res-101+SN+MS_Flip	79.90
Res-101+SN+BN+MS_Flip	80.01

4.2.3 Discriminative Feature network

With the Discriminative Feature Network (DFN), we conduct experiments about the balance parameter of the combined loss. Then we present the final results on PASCAL VOC 2012 and Cityscapes datasets.

Balance of both losses: The balance weight between the losses of two networks is crucial. To further analyze the effect of these two networks, we conduct experiments for different balance value. We test five values of $\{0.05, 0.1, 0.5, 0.75, 1\}$. As shown in Figure 8, with the same setting, our method achieves the highest performance with the value of 0.1.

Stage-wise refinement: It is worth noting that both Smooth Network and Border Network use the stage-wise mechanism. The Smooth Network utilizes a top-down stage-wise manner to transmit the context information from high stage to low stage, to ensure the inter-class consistency. On the other hand, the Border Network uses a bottom-up stage-wise manner to refine the semantic boundary with the edge information in the lower stage. With the bidirectional



(a) (b) (c) (d) (e)

Figure 7. Example results of DFN in the stage-wise refinement process on PASCAL VOC 2012 dataset. The first column is the original image and groundtruth. The last is the refinement process of two networks. The segmentation prediction in lower stage is more spatial coarse, and the higher is finer. While the boundary prediction in lower stage contains more edges not belong to semantic boundary, the semantic boundary in higher stage is more pure.

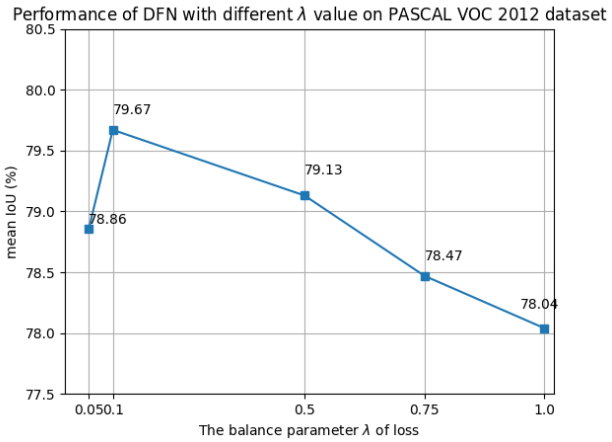


Figure 8. Results of DFN with different λ value on PASCAL VOC 2012 dataset.

stage-wise mechanism, the Smooth Network and Border Network respectively refine the segmentation and boundary prediction, as shown in Figure 7. The gradually accurate predictions validate the effectiveness of the stage-wise mechanism.

Performance evaluation on PASCAL VOC 2012: In evaluation, we apply the multi-scale inputs (with scales $\{0.5, 0.75, 1.0, 1.5, 1.75\}$) and also horizontally flip the in-

Table 4. Validation strategy on PASCAL VOC 2012 dataset. **MS_Flip:** Multi-scale and flip evaluation.

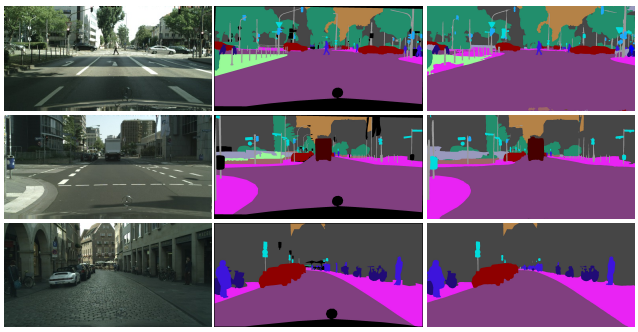
Method	train_data	MS_Flip	Mean IOU(%)
DFN			79.67
DFN	√		80.46
DFN	√	√	80.60

Table 5. Performance on PASCAL VOC 2012 test set. Methods pre-trained on MS-COCO are marked with $^+$.

Method	Mean IOU(%)
FCN [27]	62.2
Zoom-out [29]	69.6
ParseNet [24]	69.8
Deeplab v2-CRF [5]	71.6
DPN [26]	74.1
Piecewise [20]	75.3
LRR-CRF [11]	75.9
PSPNet [40]	82.6
Ours	**82.7**
DLC$^+$ [18]	82.7
DUC$^+$ [34]	83.1
GCN$^+$ [30]	83.6
RefineNet$^+$ [19]	84.2
ResNet-38$^+$ [35]	84.9
PSPNet$^+$ [40]	85.4
Deeplab v3$^+$ [6]	85.7
Ours$^+$	**86.2**

puts to further improve the performance. In addition, since the PASCAL VOC 2012 dataset provides higher quality of annotation than the augmented datasets [12], we further fine-tune our model on PASCAL VOC 2012 *train* set for evaluation on validation set. More performance details are listed in Table 4. And then for evaluation on test set, we use the PASCAL VOC 2012 *trainval* set to further fine-tune our proposed method. In the end, our proposed approach respectively achieves performance of 82.7% and 86.2% with and without MS-COCO [23] fine-tuning, as shown in Table 5. Note that, we do not use Dense-CRF [4] post-processing for our method.

Performance evaluation on Cityscapes: We also evaluate our approach on the Cityscapes dataset [8]. In training, our crop size of image is 800×800. We observe that for the high resolution of image the large crop size is useful. The test performance results are specifically reported in Table 6. We visualize the results of our approach on the Cityscapes dataset, as shown in Figure 9.



(a) Input (b) GroundTruth (c) DFN

Figure 9. Example results of DFN on Cityscapes dataset.

Table 6. Performance on Cityscapes test set. The "-" indicates that the method do not present this result in its paper.

Method	Mean IOU(%)	
	w/o coarse	w/ coarse
CRF-RNN [41]	62.5	-
FCN [27]	65.3	-
DPN [26]	66.8	59.1
LRR [11]	69.7	71.8
Deeplab v2-CRF [5]	70.4	-
Piecewise [20]	71.6	-
RefineNet [19]	73.6	-
SegModel [10]	78.5	79.2
DUC [34]	77.6	80.1
PSPNet [40]	78.4	80.2
Ours	**79.3**	**80.3**

5. Conclusion

We redefine the semantic segmentation from a macroscopic view of point, regarding it as a task to assign a consistent semantic label to one category of objects, rather than to each single pixel. Inherently, this task requires the intra-class consistency and inter-class distinction. Aiming to consider both sides, we propose a Discriminative Feature Network, which contains two sub-networks: Smooth Network and Border Network. With the bidirectional stage-wise mechanism, our approach can capture the discriminative features for semantic segmentation. Our experimental results show that the proposed approach can significantly improve the performance on the PASCAL VOC 2012 and Cityscapes benchmarks.

Acknowledgment

This work has been supported by the Project of the National Natural Science Foundation of China No.61433007 and No.61401170.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. 2

[2] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, June 1986. 5

[3] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *International Conference on Learning Representations*, 2015. 8

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv*, 2016. 2, 3, 5, 8

[6] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv*, 2017. 1, 2, 3, 8

[7] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2, 5, 6, 8

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html. 2, 5, 6

[10] S. Y. Falong Shen, Gan Rui and G. Zeng. Semantic segmentation via structured patch prediction, context crf and guidance crf. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 8

[11] G. Ghiasi and C. C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision*, 2016. 2, 8

[12] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik. Semantic contours from inverse detectors. In *IEEE International Conference on Computer Vision*. IEEE, 2011. 5, 8

[13] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, 2014. 2, 3

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 4, 5

[15] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, 2016. 4

[16] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv*, 2017. 2

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, 2012. 5

[18] X. Li, Z. Liu, P. Luo, C. C. Loy, and X. Tang. Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 8

[19] G. Lin, A. Milan, C. Shen, and I. Reid. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 8

[20] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 8

[21] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations*, 2014. 2, 3

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2017. 5

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 2014. 8

[24] W. Liu, A. Rabinovich, and A. C. Berg. Parsenet: Looking wider to see better. In *International Conference on Learning Representations*, 2016. 2, 3, 5, 8

[25] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai. Richer convolutional features for edge detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[26] Z. Liu, X. Li, P. Luo, C.-C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *IEEE International Conference on Computer Vision*, 2015. 8

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 3, 5, 6, 8

[28] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Neural Information Processing Systems*, 2014. 2

[29] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 8

[30] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters–improve semantic segmentation by global convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 8

[31] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015. 2

[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual

Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 5

[33] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[34] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 8

[35] Z. Wu, C. Shen, and A. v. d. Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv*, 2016. 8

[36] S. Xie and Z. Tu. Holistically-nested edge detection. In *IEEE International Conference on Computer Vision*, 2015. 2, 3, 5, 6

[37] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang. Object contour detection with a fully convolutional encoder-decoder network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[38] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016. 2

[39] Z. Yu, C. Feng, M.-Y. Liu, and S. Ramalingam. Casenet: Deep category-aware semantic edge detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[40] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 8

[41] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr. Conditional random fields as recurrent neural networks. In *IEEE International Conference on Computer Vision*, 2015. 8

[42] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2