Learning Adaptive Receptive Fields for Deep Image Parsing Network

Zhen Wei[1,4], Yao Sun[1*], Jinqiao Wang[2], Hanjiang Lai[3], Si Liu[1]
1. State Key Laboratory of Information Security (SKLOIS),
Institute of Information Engineering, Chinese Academy of Sciences, Beijing, 100093, China
2. National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
3. School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510275, China
4. University of Chinese Academy of Sciences, Beijing, 101408, China
{weizhen,sunyao,liusi}@iie.ac.cn, jqwang@nlpr.ia.ac.cn, laihanj@gmail.com

Abstract

In this paper, we introduce a novel approach to regulate receptive field in deep image parsing network automatically. Unlike previous works which have stressed much importance on obtaining better receptive fields using manually selected dilated convolutional kernels, our approach uses two affine transformation layers in the network's backbone and operates on feature maps. Feature maps will be inflated/shrinked by the new layer and therefore receptive fields in following layers are changed accordingly. By end-to-end training, the whole framework is data-driven without laborious manual intervention. The proposed method is generic across dataset and different tasks. We conduct extensive experiments on both general image parsing task and face parsing task as concrete examples to demonstrate the method's superior regulation ability over manual designs.

1. Introduction

In deep neural network, the notion of receptive field refers to the extent of data that are path-connected to a neuron [12]. After the introduction of Fully Convolutional Network (FCN) [11], receptive field has become especially important for deep image parsing network and could significantly affect the network's performance. As discussed in [14], a small receptive field may lead to inconsistent parsing results on large objects while a large receptive field often ignores small objects and classify them as background. Even not to such extreme extents, unsuitable receptive fields can also impair performance.

Recent works such as [2, 19] have already accentuated on adapting network's structures to realizing different receptive fields. Dilated convolutional kernels are often used to achieve this kind of adaptation. By setting different dilation values (mostly integers), the convolutional kernels could expand its receptive field accordingly. However, there are several main drawbacks in this approach that should be addressed. Firstly, these dilation values are always treated as hyper-parameters in network design. The selection of dilation values is based on designers' observation or results of a series of trials on a certain dataset, which is laborious and time-consuming. Secondly, such selection results are not generic across different image parsing tasks or even various dataset under the same task. During network transfer, such selection procedure would be repeated again. Thirdly, dilated convolutional kernels only produce discrete values of receptive fields. When dilation value is added by 1, the receptive field (e.g. the fc6 layer in VGG [15]) may expand by tens or even hundreds of pixels, making it even harder to find a finer receptive field.

The contribution of this paper is to propose a learning based, data-driven method for regulating receptive field in deep image parsing network automatically. The main idea is to introduce a novel affine transformation layer (the *'inflation layer'*) before the convolutional layer whose receptive field needs to be regulated. This inflation layer uses derivable interpolation algorithms to enlarge or shrink feature maps. The following layers perform inference on these inflated features and thus receptive fields after the inflation layer are changed. Then, inference results (before SoftMax normalization) will be resized to a fixed size by *'interpolation layer'*. During training, the *'inflation factor'* (denoted as f) embedded in both inflation layer and interpolation layer is derivable and is trained end-to-end together with the network backbone. As f may be a float number, the inflation layer is able to produce a more fine-grained receptive field and is only trained once. To corroborate the method's effectiveness, we conduct experiments on both general im-

*corresponding author

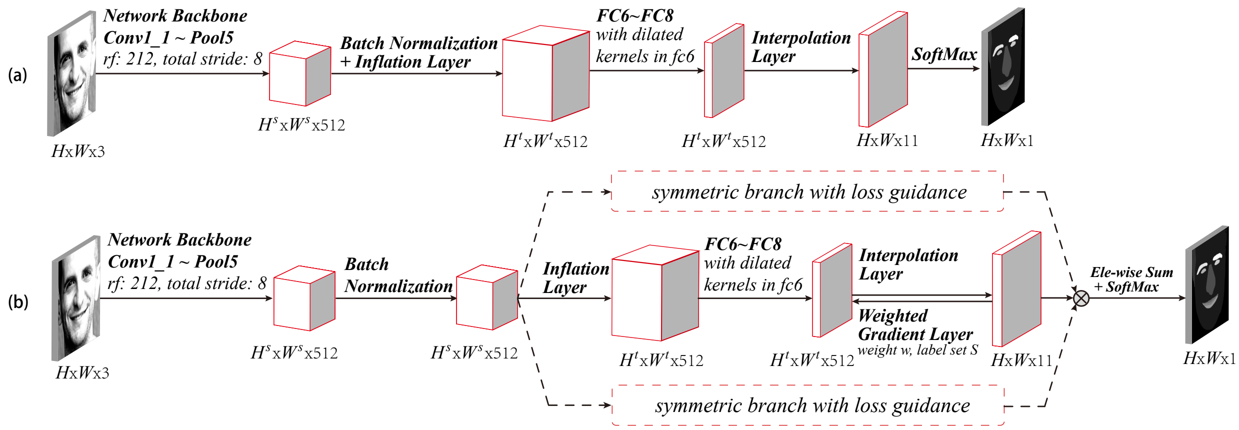

Figure 1. The framework of our method. (a): modified single path network. New layers are inserted before fc6 layer and after fc8 layer. (b): modified multi-paths network where all branches are with the same structure and initialization. *Weighted Gradient Layers* are used to break the symmetry during training. The specific settings of the single-path network are shown in Table 1 in supplementary file.

age parsing task as well as face parsing task. With proper initialization settings, the proposed method could achieve compatible, or even superior performance comparing to the best manually selected dilated convolutions. Additionally, due to the strong regulation ability brought by our method, the improved model achieves state-of-the-art face parsing accuracy on Helen dataset [8, 16].

The rest of this paper is organized as follows. In Section 2, we will review related works on image parsing tasks, especially focusing on issues related to receptive field. The Section 3 will further elaborate on implementation details of the new affine transformation layer and the derivatives of the inflation factor f. Section 4 will describe experimental settings. In Section 5, experimental results are further discussed. And Section 6 concludes the paper.

2. Related Work

A brief review on related works and contrastive discussions are made in this section.

2.1. FCNs and Dilated Convolutions

The introduction of FCN [11] has placed the receptive field in a prominent position. The forward process of FCN to generate dense classification result is equal to a series of inference using sliding windows on input image. With the sliding stride fixed, inference at pixel-level is solely based on data inside the window. The size of window is exactly the receptive field of the network. In [11], the authors discuss on dilated convolutions but do not use it in network. Then DeepLab [2] uses dilated convolutions to reduce pooling strides while expanding receptive field and reducing parameters in fc6 layer. [19] appends a series of dilated convolutional layers after a FCN backbone (or *the 'frontend'*) to expand receptive field. Recently, in DeepLab v2 [3], the authors manually design four different dilated convolutions in parallel to achieve multi-scale parsing.

However, dilation designs in question are all based on trials or designers' observation on dataset. This is not difficult, but rather laborious and time-consuming. This paper is the first trial to replace such process with an automatic way.

2.2. Regulating Receptive Field with Input Variance

Adding input variance can also achieve dynamic receptive fields for a network. Zoomout [13] uses 4 input with different scales during inference to capture both contextual and local information. The DeconvNet [14] applies prepared detection bounding boxes and crops out object instances. Inferences are conducted on both these sub-images and the whole image.

Such approaches require complex pre- and post-processing. Meanwhile, they are computationally expensive as tens or even hundreds of forward propagations may be needed for one input image.

2.3. Affine Transformation in Deep Network

Affine transformations are usually seen in deep networks. Spatial Transformer Network (STN) [7] for character recognition uses a side branch to regress a set of affine parameters and applies corresponding transformation on feature maps. In [1], the network predicts both facial landmarks in original image and transformed sub-image. Then affine transformation parameters are obtained through the projection of these two sets of landmarks.

Our method is intrinsically different with the related works in question. Take STN for an example:

- Affine transformation is only the tool to solve different problems in these works. STN uses affine transformation to correct spatial variance of input data for recognition while our method is to regulate receptive field in parsing network.

- The different aims result in different network structures. Affine parameters in STN are data-dependent (obtained by forward) as each input is different. The parameter f in our method is embedded, knowledge-dependent (obtained by training) as receptive field should be stable during inference. Note that this work focuses on replacing manual receptive field selection process. Studies on using dynamic receptive fields are not taken into consideration here.
- As receptive field is all about sizes, the rotation functionality is discarded in this work, which is another noticeable difference with related works.

3. Approach

In this section, we will further elaborate on the details of our methods, including an overview on modified network structure, implementation of the inflation layer and interpolation layer and a loss guidance for multi-path network to realize a multi-scale inference with our data-driven method.

In this paper, we use both single-path and multi-path structures. The motivation is that almost all state-of-the-art deep image parsing networks are either single-path [11, 19, 2, 21] or multi-path [3]. We use these two structures to show that our method is effective and compatible with the state-of-the-arts.

3.1. Framework

Figure 1 presents the details of the framework. The specific settings of network backbone is listed in Table 1 in supplementary file. Using dilated convolutions, pooling strides in pool4 and pool5 are removed. The extent of receptive field for fc6 layer is 212×212. Note that we still use dilated convolutions in fc6 layers to generate different initial receptive fields. In experiment section we will present the improved performance brought by our method with improper initial receptive fields.

In the single path network, the inflation layer and the interpolation layer are inserted before fc6 layer and after fc8 layer respectively. The regulation of receptive field is operated on pool5 features. To reduce feature variance and add more robustness during optimization, we add a batch normalization (BN) [6] layer in front of the inflation layer.

While in multi-paths version, layers from BN to interpolation layer are paralleled, followed by a summation operation as feature fusion. The initializations of each parallels are the same. In order to break this symmetry and achieve discriminative, multi-scale inference, a loss guidance layer is added to enforce each parallel focus on different scales. These issues will be specified in the following subsections.

3.2. The Affine Transformation Layers

The affine transformation layers include *the inflation layer* and *interpolation layer*.

The inflation layer learns a parameter f, standing for *the inflation factor*. That is, the feature map will be enlarged by f times before the following convolution operations. Different from other deep networks with affine operations [7, 1], regulating receptive fields does not require cropping or rotations. Consequently, there is only one parameter in the inflation layer.

There are two steps in the inflation operation, namely coordinate transformation and sampling. To formulate the first process, let (x_i^s, y_i^s) and (x_i^t, y_i^t) to be the coordinates in the source feature map (input) and the target feature map (output) respectively. The inflation process builds up an element-wise coordinate projection as:

$$x_i^t = f \cdot x_i^s, \quad y_i^t = f \cdot y_i^s \ . \tag{1}$$

Also, the size of the feature map changes accordingly:

$$H^t = f \cdot (H^s - 1) + 1, \quad W^t = f \cdot (W^s - 1) + 1 \ . \tag{2}$$

where H and W are the height and width of feature maps, superscript t means 'target' and s means 'source'.

In the second step, we use a sampling kernel $\mathrm{k}(\cdot)$ to assign pixel values in target feature maps, which is denoted as V_i^c where i is pixel index, c is the channel index. Let U_i^c to be a pixel value in source feature maps, then we have:

$$\begin{gathered} V_i^c = \sum_n^{H^s} \sum_m^{W^s} U_{nm}^c \mathrm{k}(x_i^t, f, m) \mathrm{k}(y_i^t, f, n), \\ \forall i \in [1, ..., H^t W^t], \quad \forall c \in [1, ..., C] \ . \end{gathered} \tag{3}$$

Note that this operation is identical for each input channel. The sampling kernel $\mathrm{k}(\cdot)$ could be any differentiable image interpolation kernel. Here we use the bilinear kernel, where $\mathrm{k}(x, f, m) = \max(0, 1 - |\frac{x}{f} - m|)$, and we get:

$$\begin{gathered} V_i^c = \sum_n^{H^s} \sum_m^{W^s} U_{nm}^c \max(0, 1 - |\frac{x_i^t}{f} - m|) \max(0, 1 - |\frac{y_i^t}{f} - n|), \\ \forall i \in [1, ..., H^t W^t], \quad \forall c \in [1, ..., C] \ . \end{gathered} \tag{4}$$

The differential of V_i^c can also be obtained below.

$$\begin{aligned} \frac{\partial V_i^c}{\partial f} = \sum_n^{H^s} \sum_m^{W^s} U_{nm}^c \\ \cdot \left[\mathrm{k}(y_i^t, f, n) \frac{\partial \mathrm{k}(x_i^t, f, m)}{\partial f} + \mathrm{k}(x_i^t, f, m) \frac{\partial \mathrm{k}(y_i^t, f, n)}{\partial f} \right], \end{aligned} \tag{5}$$

where

$$\frac{\partial \mathrm{k}(x_i^t, f, m)}{\partial f} = \begin{cases} 0, & \text{if } |m - x_i^t/f| \geq 1 \\ -x_i^t/f^2, & \text{if } m \geq x_i^t/f \\ x_i^t/f^2, & \text{if } m < x_i^t/f \end{cases} \text{, and}$$

$$\frac{\partial \mathrm{k}(y_i^t, f, n)}{\partial f} = \begin{cases} 0, & \text{if } |n - y_i^t/f| \geq 1 \\ -y_i^t/f^2, & \text{if } n \geq y_i^t/f \\ y_i^t/f^2, & \text{if } n < y_i^t/f \end{cases} .$$

Together with the chain rule, the gradient from the inflation layer G_{inf} is:

$$G_{inf} = \sum_{c}^{C} \sum_{i}^{H^t \times W^t} \frac{\partial Loss}{\partial V_i^c} \cdot \frac{\partial V_i^c}{\partial f} . \tag{6}$$

Additionally, we normalize G_{inf} by dividing $H^t \times W^t$, which is the number of pixels in a target feature map.

$$G_{inf} = \frac{1}{H^t W^t} \sum_{c}^{C} \sum_{i}^{H^t \times W^t} \frac{\partial Loss}{\partial V_i^c} \cdot \frac{\partial V_i^c}{\partial f} . \tag{7}$$

The interpolation layer has almost the opposite functionality. In this layer, feature maps are resized back to a fixed size. The resize factor f' in interpolation layer is:

$$f' = F/f . \tag{8}$$

where F is a constant and is determined by desired output size. In our implementation F is 8.11 to resize the final result as large as label map or input image.

The interpolation layer is another source of the inflation factor's gradient:

$$\begin{aligned} G_{inter} &= \frac{\partial Loss}{\partial f'} \frac{\partial f'}{\partial f} \\ &= \frac{\partial Loss}{\partial f'} \left(\frac{-F}{f^2}\right) . \end{aligned} \tag{9}$$

where $\partial Loss/\partial f'$ has exactly the same form as (7). In practice, we simply add these two gradients together to update the inflation factor f:

$$\frac{\partial Loss}{\partial f} = G_{inf} + G_{inter} . \tag{10}$$

And when considering specific layers in our implementation, we can get:

$$\begin{aligned} \frac{\partial Loss}{\partial f} = & \frac{1}{H^{fc6} W^{fc6}} \sum_{c}^{C} \sum_{i}^{H^{fc6} W^{fc6}} \frac{\partial Loss}{\partial V_{bn,i}^c} \cdot \frac{\partial V_{bn,i}^c}{\partial f} \\ & - \frac{F}{H^{img} W^{img} f^2} \sum_{c}^{C} \sum_{i}^{H^{img} W^{img}} \frac{\partial Loss}{\partial V_{fc7,i}^c} \cdot \frac{\partial V_{fc7,i}^c}{\partial f'} . \end{aligned} \tag{11}$$

where C is channel amount in BN layer, subscript bn and img refer to BN layer and input image respectively.

In this way, it is possible to learn the inflation factor during the end-to-end training.

3.3. The New Receptive Field

To calculate the range of new receptive fields, we can transform the question to obtain an equivalent kernel size of fc6 layer while feature maps are unchanged. Denote the original kernel size as k, the new equivalent size is $k' = \lceil (k+1)/f \rceil$ according to Equation (2). Thus the extent of the new receptive field is $212 + 8 \times (k' - 1)$, where 212 is the receptive field in pool5 layer, 8 is the overall stride from conv1_1 layer to pool5 layer in the network backbone.

3.4. Loss Guidance for Multi-paths Network

Deep networks with multi-scale receptive fields have brought performance improvement in image parsing task [3]. This kind of network usually has several slightly different parallels to achieve multiple receptive fields. Our method can be also used in similar structures to realize further improvement and take place of hand-craft dilated convolutional kernels.

To achieve this, as shown in Figure 1(b), fc6, fc7 and fc8 layers are first copied to make parallels. The output of fc8s are fused by a summation operation. Then, inflation and interpolation layers are inserted before each fc6 layers and after fc8 layers. A shared BN layer is appended after pool5.

However, this framework is symmetric and is bad for learning discriminative features. To break this symmetry, a *weighted gradient layer* is added behind each interpolation layers during training. Similar to the *class-rebalancing* strategy in [20] and the *weighted loss* in [18], the weighted gradient layer multiplies a weight w (usually greater than 1) on the gradient values G_i^c if the ground truth label l_i of the correspondent pixel (i-th pixel in c-th channel) is in a given label set S. To formulate this process, we have:

$$G_{s,i}^c = w \, G_{t,i}^c , \quad w.r.t. \quad w = \begin{cases} W, \text{ if } l_i \in S \\ 1, \text{ if } l_i \notin S \end{cases} . \tag{12}$$

$G_{s,i}^c$ comes from source feature maps while $G_{t,i}^c$ comes from target feature maps. The set S contains labels that have similar sizes. For example, in face parsing experiment, we use{*eyes*, *eyebrows*} and $\varnothing$ for each parallels in bi-path model and {*eyes*, *eyebrows*}, {*nose*, *mouth*, *lips*} and $\varnothing$ in the tri-path model. Such weighted gradients will induce each branch to focus on different label, scales and thus lead to obtain discriminative receptive fields.

4. Experiment

We conduct experiments to show the superiority of our method on selecting a finer receptive field. The experiment consists of three parts:

- We first reproduce the receptive field searching process by using dilated convolutional kernels and find the optimal receptive field manually.

Table 1. Quantitative evaluation results of baseline models and modified models on Helen [8, 16] dataset. *'dilation'* means dilation values in fc6 layer. *'rf-fc6'* means the extent of receptive field in fc6 layer. $'*'$ means the inflation factor begins to be updated after 10,000 iterations in training. For modified models with large initial receptive fields, their performances are obviously improved. Please refer to Table 2 in supplementary file for full results.

Single Path Baseline Model		
dilation	rf-fc6	F-score
2	260	0.8995
4	**308**	**0.9012**
6	356	0.9001
8	404	0.8983
10	452	0.8965
12	500	0.8924
14	548	0.8849

Single Path Modified Model			
init dilation	f	rf-fc6	F-score
2	2.44	236	0.8964
2	0.88*	284	0.8952
6	1.82	292	0.8995
8	2.61	284	0.9021
10	2.44	336	0.9000
12	3.60	292	0.9005

Table 2. Quantitative evaluation results of baseline models and modified models on PASCAL VOC 2012 [4] validation set. Please refer to Table 5 in supplementary file for full results.

Single Path Baseline Model		
dilation	rf-fc6	mean IOU (%)
4	276	61.310
6	308	64.040
8	340	65.200
10	372	65.580
12	**404**	**65.540**
14	436	64.680
16	468	64.190
18	500	63.860
20	532	63.393

Single Path Modified Model			
init dilation	f	rf-fc6	mean IOU (%)
4	0.73	332	64.536
6	0.76	364	65.080
16	1.46	396	66.030
18	1.56	404	67.780
20	1.61	420	66.530

Table 3. Quantitative evaluation results of multi-paths versions of baseline models and modified models on Helen dataset [8, 16]. Each parallels in the modified network is initialized with dilation value of 8. The loss guidance has helped the network to break symmetry and acquire discriminative features. Please refer to Table 3 in supplementary file for full results.

Multi-paths Baseline Model			
model	dilation	rf-fc6	F-score
bipath	4,6	308,356	0.8964
tripath	4,6,8	308,356,404	0.8894

Multi-paths Modified Model			
model	f	rf-fc6	F-score
bipath	3.32,1.27	268,372	0.9008
tripath	1.61, 1.12, 1.11	340, 396, 396	0.8983

- With the network backbone intact, the single-path network is modified by inserting new affine transformation layers. The inflation factor is learned with different initial dilation values.
- We adopt the best two and best three receptive field settings according to results in the first experiment and build up a bi-path network and a tri-path network as baseline models. For modified models, paralleled paths are initiated with the same structure. By deploying the loss guidance, each parallel learns discriminative inflation factor and feature.

Results demonstrate the effectiveness of the proposed method to learn and obtain better receptive fields without much manual intervention.

4.1. Dataset and Data Pre-processing

The Helen dataset [8, 16] is used for face parsing task. The Helen dataset contains 2330 face images with 11 manually labelled facial components including eyes, eyebrows, noses, lips and mouths. The hair region is annotated through a matting algorithm without human correction so that it is not accurate enough comparing to other annotations. We adopt the same dataset division setting as in [18, 10] that uses 100 images for testing.

All images are aligned following similar steps in [10]. We use [17] to generate facial landmarks and align each image to a canonical position. After alignment, each image is cropped or padded and then resized to 500×500 pixels.

The augmented PASCAL VOC 2012 segmentation dataset is used for general image parsing task. The augmented PASCAL VOC 2012 segmentation dataset is composed of PASCAL VOC 2012 segmentation benchmark [4] and extra annotation provided by [5]. There are $12,031$ images for training and 1499 images for validation, consisting of 20 foreground object classes and one background class.

4.2. Implementation Details

Structures of models modified by our method are shown in Table 1 in supplementary file and Figure 1.

For face parsing task, we train each models with mini-batch gradient descent. The momentum, weight decay and batch size are set to be 0.9, 0.0005 and 2 respectively. The base learning rate is $1e^{-7}$ while the softmax loss is normalized by batch size. The total iteration is 55000 and the training process steps after 50000 iterations.

Meanwhile, the batch normalization layer uses its default settings. Inflation factors are initialized by 1 and their learning rates are base learning rates multiplied by a weight that ranges from $3e^4$ to $9e^4$. No weight decays are applied on inflation factors during training. Additionally, inflation factors are restricted within the range of [0.25, 4] in order to avoid numerical problems or exceptional memory usage.

For general image parsimg task, we realize its single-path version. The batch size is 20 and the learning rate multiplier of f is $3e^5$. The total iteration is 9600 with 3 steps. The great data variance in VOC dataset as well as data shuffle and random cropping strategies bring lots of obstacles for optimizing f. To add more robustness, some tricks are used during training: (a) clip exceptional $\partial Loss/\partial f$ values; (b) when updating f, gradients from background areas are multiplied by a weight (less than 1) to avoid the background area to be dominant (background mask); (c) the original step using a gamma of 0.1 is replaced with two smaller steps 200 iterations apart with gammas of 0.32.

Table 4. Quantitative evaluation on Helen dataset[8, 16]. Our method achieves state-of-the-art performance on face parsing task. Full results are in Table 4 in supplementary file.

Model	F-score
Liu et al.[9]	0.738
Smith et al.[16]	0.804
Liu et al.[10]	0.847
Our method	**0.9021**

4.3. Comparison with Manual Selection Method

4.3.1 Single Path Models

For face parsing task, we quantitatively evaluate and compare our model with baseline models using F-measures, as shown in Table 1. First, we manually search the best receptive field using dilated convolutional kernels based on baseline models. That is, set a series of dilation values on each model and evaluate their performance successively. The one with the highest F-score is selected as the optimal manually designed model.

Then the other unselected networks are modified with the proposed method where their receptive fields are treated as initializations. Results in Table 1 show that almost all modified models (except dilation value 2, which will be further discussed in Section 5.1) have witnessed improvement and their performances are compatible with that of the optimal manually designed model. The new receptive fields, e.g. 292, are more fine-grained and cannot be obtained by using dilation algorithm. And their performances stay abreast, or even has surpassed the best manually design model.

Qualitative comparisons for face parsing task is shown in Figure 1 in supplementary file. Results in Figure 1(d) and (e) show the improvements brought by our method. Smaller semantic areas have better parsing results, especially in eyebrows and nose. Face boundaries are smoother and more accurate. Results in Figure 1(c) and (d) show that the proposed models have compatible performance with manually designed models, which means the proposed method can replace previous receptive field selection methods.

For general image parsing task, the similar process is repeated. Evaluations are conducted on VOC validation set under mean IOU metric (or average Jaccard distance).

Table 2 demonstrates quantitative evaluation results. Modified models with initial dilation values of 16, 18 and 20 witness noticeable performance improvement that is compatible with the best manually designed model, and their receptive fields are regulated to an optimal range. Note that dilation convolutional kernels with current network backbone can not generate receptive field of 396, showing that the proposed method is able to generate receptive fields at a finer granularity.

Choosing different dilation values when initializing the modified models determines how much potential could be excavated from the proposed method. The modified models with small initial dilation values have improved parsing accuracy but still perform worse than the best manually designed one, which are mainly due to the shrinkage of features and the information lost. On the other hand, models with large initial dilation values perform better than the optimal baseline model. The reasons may vary, but one possible reason is that the modified models learn from data with dynamic sizes while f is changing, which has similar effect of data augmentation methods. These phenomena will be further discussed in Section 5.1.

Qualitative comparisons for general images parsing task is demonstrated in Figure 3. Results in (d) and (e), (f) and (g) show the improvements brought by our method. With finer receptive fields, results from modified model are generally more consistent. Results in (d) have clearer shapes and boundaries than results in (e). Results in (c), (f) and (g) show that if initial receptive field is not proper, performances of modified models are improved but still not compatible with the best manually designed models. Results in

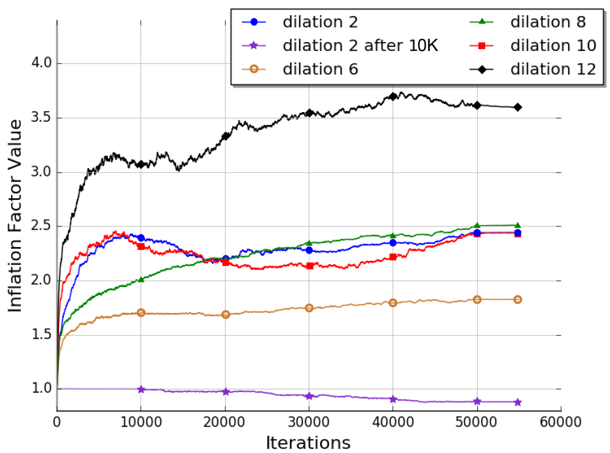

Figure 2. The fluctuation of f during training in face parsing task.

(c) and (d) show that, if initial receptive field is properly set, our models have very close performance with manually designed models, which means the proposed method can replace previous manual method on receptive field design.

These phenomena have proven that, with proper initial settings, the proposed method is able to help deep image parsing network find better receptive field automatically and guarantee to acquire a good performance that is equivalent to, or better than, the best manually designed one.

4.3.2 Multi-path Models

A bi-path network and a tri-path network are built for face parsing experiment. For baseline models, dilated convolutional kernels with top accuracy are selected, namely kernels with dilation value 4 (best overall performance with the highest eye F-score) and 6 (the highest nose and mouth F-score) for bi-path network, and dilation value 4, 6 and 8 (the highest face F-score) for tri-path network.

By comparison, the parallels in both modified bi-path and tri-path networks are symmetric with initial dilation value of 8. Weight w in weighted gradient layer is 1.2.

Results in Table 3 show that the proposed method is able to obtain better receptive field for each parallel with superior performance than the manually designed network. Also, the loss guidance manages to break symmetry in network structure and learn discriminative features.

4.3.3 Comparison with Previous Face Parsing Method

Table 4 shows a quantitative face parsing comparison between our method and other state-of-the-art methods. We use reported results from [9], [16] and [10]. Our method uses the single path network with the initial dilation value of 8. Even without CRF or RNN post-process, our method still achieves the highest accuracy.

5. Discussion

5.1. Choosing Proper Initial Receptive Fields

Although our method has strong ability on regulating receptive fields, but to make the best use, not all initial dilation values are good choices. Figure 2 and Figure 2 in supplementary file demonstrate some typical fluctuations of f during training in the both tasks.

For initial receptive field much smaller than the desired one, f is hard to optimize as the network will try to keep it larger than 1 (see line 'dilation 2'). The shrinkage of features will result in losing information and thus impair parsing performance. In face parsing task, even with some tricks, e.g. begin to update f after 10k iterations (see line 'dilation 2 after 10k' in Figure 2), f goes down but won't reach the value as expected. Consequently, performances of modified models with small initial receptive field are improved but still not compatible with the best manually designed models. When it comes to general image parsing task, models with small initial dilation values sometimes are trapped in local minimums where f fluctuates within the vicinity larger than 1 (see Figure 3 in supplementary file). On the other hand, using extremely greater initial dilations requires learning greater f, which means unaffordable memory load and time cost as feature maps become much larger accordingly. In summary, our suggestion is: use large dilation values for initialization, but not arbitrarily large.

5.2. Optimization in General Dataset

Unlike face parsing task where images are coarsely aligned and semantic constituents from different images are of similar sizes (e.g. eyes, lips), object sizes in general dataset have much greater variance, making optimizing f rather more difficult. Even with proper initialization and the same network settings, f stays in a certain range but not a specific value (see Figure 4 in supplementary file). Results shown in Table 2 are typical examples.

6. Conclusion

In this paper, we introduce a new regulation method for receptive fields in deep image parsing network automatically. This data-driven approach is able to replace the existing hand-craft receptive field selection methods as it enables a deep image parsing network obtains better receptive fields at finer granularity in only one training process. Experimental results on Helen dataset and PASCAL VOC 2012 dataset demonstrate the efficiency and effectiveness of our method over existing methods.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No.U1536203, Grant 61572493

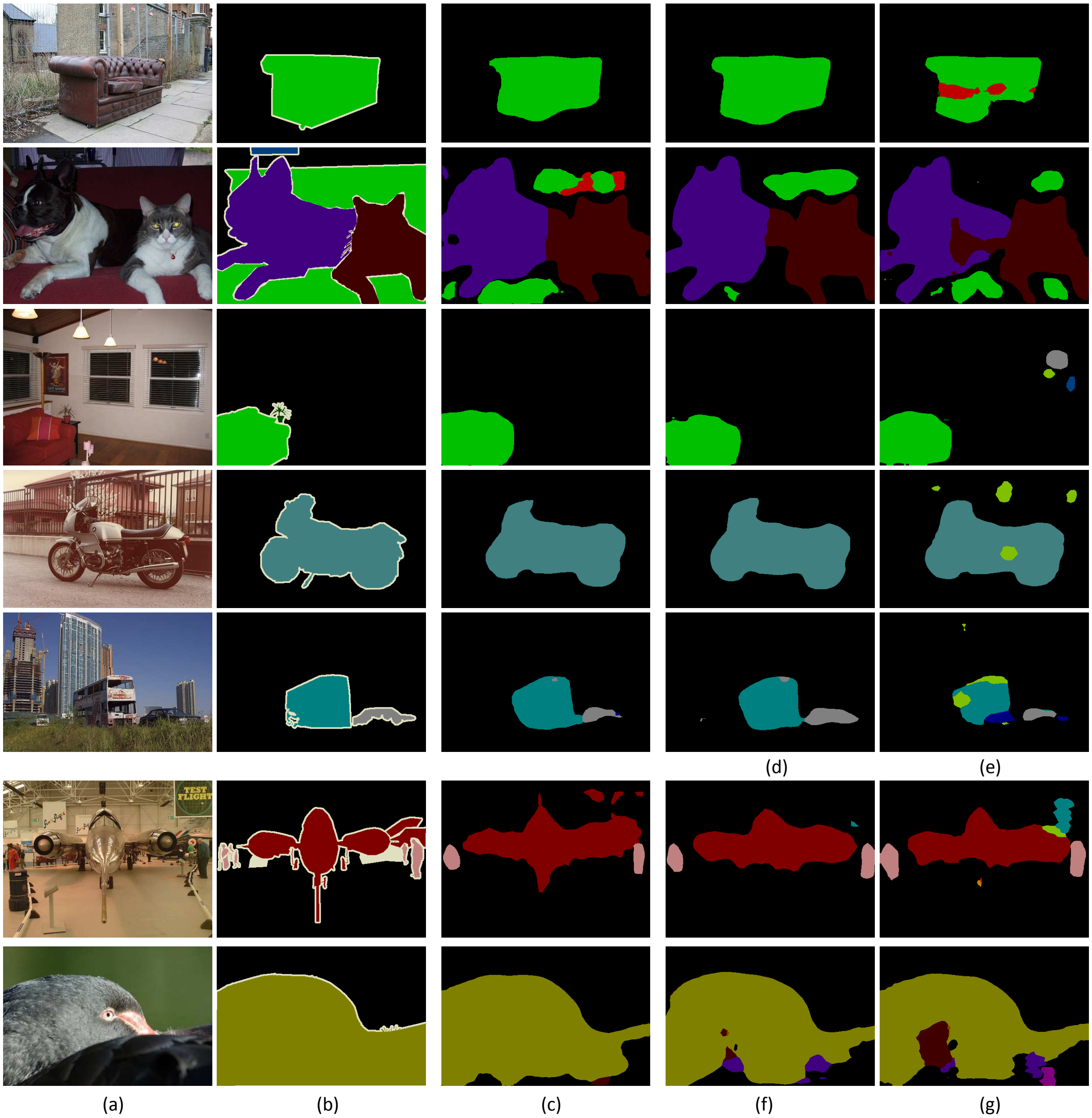

Figure 3. General image parsing results on PASCAL VOC 2012 validation set[4]. (a): original images. (b): ground truth. (c): results from baseline model with dilation value of 12(with best manually selected receptive field). (d): results from *modified model* with initial dilation value of 20. (e): results from *baseline model* with dilation value of 20. (f): results from *modified model* with initial dilation value of 4. (g): results from *baseline model* with dilation value of 4. Results in (d) and (e), (f) and (g) show the improvements brought by our method. With finer receptive fields, results from modified model are generally more consistent. Results in (d) have clearer shapes and boundaries than results in (e). Results in (c), (f) and (g) show that with improper initial receptive field, performances of modified models are improved but still not compatible with the best manually designed model. Results in (c) and (d) show that, if initial receptive field is properly set, the proposed models have compatible performance with manually designed model, which means our method can replace previous receptive field design process. Best view in color.

61602530, U161126411301523) and the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) 201600035. We also would like to thank NVIDIA for GPU donation.

References

[1] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. *arXiv:1607.05477*, 2016.

[2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv:1412.7062*, 2014.

[3] L. 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:1606.00915*, 2016.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, IJCV*, 88(2):303–338, June 2010.

[5] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *13th European Conference on Computer Vision ECCV*, 2014.

[6] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, pages 448–456, 2015.

[7] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, NIPS 2015*, pages 2017–2025, 2015.

[8] V. Le, J. Brandt, Z. Lin, L. D. Bourdev, and T. S. Huang. Interactive facial feature localization. In *12th European Conference on Computer Vision ECCV*, pages 679–692, 2012.

[9] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE Transaction on Pattern Analysis and Machine Intelligence, TPAMI*, 33(12):2368–2382, 2011.

[10] S. Liu, J. Yang, C. Huang, and M. Yang. Multi-objective convolutional learning for face labeling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3451–3459, 2015.

[11] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3431–3440, 2015.

[12] J. Long, N. Zhang, and T. Darrell. Do convnets learn correspondence? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, NIPS 2014*, pages 1601–1609, 2014.

[13] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3376–3385, 2015.

[14] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision, ICCV 2015*, pages 1520–1528, 2015.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[16] B. M. Smith, L. Zhang, B. Jonathan, Z. Lin, and J. Yang. Exemplar-based face parsing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pages 3484–3491, 2013.

[17] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3476–3483, 2013.

[18] T. Yamashita, T. Nakamura, H. Fukui, Y. Yamauchi, and H. Fujiyoshi. Cost-alleviative learning for deep convolutional neural network-based facial part labeling. *Ipsj Transactions on Computer Vision and Applications*, 7:99–103, 2015.

[19] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015.

[20] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. *arXiv:1603.08511*, 2016.

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