Learning Adaptive Receptive Fields for Deep Image Parsing Network Supplementary File

Table 1. Specifications of network structures used in this paper, including the network backbone, single-path baseline model and single-path modified model.

<table>
<tr><th colspan="4">Network Backbone</th></tr>
<tr><td colspan="2">conv 1_1 conv 1_2</td><td colspan="2">output dim: 64, kernel size: 3, pad: 1</td></tr>
<tr><td colspan="2">pool 1</td><td colspan="2">MAX pooling, stride: 2, kernel size: 2, pad: 1</td></tr>
<tr><td colspan="2">conv 2_1 conv 2_2</td><td colspan="2">output dim: 128, kernel size: 3, pad: 1</td></tr>
<tr><td colspan="2">pool 2</td><td colspan="2">MAX pooling, stride: 2, kernel size: 2, pad: 1</td></tr>
<tr><td colspan="2">conv 3_1 conv 3_3</td><td colspan="2">output dim: 256, kernel size: 3, pad: 1</td></tr>
<tr><td colspan="2">pool 3</td><td colspan="2">MAX pooling, stride: 2, kernel size: 2, pad: 1</td></tr>
<tr><td colspan="2">conv 4_1 conv 4_3</td><td colspan="2">output dim: 512, kernel size: 3, pad: 1</td></tr>
<tr><td colspan="2">pool 4</td><td colspan="2">MAX pooling, stride: 1, kernel size: 2, pad: 1</td></tr>
<tr><td colspan="2">conv 5_1 conv 5_3</td><td colspan="2">output dim: 512, kernel size: 3, pad: 2, dilation: 2</td></tr>
<tr><td colspan="2">pool 5</td><td colspan="2">MAX pooling, stride: 1, kernel size: 3, pad: 1</td></tr>
<tr><th colspan="2"></th><th>Single-path Baseline Model</th><th>Single-path Modified Model</th></tr>
<tr><td colspan="2">batch norm</td><td>\</td><td>default parameters</td></tr>
<tr><td colspan="2">inflation layer</td><td>\</td><td>✓</td></tr>
<tr><td rowspan="3">conv 6</td><td>output dim</td><td colspan="2">1024</td></tr>
<tr><td>kernel size</td><td colspan="2">4 (Helen), 3 (VOC)</td></tr>
<tr><td>pad</td><td colspan="2">(dilation*(kernel size-1))/2</td></tr>
<tr><td rowspan="2">conv 7</td><td>output dim</td><td colspan="2">512 (Helen), 1024 (VOC)</td></tr>
<tr><td>kernel size</td><td colspan="2">1</td></tr>
<tr><td colspan="2">interpolation layer</td><td>\</td><td>✓</td></tr>
<tr><td>output layer</td><td>output dim</td><td colspan="2">11 (Helen), 21 (VOC)</td></tr>
</table>

Table 2. Quantitative evaluation results of baseline models and modified models on Helen [2, 5] 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 10000 iterations in training.

Single Path Baseline Model

network settings		F-score					
dilation	rf-fc6	eye	eyebrow	nose	mouth	face	ovearall
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	ovearall
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

Table 3. Quantitative evaluation results of multi-paths versions of baseline models and modified models on Helen dataset [2, 5]. Each parallels in the modified network is initialized with dilation value of 8.

Multi-paths Baseline Model

network settings			F-score					
model	dilation	rf-fc6	eye	eyebrow	nose	mouth	face	ovearall
bipath	4,6	308,356	0.8368	0.7757	0.9309	0.9104	0.9423	0.8964
tripath	4,6,8	308,356,404	0.8315	0.7638	0.9257	0.9044	0.9402	0.8894

Multi-paths Modified Model

model	f	rf-fc6	eye	eyebrow	nose	mouth	face	ovearall
bipath	3.32,1.27	268,372	0.8401	0.7888	0.9316	0.9129	0.9418	0.9008
tripath	1.61, 1.12, 1.11	340, 396, 396	0.8413	0.7763	0.9365	0.9098	0.9430	0.8983

Table 4. Quantitative evaluation results of our method and other face parsing models. Our method has achieved state-of-the-art performance on face parsing task.

Model	F-score					
	eye	brows	nose	mouth	face	overall
Liu et al.[3]	0.770	0.640	0.843	0.742	0.886	0.738
Smith et al.[5]	0.785	0.722	0.922	0.857	0.882	0.804
Liu et al.[4]	0.768	0.713	0.909	0.841	0.910	0.847
Ours	0.8466	0.7861	0.9365	0.9148	0.9148	**0.9021**

Table 5. Quantitative evaluation results of baseline models and modified models on VOC 2012 validation set [1]. d means dilation values in fc6 layer. *'rf-fc6'* means the extent of receptive field in fc6 layer.

Single Path Baseline Model

network settings		IOU (%)																					
d	rf-fc6	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
4	276	88.7	67.4	24.4	75.3	54.9	64.1	79.8	72.6	76.1	27.7	65.6	47.1	71.5	61.7	63.1	71.7	41.5	65.3	46.2	73.7	49.1	61.310
6	308	89.1	68.4	25.3	76.6	56.7	68.7	83.0	74.0	78.6	30.1	70.6	50.6	74.5	67.1	66.1	73.0	44.9	70.7	49.1	76.9	50.9	64.040
8	340	89.2	69.2	25.5	77.0	58.2	67.7	84.3	74.9	79.7	31.5	72.0	53.9	75.8	67.6	67.5	73.6	47.1	72.6	50.8	78.5	52.6	65.200
10	372	89.2	69.2	25.2	76.7	57.8	70.7	84.2	75.5	79.8	31.1	73.4	56.0	76.3	68.6	68.0	73.7	44.8	72.9	51.4	78.4	54.2	65.580
12	404	89.1	68.3	25.0	76.1	57.2	69.6	84.4	76.3	79.8	31.0	73.6	57.0	76.2	67.9	67.8	73.5	46.8	73.9	52.0	77.8	53.3	65.540
14	436	88.8	67.4	25.1	74.8	55.1	68.3	83.3	75.6	79.9	29.9	71.1	56.1	75.1	66.8	67.3	73.2	46.2	72.8	50.3	76.7	54.5	64.680
16	468	88.7	66.5	25.0	74.1	54.3	67.8	83.6	75.3	80.1	29.3	69.3	56.7	75.9	65.9	66.2	73.1	45.5	71.6	50.2	77.2	51.6	64.190
18	500	88.7	66.2	24.5	74.2	53.8	67.1	83.5	74.9	79.8	29.1	69.1	55.1	75.3	64.8	65.4	72.4	45.3	70.7	51.0	76.6	53.6	63.860
20	532	88.7	67.6	24.6	72.7	52.9	68.0	83.6	74.5	79.4	28.4	67.3	53.6	74.4	62.9	65.0	72.2	45.2	69.9	50.2	76.7	53.4	63.393

Single Path Modified Model

init d	f	rf-fc6	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
4	0.73	332	90.2	70.1	24.3	76.5	59.9	68.1	81.7	76.9	78.4	32.0	61.8	55.2	73.7	61.8	66.2	72.4	48.1	69.3	50.7	73.9	64.2	64.536
6	0.76	364	89.6	69.9	24.5	75.8	58.0	69.1	83.6	76.8	79.8	32.1	65.1	56.9	73.1	63.6	68.1	73.4	47.6	68.1	49.6	77.4	64.6	65.080
16	1.46	396	89.1	66.4	23.9	79.4	54.9	72.7	85.2	74.3	81.3	31.8	72.9	59.3	76.1	68.2	64.4	73.1	51.3	71.6	52.7	80.0	58.1	66.030
18	1.56	404	89.9	70.2	25.3	76.3	61.8	73.3	85.9	79.2	81.4	33.9	73.9	61.4	75.9	68.4	68.4	75.5	50.8	72.3	56.8	78.8	64.3	67.780
20	1.61	420	89.6	67.8	23.5	78.7	59.8	75.6	82.2	78.8	80.9	31.8	71.1	59.7	73.3	67.8	68.7	74.5	50.9	70.3	54.6	78.1	59.5	66.530

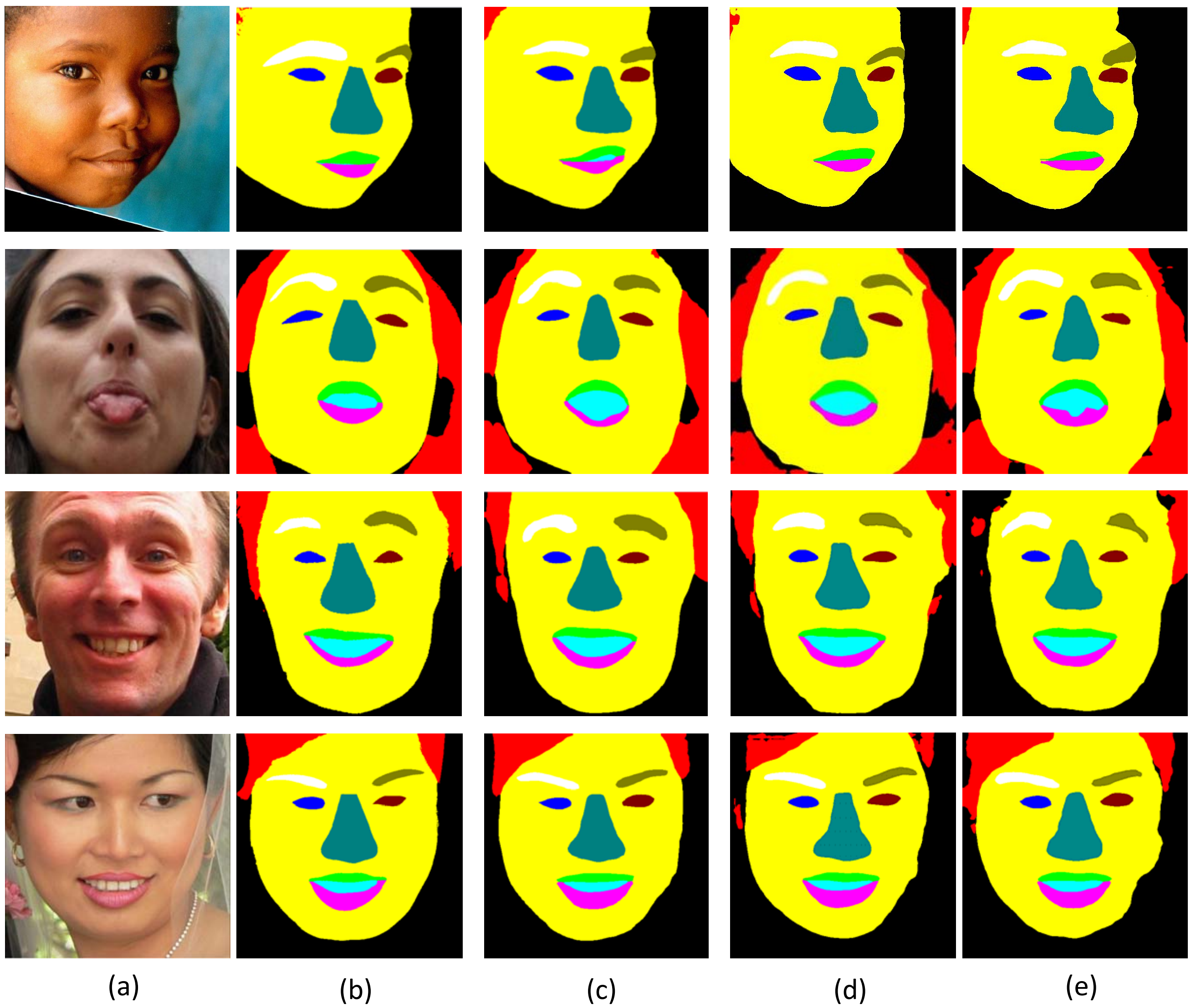

Figure 1. Face parsing results on Helen dataset [2, 5]. (a): original images. (b): ground truth. (c): results from baseline model with dilation value of 4 (with best manually selected receptive field). (d): results from modified model with initial dilation value of 12. (e): results from baseline model with dilation value of 12. Results in (d) and (e) show the improvements brought by our method. Smaller semantic areas have better parsing results, especially **eyebrows and nose**. **Face boundaries** are smoother and more accurate. Results in (c) and (d) show that our models have very close performance with manually designed models, which means our method can replace previous receptive field design process. Best view in color.

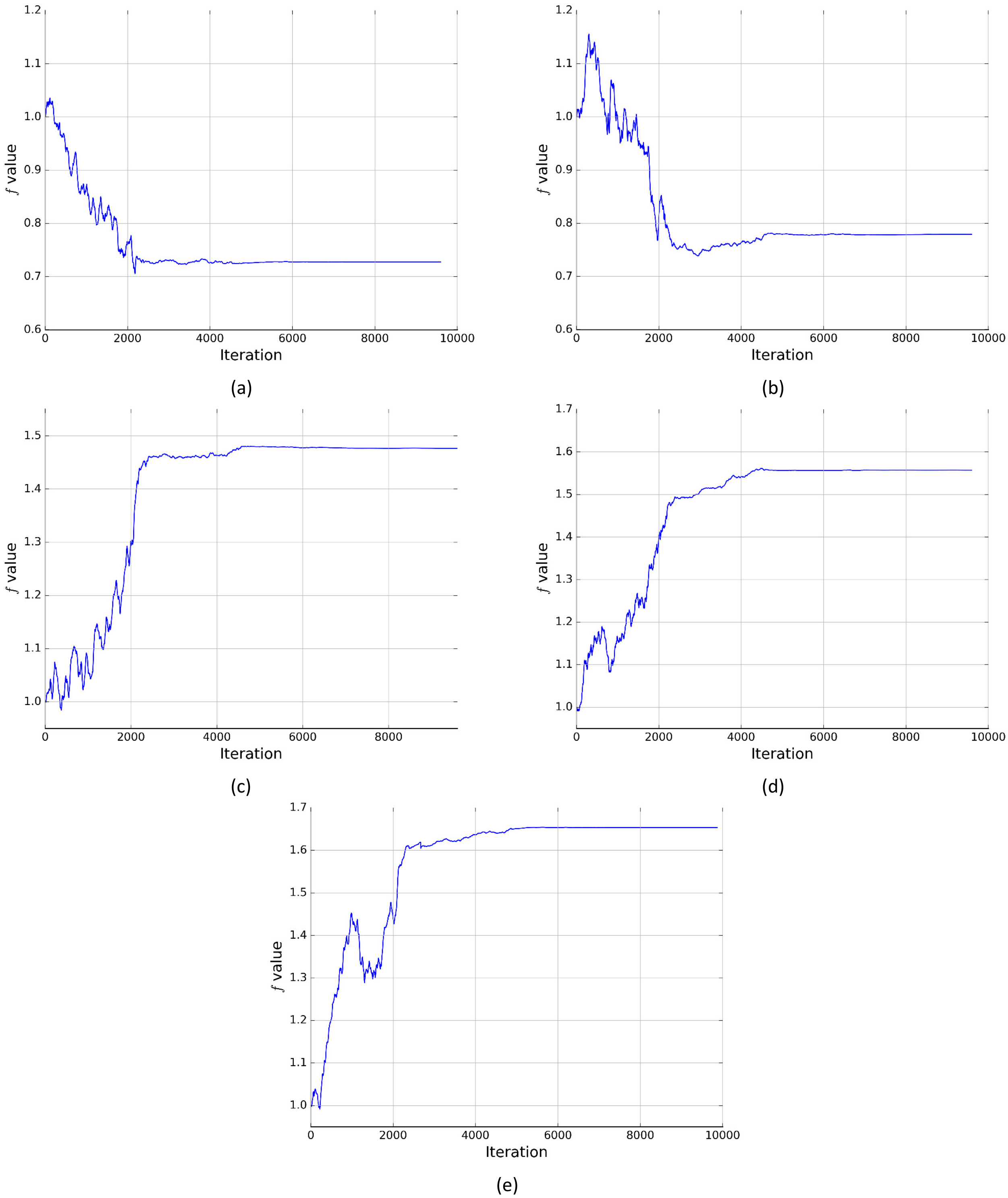

Figure 2. The typical fluctuation of f during training in general image parsing task. f come from the modified models with initial dilation values of: (a)4, (b)6, (c)16, (d)18, (e)20. Unlike the training process in face parsing task, f have more noticeable fluctuations due to great data variance on VOC dataset.

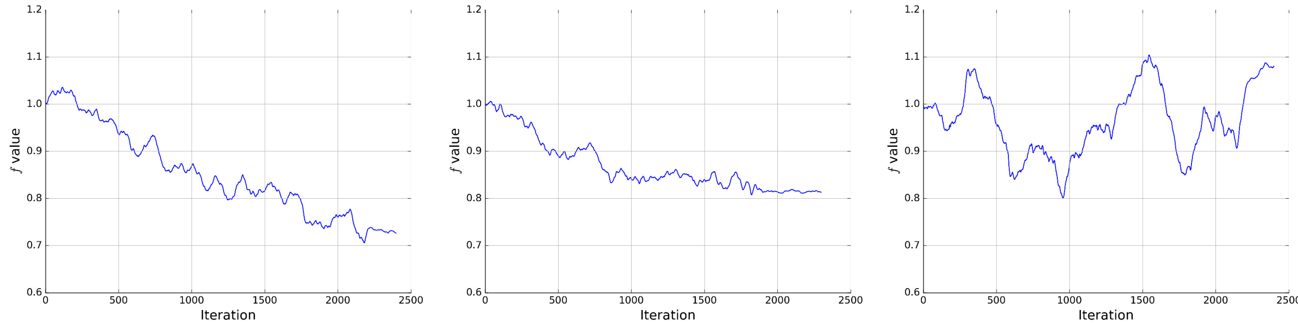

Figure 3. The fluctuation of f during training in general image parsing task with the same initial network settings. Only changes in the first 2,500 iterations are plotted here. The initial dilation value is 4, which is much smaller than the optimal value. In this case, f sometimes may trapped in local minimums and stay within the vicinity of 1. Small initial dilation values are not preferable.

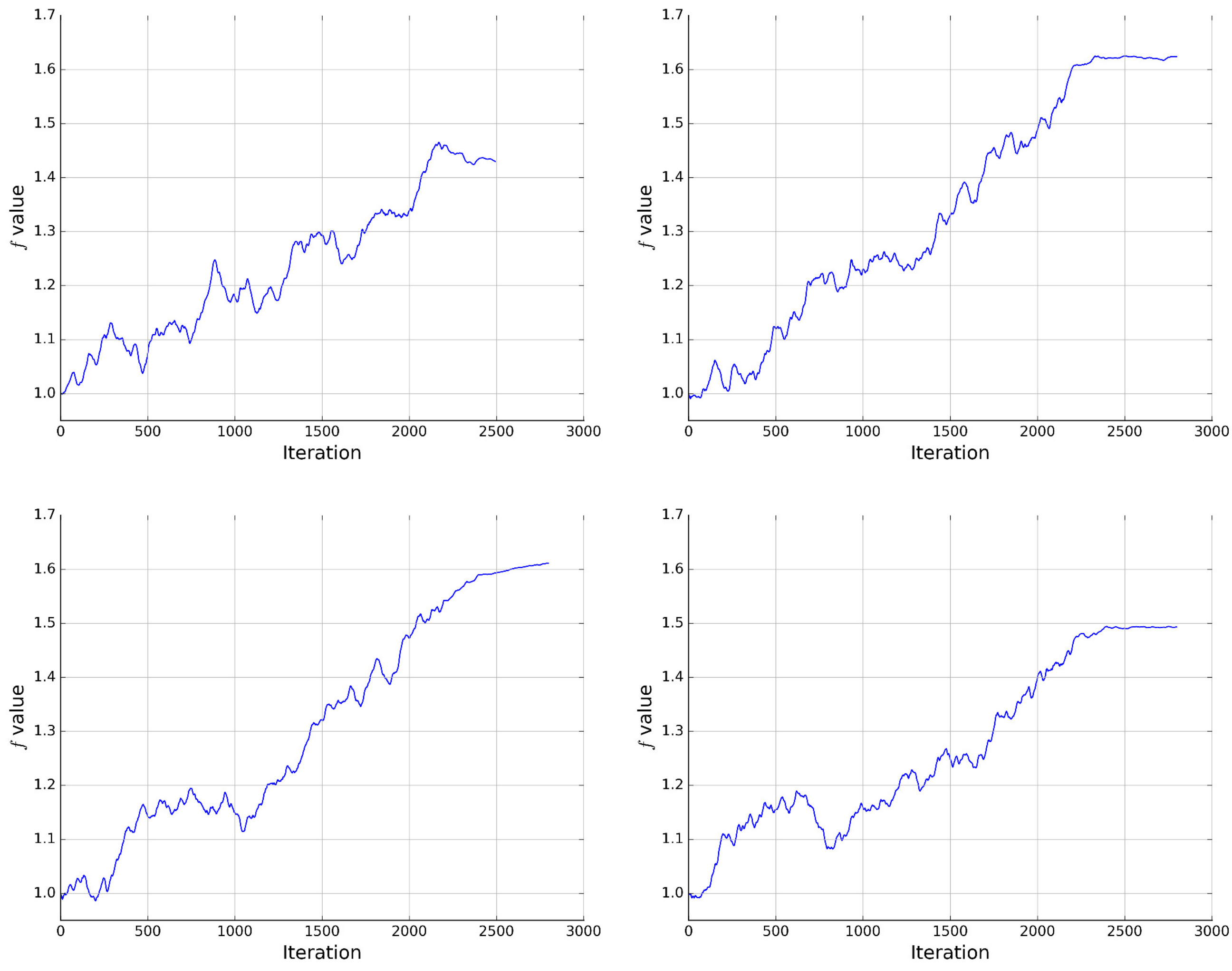

Figure 4. The fluctuation of f during training in general image parsing task with the same initial network settings. Only changes in the first 3,000 iterations are plotted here. The initial dilation value is 18. Due to the great variance during optimization, f will fall into a range of values, instead of stopping at a specific number.

References

[1] 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*, 88(2):303–338, June 2010. 3

[2] 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. 2, 4

[3] 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. 2

[4] 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. 2

[5] 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. 2, 4