Computational Visual Media
https://doi.org/10.1007/s41095-018-0112-1

Research Article

Learning adaptive receptive fields for deep image parsing networks

Zhen Wei[1,2], **Yao Sun**[1] (✉), **Junyu Lin**[3], **and Si Liu**[1,4]

Abstract In this paper, we introduce a novel approach to automatically regulate receptive fields in deep image parsing networks. Unlike previous work which placed 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 are inflated or shrunk by the new layer, thereby changing the receptive fields in the following layers. By use of end-to-end training, the whole framework is data-driven, without laborious manual intervention. The proposed method is generic across datasets and different tasks. We have conducted extensive experiments on both general image parsing tasks, and face parsing tasks as concrete examples, to demonstrate the method's superior ability to regulate over manual designs.

Keywords semantic segmentation; receptive field; data-driven; face parsing

1 Introduction

In deep neural networks, the notion of a *receptive field* refers to the data that are path-connected to a neuron [1]. After the introduction of *fully convolutional networks* (FCN) [2], receptive fields have become especially important for deep image parsing networks; they can significantly affect the network's performance. As discussed in Ref. [3], a small receptive field may lead to inconsistent parsing results for large objects while a large receptive field may ignore small objects and classify them as background. Even if such extreme problems do not arise, unsuitable receptive fields can still impair performance.

Recent works such as Refs. [4, 5] have already discussed adapting network structures to use different receptive fields. Dilated convolutional kernels are often used for this purpose. The convolutional kernels' receptive field size can be controlled by appropriate choice of dilation values (typically integers). However, this approach has several main drawbacks that should be addressed. Firstly, these dilation values are treated as hyper-parameters in network design. The selection of dilation values is based on the designer's observations or the results of a series of trials on a certain dataset, which is laborious and time-consuming. Secondly, such choices are not generic across different image parsing tasks, or even various datasets given the same task. During network transfer, the selection procedure must be performed again. Thirdly, dilated convolutional kernels only lead to discrete values for receptive fields. When a dilation value is incremented, the corresponding receptive field (e.g., the fc6 layer in VGG [6]) may expand by tens or even hundreds of pixels, making it hard to accurately control the receptive field.

The contribution of this paper is a learning-based, data-driven method for automatically regulating receptive fields in deep image parsing networks. The main idea is to introduce a novel affine transformation layer, the *inflation layer*, before the convolutional layer whose receptive field is to be regulated. This

1 State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China. E-mail: Z. Wei, weizhen@iie.ac.cn; Y. Sun, sunyao@iie.ac.cn (✉); S. Liu, liusi@iie.ac.cn.

2 University of Chinese Academy of Sciences, Beijing 101408, China.

3 Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China. E-mail: linjunyu@iie.ac.cn.

4 Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, Nanjing 210094, China.

Manuscript received: 2017-12-06; accepted: 2018-01-14

inflation layer uses interpolation algorithms to enlarge or shrink feature maps. The following layers perform inference on these inflated features, thus changing the receptive fields after the inflation layer. Then, inference results (before softmax normalization) are resized to a fixed size by an *interpolation layer*. During training, the *inflation factor*, f, embedded in both the inflation layer and the interpolation layer has computable derivatives, and is trained end-to-end together with the network backbone. As f may be a real number, the inflation layer can produce a more fine-grained receptive field, and is trained only once.

To corroborate the method's effectiveness, we have conducted experiments on both general image parsing tasks as well as face parsing. With proper initialization, the proposed method can achieve comparable, or even superior, results compared to the best manually selected dilated convolutions. In particular, due to the strong regulation ability brought by our method, the improved model achieves state-of-the-art face parsing accuracy on the Helen dataset [7, 8].

The rest of this paper is organized as follows. In Section 2, we review related work on image parsing, especially focusing on issues relevant to receptive fields. Section 3 provides details of the new affine transformation layer and derivatives of the inflation factor f. Section 4 describes the experimental settings, while Section 5 discusses our experimental results. Section 6 concludes the paper.

This paper extends our former conference publication [9]. Additional content here mainly includes: (i) more elaborate discussion of several issues during optimization (in Section 5), (ii) detailed network settings used in experiments (in Table 1), and (iii) further qualitative and quantitative results (in Tables 2–5 and Figs. 2 and 3).

2 Related work

This section provides a brief review and discussion of related work.

2.1 FCNs and dilated convolution

The introduction of FCNs [2] has emphasised the importance of receptive fields. The forward process of FCNs to generate dense classification results is equivalent to a series of inferences using sliding windows on the input image. Using fixed stride sliding, inference at the pixel-level is solely based on data inside the window. The window is, in this case, the receptive field of the network. In Ref. [2], the authors discuss dilated convolution, but do not make use of it in their network. DeepLab [4] uses dilated convolutions to reduce pooling strides while expanding receptive fields and reducing the number of parameters in the fc6 layer. In Ref. [5], the authors append a series of dilated convolutional layers after an FCN backbone (the *frontend*) to expand the receptive field. Recently, in DeepLab v2 [10], the authors manually designed four different dilated convolutions which are used in parallel to achieve multi-scale parsing.

However, these dilation designs are all based on trials or the designers' observations of the dataset. This is not difficult, but nevertheless is laborious and time-consuming. This paper offers the first way to replace such a process with an automatic method.

2.2 Regulating receptive fields with input variance

Adding input variability can also be used to provide dynamic receptive fields for a network. Zoomout [11] uses 4 inputs at different scales during inference to capture both contextual and local information. DeconvNet [3] applies prepared detection bounding boxes and crops out object instances. Inference is conducted on both these sub-images and the whole image.

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

2.3 Affine transformation in deep networks

Affine transformations are usually seen in deep networks. Spatial transformer network (STN) [12] for character recognition uses a side branch to regress a set of affine parameters and applies the corresponding transformation to feature maps. In Ref. [13], the network predicts both facial landmarks in both the original image and the transformed sub-images. Affine transformation parameters are then obtained by projection of these two sets of landmarks.

Our method intrinsically differs from such related work. Taking STN for example:

- Affine transformation is only the tool used by STN to solve various problems; in particular, STN uses

affine transformation to correct spatial variability of input data for recognition. Our method regulates the receptive field in the parsing network.

- The different aims result in different network structures. Affine parameters used in STN are data-dependent, as each input is different. The parameter f in our method is embedded, and knowledge-dependent (obtained by training): the receptive field should be stable during inference. Our work focuses on replacing the manual receptive field selection process. Studies on use of dynamic receptive fields are not taken into consideration here.
- As the receptive field depends only on size, rotation functionality is discarded in this work, unlike other work.

In Ref. [14], *deformable convolutions* are used to reformulate the sampling process in convolutions in a learning-based approach. Deformable convolutions can also be regarded as a way of reallocating convolutional weights. If nearby weights in lower layers are increased, the receptive fields of the corresponding weights in higher layers become smaller, and vice versa.

3 Approach

In this section, we provide details of our methods, including the modified network structure, implementation of the inflation and interpolation layers, and loss guidance for our multi-path network. These allow us to realize multi-scale inference with our data-driven method.

We use both single-path and multi-path structures. Almost all state-of-the-art deep image parsing networks are either single-path [2, 4, 5, 15] or multi-path [10], so we use these two structures to show that our method is effective and compatible with such state-of-the-art methods.

3.1 Framework

Figure 1 presents the details of our framework. The specific settings for the network backbone are provided in Table 1. Using dilated convolutions, pooling strides in pool4 and pool5 are removed. The extent of the receptive field for the fc6 layer is 212×212. Note that we still use dilated convolutions in the fc6 layer to generate different initial receptive fields.

In the single-path network, the inflation layer and the interpolation layer are inserted before layer fc6 and after layer fc8 respectively. The receptive field is regulated by pool5 features. To reduce feature variability and increase robustness during optimization, we add a batch normalization (BN) [16] layer before the inflation layer.

In the multi-path version, layers from BN to the interpolation layer are duplicated, and followed by a summation operation for feature fusion. Each duplicate is initialized in the same way. In order

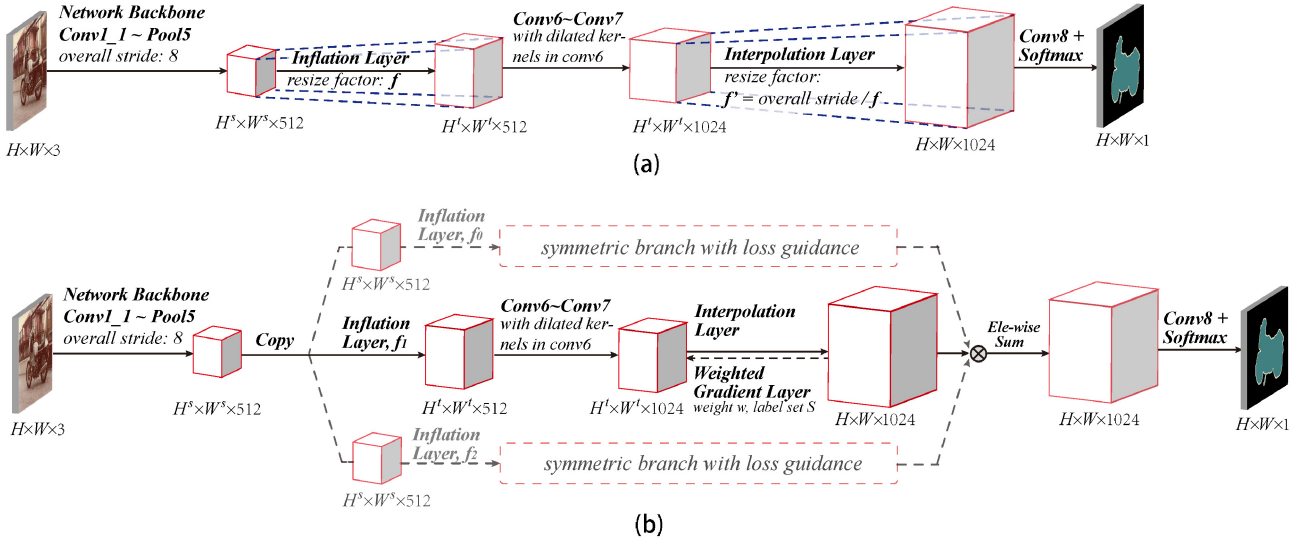


Fig. 1 Framework. (a) Modified single-path network. New layers are inserted before layer fc6 and after layer fc8. (b) Modified multi-path network in which all branches have the same structure and initialization. *Weighted gradient layers* are used to break symmetry during training. The specific settings for the single-path network are given in Table 1.

Table 1 Network structures used, including network backbone, single-path baseline model, and single-path modified model

Network backbone		
conv 1_1–conv 1_2	output dim: 64, kernel size: 3, pad: 1	
pool 1	MAX pooling, stride: 2, kernel size: 2, pad: 1	
conv 2_1–conv 2_2	output dim: 128, kernel size: 3, pad: 1	
pool 2	MAX pooling, stride: 2, kernel size: 2, pad: 1	
conv 3_1–conv 3_3	output dim: 256, kernel size: 3, pad: 1	
pool 3	MAX pooling, stride: 2, kernel size: 2, pad: 1	
conv 4_1–conv 4_3	output dim: 512, kernel size: 3, pad: 1	
pool 4	MAX pooling, stride: 1, kernel size: 2, pad: 1	
conv 5_1–conv 5_3	output dim: 512, kernel size: 3, pad: 2, dilation: 2	
pool 5	MAX pooling, stride: 1, kernel size: 3, pad: 1	
	Single-path baseline model	**Single-path modified model**
inflation layer	—	✓
conv 6	output dim: 1024, kernel size: 4 (Helen), 3 (VOC), pad: (dilation*(kernel size−1))/2	
conv 7	output dim: 512 (Helen), 1024 (VOC), kernel size: 1	
interpolation layer	—	✓
output layer	output dim: 11 (Helen), 21 (VOC)	

to break this symmetry and achieve discriminative multi-scale inference, a loss guidance layer is added to enforce each duplicate to focus on a different scale. These issues will be explained in detail in the following subsections.

3.2 Affine transformation layers

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

The inflation layer learns a parameter f, the *inflation factor*. The feature map is enlarged by the factor f before the following convolution operations. Unlike other deep networks with affine operations [12, 13], regulating receptive fields does not require cropping or rotation, so only one parameter is needed in the inflation layer.

There are two steps in the inflation operation, coordinate transformation and sampling. To formulate the first process, let $(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})$ and $(x_i^{\mathrm{t}}, y_i^{\mathrm{t}})$ be coordinates in the source feature map (input) and target feature map (output) respectively. The inflation process performs element-wise coordinate projection using:

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

The size of the feature map changes accordingly:

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

where H and W are the height and width of the feature maps, superscripts s and t meaning "source" and "target", respectively.

In the second step, we use a sampling kernel $k(\cdot)$ to assign pixel values in target feature maps. It is denoted by V_i^c where i is the pixel index and c is the channel index. Let U_i^c be a pixel value in a source feature map. Then we have

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

This operation is identical for each input channel. The sampling kernel $k(\cdot)$ could be any differentiable image interpolation kernel. Here we use the bilinear kernel: $k(x, f, m) = \max(0, 1 - |x/f - m|)$, giving

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

The derivative of V_i^c can be obtained as

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

where

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

and

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

Using the chain rule, the gradient from the inflation layer G_{inf} is

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

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

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

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

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

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

The interpolation layer provides a further contribution to the inflation factor's gradient:

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

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

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

When considering specific layers in our network, we obtain:

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

where C is the number of channels in the BN layer, and subscripts bn and img refer to the BN layer and input image respectively.

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

3.3 New receptive field

To calculate the ranges of the new receptive fields, we can transform the question to one of obtaining an equivalent kernel size for the fc6 layer while leaving the feature maps unchanged. Denoting the original kernel size by k, Eq. (2) gives the new equivalent size: $k' = \lceil (k+1)/f \rceil$. Thus the extent of the new receptive field is $212 + 8 \times (k' - 1)$, where 212 is the receptive field in the pool5 layer, and 8 is the overall stride from the conv1_1 layer to the pool5 layer in the network backbone.

3.4 Loss guidance for multi-path networks

Deep networks with multi-scale receptive fields have brought performance improvements in image parsing tasks [10]. Such networks usually use several slightly different parallel paths to achieve multiple receptive fields. Our method can be also used in similar structures to realize further improvements, taking the place of hand-craft dilated convolutional kernels.

To achieve this, as shown in Fig. 1(b), layers fc6, fc7, and fc8 are first copied in parallel. The outputs of the fc8 layers are fused by summation. Then, inflation and interpolation layers are inserted before each fc6 layer and after each fc8 layer. A shared BN layer is appended after pool5.

However, this framework is symmetric and is unsuited to learning discriminative features. To break this symmetry, a *weighted gradient layer* is added behind each interpolation layer during training. Following the *class-rebalancing* strategy in Ref. [17] and the use of *weighted loss* in Ref. [18], the weighted gradient layer weights the gradient values G_i^c if the ground truth label l_i of the corresponding pixel (the ith pixel in the cth channel) is in a given label set S. The weight w is usually greater than 1. Thus

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

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

4 Experiments

We conducted experiments to show the superiority of our method, in its ability to select a finer receptive

field. The experiment consisted of three parts:

- We first reproduced the receptive field search process by using dilated convolutional kernels and found the optimal receptive field manually.
- Leaving the network backbone intact, the single-path network was modified by inserting new affine transformation layers. The inflation factor was learned with different initial dilation values.
- We used the best two and three receptive field settings according to the results in the first experiment to build a bi-path network and a tri-path network as baseline models. For modified models, parallel paths were constructed with the same structure. By deploying loss guidance, each parallel path learned a discriminative inflation factor and feature.

Results demonstrate the effectiveness of the proposed method to learn and obtain better receptive fields with little manual intervention.

4.1 Dataset and data preprocessing

The *Helen dataset* [7, 8] was used in the face parsing task. It contains 2330 facial images with 11 manually labelled facial components including eyes, eyebrows, nose, lips, and mouth. The hair region is annotated b; it is thus not accurate enough for comparison. We adopted the same dataset division setting as in Refs. [18, 19], using 100 images for testing.

All images were aligned using similar steps to those in Ref. [19]. We used Ref. [20] to generate facial landmarks and align each image to a canonical position. After alignment, each image was cropped or padded and then resized to 500×500 pixels.

The *augmented PASCAL VOC 2012 segmentation dataset* was used in the general image parsing task. The augmented PASCAL VOC 2012 segmentation dataset is based on the PASCAL VOC 2012 segmentation benchmark [21] with extra annotation provided by Ref. [22]. It has 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 and Fig. 1.

In the *face parsing task*, we trained each model with mini-batch gradient descent. The momentum, weight decay, and batch size were set to be 0.9, 0.0005, and 2 respectively. The base learning rate was $1e^{-7}$ while the softmax loss was normalized by batch size. A total of 55,000 iterations were used; training stopped after 50,000 iterations.

The batch normalization layer used default settings. Inflation factors were initialized to 1 and their learning rates were base learning rates multiplied by a weight ranging from 3×10^4 to 9×10^4. No weight decays were applied to inflation factors during training. Inflation factors were restricted to the range [0.25, 4] in order to avoid numerical problems or exceptional memory usage.

In the *general image parsing task*, we realized its single-path version with a batch size of 20 and a learning rate multiplier f of 3×10^5. A total of 9600 iterations were used with 3 steps. The great data variability in the VOC dataset as well as data shuffling and random cropping strategies provided significant obstacles to optimizing f. To increase robustness, the following strategies were used during training: (a) clipping exceptional $\partial \text{Loss}/\partial f$ values; (b) when updating f, gradients from background areas were masked by multiplying them by a weight less than 1, preventing them from becoming dominant; (c) the original step using a γ value of 0.1 was replaced by two smaller steps 200 iterations apart with γ values of 0.32.

4.3 Comparison with manual selection

4.3.1 Single-path models

In the *face parsing task*, we quantitatively evaluated and compared our model with baseline models using F-measures: see Table 2. First, we manually determined the best receptive field using dilated convolutional kernels based on the baseline models, by using a series of dilation values for each model and selecting the one providing the highest F-score as the optimal manually designed model.

Next, the other unselected networks were modified using our proposed method, and their receptive fields were used for initialization. Results in Table 2 show that almost all modified models (except for a dilation value of 2, which will be further discussed in Section 5.1) have witnessed improvement, providing results comparable to those from the optimal manually designed model. The new receptive fields, e.g., of size 292, are more fine-grained and cannot be obtained by use of the dilation algorithm. Their results are equal to, or even surpass, those of the best manually design models.

Table 2 Quantitative evaluation of baseline models and modified models using the Helen dataset. Key: *dilation*: dilation values in the fc6 layer. *rf-fc6*: the extent of the receptive field in the fc6 layer. ∗: the inflation factor starts being updated after 10,000 iterations in training

Single-path baseline model							
network setting		F-score					
dilation	rf-fc6	eye	eyebrow	nose	mouth	face	overall
2	260	0.8372	**0.7842**	0.9341	0.9073	0.9417	0.8995
4	308	**0.8459**	0.7839	0.9378	0.9103	0.9435	**0.9012**
6	356	0.8355	0.7787	**0.9385**	**0.9135**	0.9453	0.9001
8	404	0.8321	0.7703	0.9384	0.9093	**0.9453**	0.8983
10	452	0.8322	0.7713	0.9355	0.9068	0.9436	0.8965
12	500	0.8299	0.7665	0.9332	0.8991	0.9433	0.8924
14	548	0.8232	0.7486	0.9276	0.8989	0.9414	0.8849
Single-path modified model							
init dilation	f	rf-fc6	eye	eyebrow	nose	mouth	face	overall
2	2.44	236	0.8315	0.7795	0.9280	0.9052	0.9384	0.8964
2	0.88*	284	0.8295	0.7754	0.9297	0.9077	0.9389	0.8952
6	1.82	292	0.8433	0.7843	0.9310	0.9140	0.9415	0.8995
8	2.61	284	**0.8466**	**0.7861**	0.9365	**0.9148**	0.9148	**0.9021**
10	2.44	316	0.8437	0.7765	**0.9374**	0.9114	**0.9446**	0.9000
12	3.60	292	0.8412	0.7822	0.9367	0.9114	0.9441	0.9005

Qualitative comparisons for the face parsing task are provided in Fig. 2. Results in Figs. 1(d) and 1(e) show the improvements brought by our method. Smaller semantic areas are parsed better, especially the eyebrows and nose. Face boundaries are smoother and more accurate. Results in Figs. 1(c) and 1(d) show that the proposed models provide comparable results to manually designed models: our method can replace previous manual receptive field selection methods.

For the *general image parsing task*, a similar process was used. Evaluation was conducted using the VOC validation set under a mean IOU metric (average Jaccard distance).

Table 3 provides quantitative results. Modified models with initial dilation values of 16, 18, and 20 show noticeably improved results that are comparable with the best manually designed models, with receptive fields adjusted to an optimal range. Note that with the current network backbone, dilation convolutional kernels cannot generate receptive field of size 396, showing that the proposed method can generate receptive fields with a finer granularity.

Choosing different dilation values when initializing the modified models helps to evaluate the potential of the proposed method. Modified models with small initial dilation values have improved parsing accuracy but still perform worse than the best manually designed one, mainly due to the shrinkage of features and information loss. 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, with similar effects to those from data augmentation methods. These phenomena are further discussed in Section 5.1.

Qualitative comparisons for the general image parsing task are provided in Fig. 3. Results in (d) and (e), (f) and (g) show the improvements brought by our method. With finer receptive fields,

Table 3 Quantitative evaluation of baseline models and modified models using the PASCAL VOC 2012 validation set

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

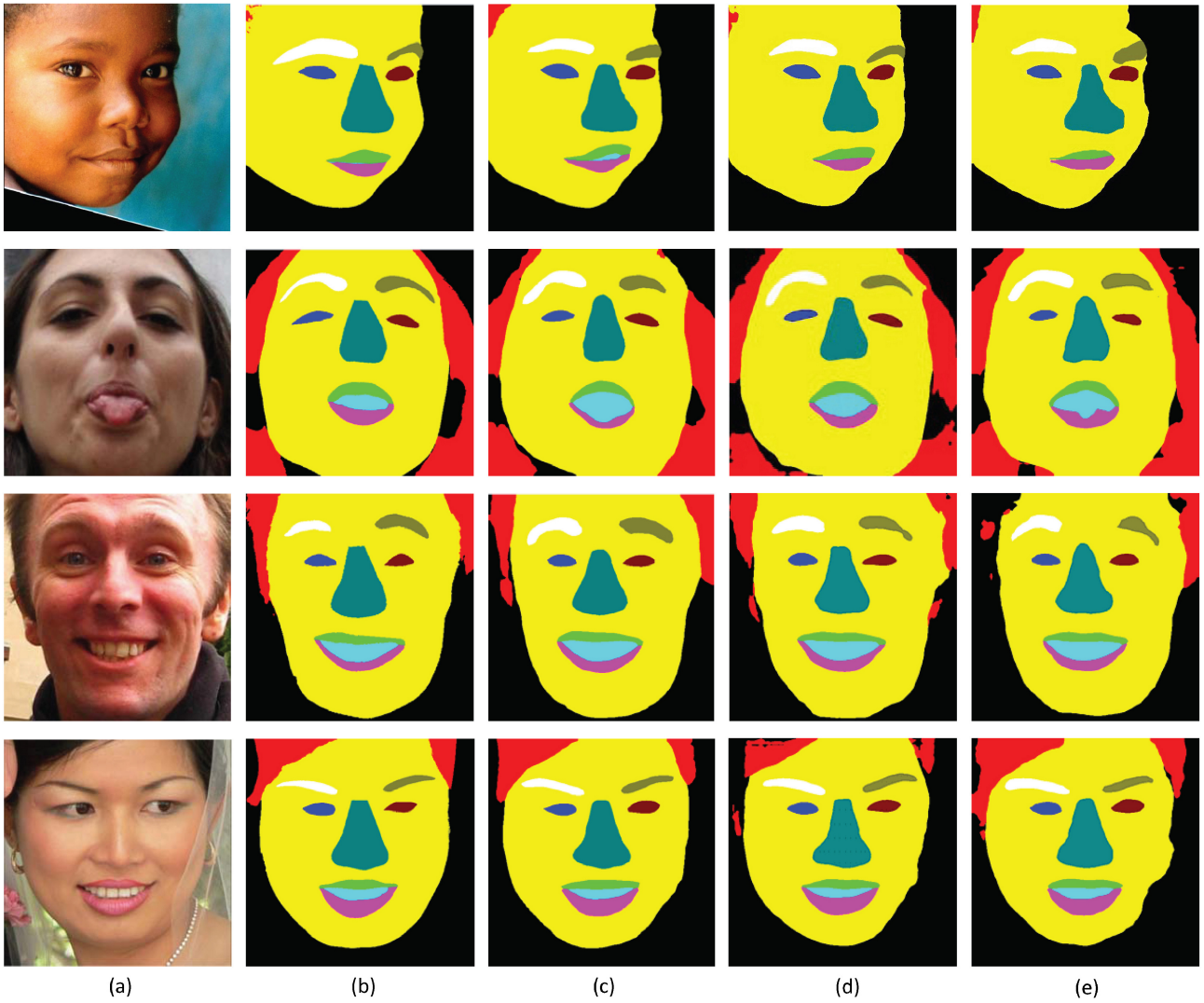


Fig. 2 Face parsing results for the Helen dataset. (a) Original images. (b) Ground truth. (c) Baseline model with dilation value of 4 (with best manually selected receptive field). (d) Modified model with initial dilation value of 12. (e) Baseline model with dilation value of 12. (d) and (e) show the improvements brought by our method. Smaller semantic areas have better parsing results, especially *eyebrows*, *nose*. *Face boundaries* are smoother and more accurate. (c) and (d) show that our models have very similar ability to manually designed models: our method can replace manual receptive field design processes.

results from modified models 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 an unsuitable initial receptive field is used, while modified models are improved, they are still not comparable to the best manually designed models. Results in (c) and (d) show that, if the initial receptive field is appropriately set, our models provide results very close to those of manually designed models: our method can replace previous manual methods of receptive field design.

These results demonstrate that, with proper initial settings, the proposed method is able to help deep image parsing networks find better receptive fields automatically, providing results that are 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 were built for use in a face parsing experiment. As baseline models, dilated convolutional kernels with best accuracy were selected: kernels with dilation values 4 (best overall results, with highest eye F-score) and 6 (highest nose and mouth F-scores) for the bi-path network, and dilation values of 4, 6, and 8 (providing highest face F-score) for the tri-path network.

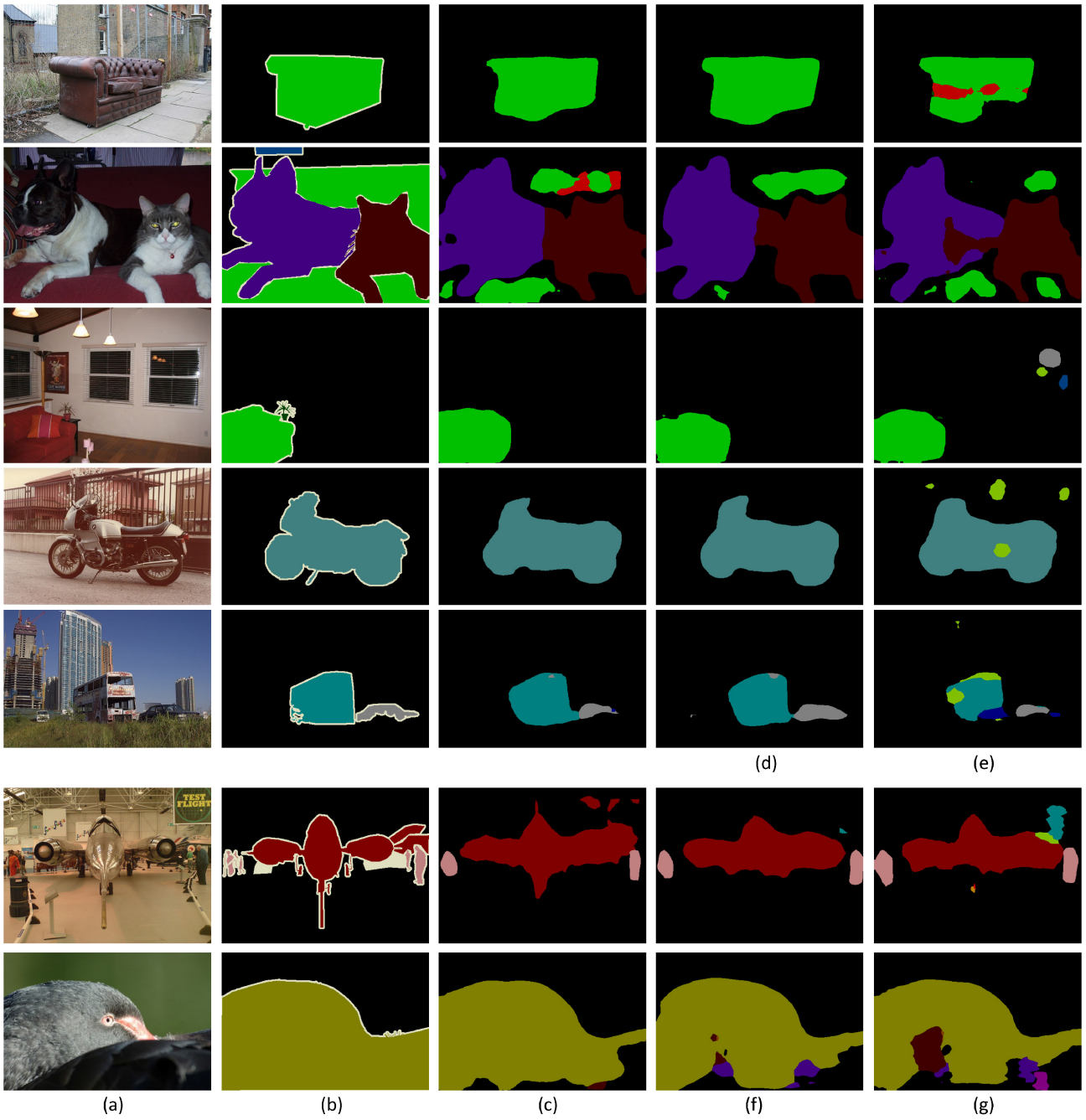


Fig. 3 General image parsing results on the PASCAL VOC 2012 validation set. (a) Original images. (b) Ground truth. (c) Baseline model with dilation value of 12 and best manually selected receptive field. (d) Modified model with initial dilation value of 20. (e) Baseline model with dilation value of 20. (f) Modified model with initial dilation value of 4. (g) 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 the 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 poor initial receptive fields, modified models are still improved but not as good as the best manually designed models. Results in (c) and (d) show that, if the initial receptive field is properly set, our model has comparable performance to the manually designed model: our method can replace previous receptive field design processes.

As a comparison, the parallel paths in both modified bi-path and tri-path networks were symmetric using an initial dilation value of 8. The weight w used in the weighted gradient layer was 1.2.

Results in Table 4 show that the proposed method is able to obtain better receptive fields for each parallel path, providing superior results to the manually designed network. We observe that the loss guidance manages to break symmetry in the network structure and learn discriminative features.

Table 4 Quantitative evaluation of multi-path versions of baseline models and modified models using Helen dataset [7, 8]. Each parallel path in the modified network was initialized with a dilation value of 8

Multi-path baseline model								
network setting			F-score					
model	dilation	rf-fc6	eye	eyebrow	nose	mouth	face	overall
bi-path	4, 6	308, 356	0.8368	0.7757	0.9309	0.9104	0.9423	0.8964
tri-path	4, 6, 8	308, 356, 404	0.8315	0.7638	0.9257	0.9044	0.9402	0.8894
Multi-path modified model								
model	f	rf-fc6	eye	eyebrow	nose	mouth	face	overall
bi-path	3.32, 1.27	268, 372	0.8401	0.7888	0.9316	0.9129	0.9418	0.9008
tri-path	1.61, 1.12, 1.11	340, 396, 396	0.8413	0.7763	0.9365	0.9098	0.9430	0.8983

4.3.3 Comparison with previous face parsing methods

Table 5 shows a quantitative comparison for face parsing between our method and other state-of-the-art methods. We use reported results from Refs. [8, 19, 23]. Our method used a single-path network with an initial dilation value of 8. Even without CRF or RNN post-processing, our method still achieves the highest accuracy.

5 Discussion

5.1 Choosing proper initial receptive fields

Although our method has a strong ability to regulate receptive fields, to get the best results, suitable initial dilation values must be chosen. Figures 4 and 5 demonstrate typical fluctuations in f during training for both tasks.

Using initial receptive fields much smaller than the desired one, f is hard to optimize as the network attempts to keep it larger than 1 (see "dilation 2" in Fig. 4). The shrinkage of features results in information loss, impairing parsing performance. In the face parsing task, even with some strategies, e.g., beginning to update f after 10k iterations (see "dilation 2 after 10k" in Fig. 4), f goes down but does not reach the value expected. Consequently, modified models with small initial receptive fields provide improved results, but they are still not comparable to those from the best manually designed models. When it comes to the general image parsing task, models with small initial dilation values are sometimes trapped in local minima where f fluctuates around a value larger than 1 (see Fig. 6). On the other hand, using extremely large initial dilations requires more extensive learning of f, leading to unaffordable memory loads and time costs, as feature maps accordingly become much larger. In summary, our suggestion is to use moderately large dilation values for initialization, but not too large.

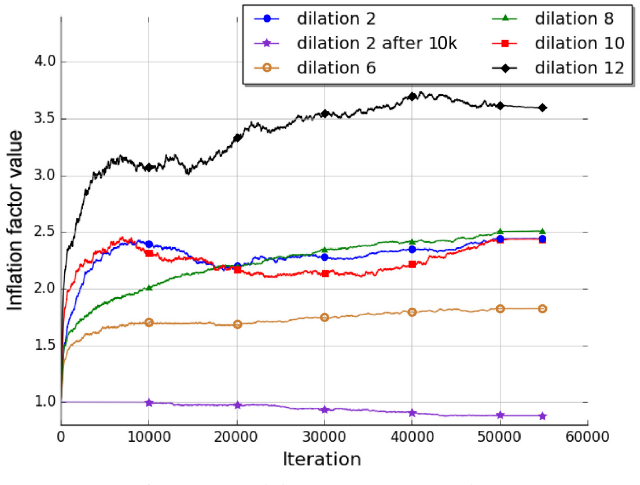


Fig. 4 The fluctuation of f during training in face parsing task.

5.2 Optimization for the general dataset

Unlike the face parsing task, in which images are coarsely aligned and semantic constituents from different images are of similar size (e.g., eyes, lips), object sizes in general datasets have much greater variability, making optimizing f rather more difficult. Even with proper initialization and identical network settings, while f stays in a certain range it does not converge to a specific value (see Fig. 7). Results shown in Table 3 are typical examples.

6 Conclusions

In this paper, we have introduced a new automatic regulation method for receptive fields in deep image parsing networks. This data-driven approach is able to replace existing hand-crafted receptive field selection methods. It enables deep image parsing networks to obtain better receptive fields with finer granularity in a single training process. Experimental results

Table 5 Quantitative comparison of our method and other face parsing models on the face parsing task. Our method performs best

Model	F-score					
	eye	brows	nose	mouth	face	overall
Liu et al. [23]	0.770	0.640	0.843	0.742	0.886	0.738
Smith et al. [8]	0.785	0.722	0.922	0.857	0.882	0.804
Liu et al. [19]	0.768	0.713	0.909	0.841	0.910	0.847
Our method	0.847	0.786	0.937	0.915	0.915	**0.902**

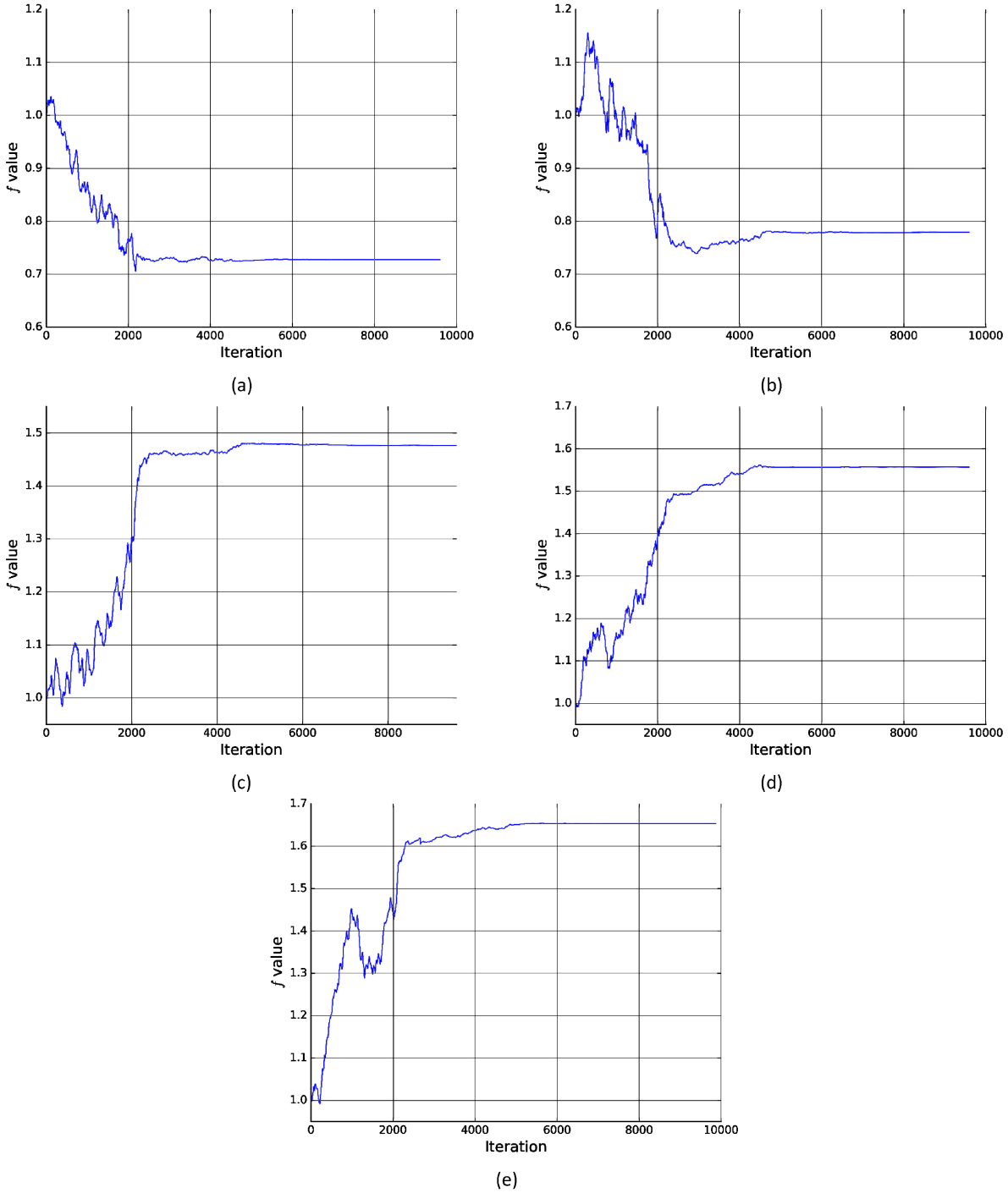


Fig. 5 Typical fluctuations in f during training for the general image parsing task. Modified models used initial dilation values of: (a) 4, (b) 6, (c) 16, (d) 18, (e) 20. Unlike the training process for the face parsing task, f shows more noticeable fluctuation due to high data variability in the VOC dataset.

Fig. 6 Fluctuation of f during training in the general image parsing task, using the same initial network settings. Only changes in the first 2500 iterations are plotted here. The initial dilation value was 4, much smaller than the optimal value. In this case, f sometimes may become trapped in local minima and stay near 1. Small initial dilation values are to be avoided.

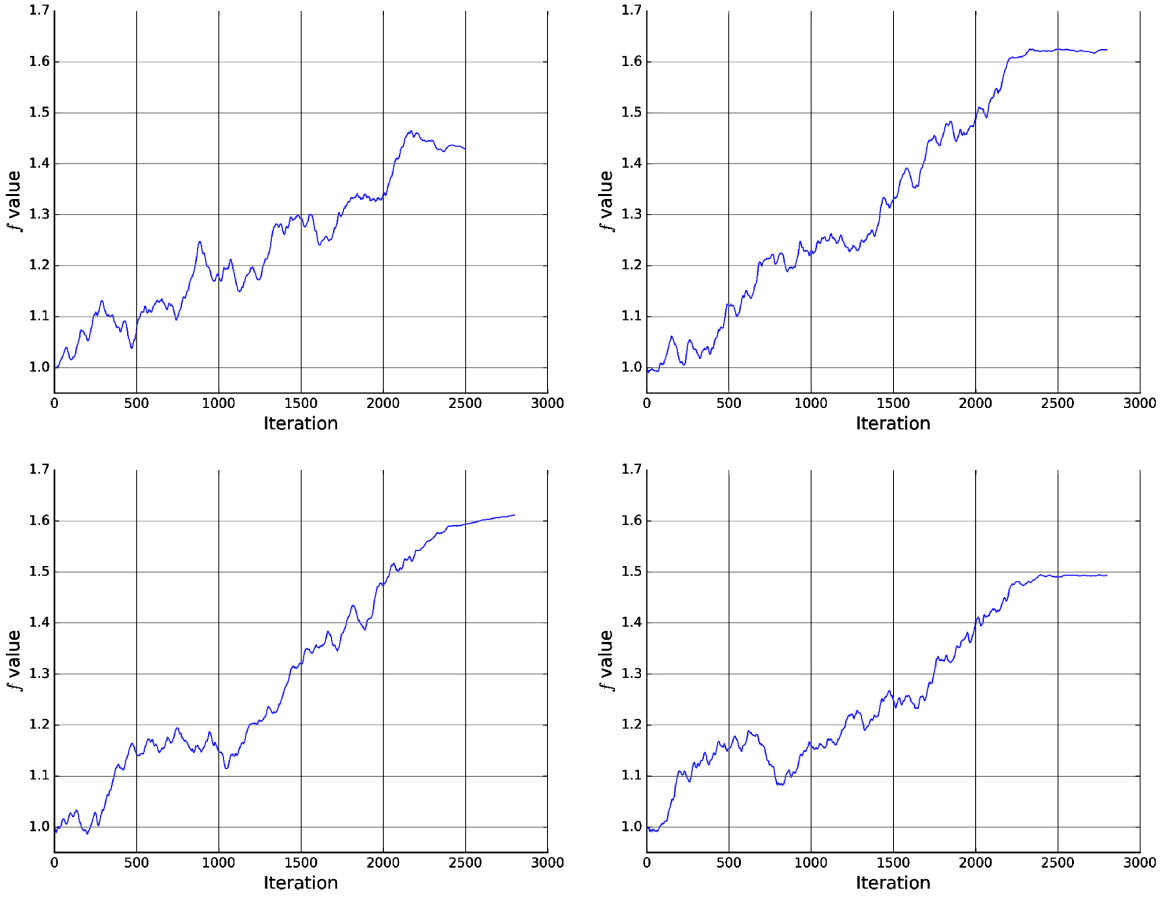


Fig. 7 Fluctuation of f during training for the general image parsing task, with identical initial network settings. Only changes in the first 3000 iterations are plotted here. The initial dilation value was 18. Due to the great variability during optimization, f will reach a range of values, instead of stopping at a specific number.

using the Helen and PASCAL VOC 2012 datasets demonstrate the effectiveness of our method in comparison to existing methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. U1536203, 61572493), the Cutting Edge Technology Research Program of the Institute of Information Engineering, CAS (No. Y7Z0241102), the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of the Ministry of Education (No. Y6Z0021102), and Nanjing University of Science and Technology (No. JYB201702).

References

[1] Long, J.; Zhang, N.; Darrell, T. Do convnets learn correspondence? In: Proceedings of the Advances in Neural Information Processing Systems 27, 1601–1609, 2014.

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

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

[4] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint* arXiv:1412.7062, 2014.

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

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

[7] Le, V.; Brandt, J.; Lin, Z.; Bourdev, L.; Huang, T. S. Interactive facial feature localization. In: *Computer Vision–ECCV 2012. Lecture Notes in Computer Science, Vol. 7574.* Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer, Berlin, Heidelberg, 679–692, 2012.

[8] Smith, B. M.; Zhang, L.; Brandt, J.; Lin, Z.; Yang, J. Exemplar-based face parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3484–3491, 2013.

[9] Wei, Z.; Sun, Y.; Wang, J.; Lai, H.; Lui, S. Learning adaptive receptive fields for deep image parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2434–2442, 2017.

[10] Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. DeepLab: Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint* arXiv:1606.00915, 2016.

[11] Mostajabi, M.; Yadollahpour, P.; Shakhnarovich, G. Feedforward semantic segmentation with zoom-out features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3376–3385, 2015.

[12] Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In: Proceedings of the Advances in Neural Information Processing Systems 28, 2017–2025, 2015.

[13] Chen, D.; Hua, G.; Wen, F.; Sun, J. Supervised transformer network for efficient face detection. In: *Computer Vision–ECCV 2016. Lecture Notes in Computer Science, Vol. 9909.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer, Cham, 122–138, 2016.

[14] Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, 764–773, 2017.

[15] Zheng, S.; Jayasumana, S.; Romera-Paredes, B.; Vineet, V.; Su, Z.; Du, D.; Huang, C.; Torr, P. H. S. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, 1529–1537, 2015.

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

[17] Zhang, R.; Isola, P.; Efros, A. A. Colorful image colorization. In: *Computer Vision–ECCV 2016. Lecture Notes in Computer Science, Vol. 9907.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer, Cham, 649–666, 2016.

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

[19] Liu, S.; Yang, J.; Huang, C.; Yang, M.-H. Multi-objective convolutional learning for face labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3451–3459, 2015.

[20] Sun, Y.; Wang, X.; Tang, X. Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3476–3483, 2013.

[21] Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; Zisserman, A. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision* Vol. 88, No. 2, 303–338, 2010.

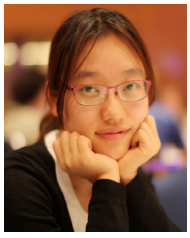
[22] Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Simultaneous detection and segmentation. In: *Computer Vision–ECCV 2014. Lecture Notes in Computer Science, Vol. 8695.* Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer, Cham, 297–312, 2014.

[23] Liu, C.; Yuen, J.; Torralba, A. Nonparametric scene parsing via label transfer. *IEEE Transaction on Pattern Analysis and Machine Intelligence* Vol. 33, No. 12, 2368–2382, 2011.

Zhen Wei received his B.S. degree in computer science and technology from Yingcai Honors School, the University of Electronic Science and Technology of China, Chengdu, China. He is now a master student in the Institute of Information Engineering, the Chinese Academy of Sciences.

Yao Sun is an associate professor in the Institute of Information Engineering, Chinese Academy of Sciences. He received his Ph.D. degree from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences.

Junyu Lin is assistant director of the Laboratory of Cyberspace Technology of the Institute of Information Engineering, Chinese Academy of Sciences. He is a member of the CCF YOCSEF academic committee and the CCF TCAPP standing committee. He is also the member of CCF council. He has more than 50 publications in *Peer to Peer Networking and Applications*, *the Journal of Software*, and IEEE conferences and journals.



Si Liu is an associate professor in the Institute of Information Engineering, Chinese Academy of Sciences. She was a research fellow in the Learning and Vision Research Group at National University of Singapore. She obtained her Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences. Her research interests include object categorization, object detection, image parsing, and human pose estimation.