Sparse Local Patch Transformer for Robust Face Alignment and Landmarks Inherent Relation Learning

Jiahao Xia[1], Weiwei Qu[2], Wenjian Huang[2], Jianguo Zhang[*,2], Xi Wang[3], Min Xu[*,1]
[1]Faculty of Engineering and IT, University of Technology Sydney
[2] Dept. of Comp. Sci. and Eng., Southern University of Science and Technology, [3]CalmCar
Jiahao.Xia@student.uts.edu.au, 11930667@mail.sustech.edu.cn, {huangwj, zhangjg}@sustech.edu.cn, Xi.Wang@calmcar.com, Min.Xu@uts.edu.au

A.1 Contructing Multi-scale Feature Maps for SLPT

As discussed in Section 4.3, we construct multi-level feature maps for ResNet34, as shown in Fig.1. Supposing the feature map size of k-th stage in ResNet34 is $W_k \times H_k \times d_k$, we firstly adopt a 1×1 CNN layer to reduce the channels from d_k to $C_I/4$. Then, the SLPT crops N patches whose size is $P_{Wk} \times P_{Hk}$ from each level and resizes these patches to $K \times K$. Note that $P_{Wk} \times P_{Hk}$ is $W_k/4 \times H_k/4$ in the initial coarse-to-fine stage and is reduced by half in each following stage. Finally, the resized patches from different levels are concatenated on the channel dimension which is C_I. As the result, the SLPT can utilize both high level and low level features for face alignment.

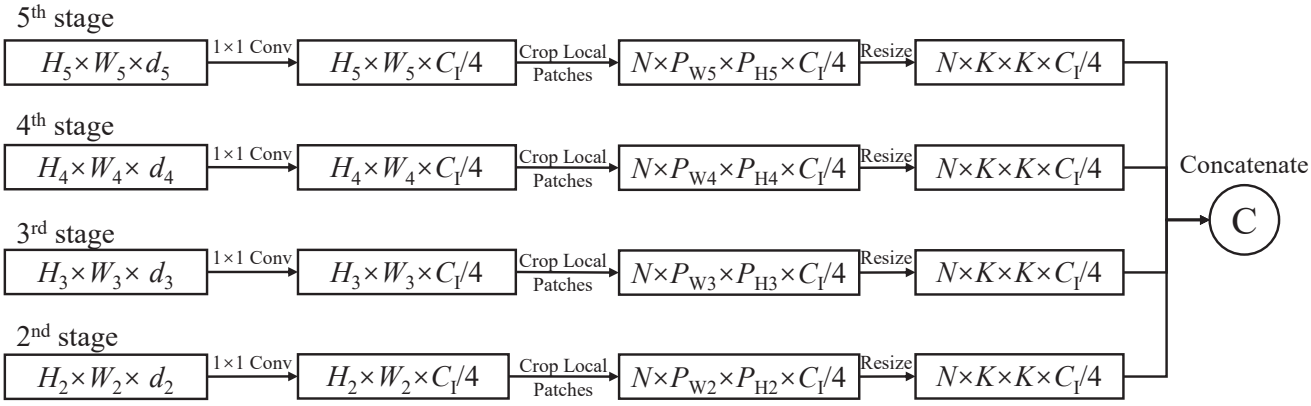

Figure 1. Constructing multi-level feature maps for SLPT

A.2 Details of comparison on WFLW

The comparison results on WFLW test set and its subsets are tabulated in Table 2. SLPT yields the best performance in NME and works at SOTA level on all subsets.

*Corresponding Author

Metric	Method	Testset	Pose	Expression	Illumination	Make-up	Occlusion	Blur
NME(%)↓	LAB [10]	5.27	10.24	5.51	5.23	5.15	6.79	6.32
	SAN [2]	5.22	10.39	5.71	5.19	5.49	6.83	5.80
	Coord* [9]	4.76	8.48	4.98	4.65	4.84	5.83	5.49
	DETR† [1]	4.71	7.91	4.99	4.60	4.52	5.73	5.33
	Heatmap* [9]	4.60	7.94	4.85	4.55	4.29	5.44	5.42
	AVS + SAN [8]	4.39	8.42	4.68	4.24	4.37	5.60	4.86
	LUVLi [4]	4.37	7.56	4.77	4.30	4.33	5.29	4.94
	AWing [11]	4.36	7.38	4.58	4.32	4.27	5.19	4.96
	SDFL* [7]	4.35	7.42	4.63	4.29	4.22	5.19	5.08
	SDL* [6]	4.21	7.36	4.49	4.12	4.05	**4.98**	4.82
	HIH [5]	**4.18**	7.20	**4.19**	4.45	**3.97**	**5.00**	**4.81**
	ADNet [3]	**4.14**	**6.96**	**4.38**	4.09	4.05	5.06	**4.79**
	SLPT‡	4.20	**7.18**	4.52	**4.07**	4.17	5.01	4.85
	SLPT†	**4.14**	**6.96**	4.45	**4.05**	**4.00**	5.06	**4.79**
$FR_{0.1}$(%)↓	LAB	7.56	28.83	6.37	6.73	7.77	13.72	10.74
	SAN	6.32	27.91	7.01	4.87	6.31	11.28	6.60
	Coord*	5.04	23.31	4.14	3.87	5.83	9.78	7.37
	DETR†	5.00	21.16	5.73	4.44	4.85	9.78	6.08
	Heatmap*	4.64	23.01	3.50	4.72	2.43	8.29	6.34
	AVS + SAN	4.08	18.10	4.46	2.72	4.37	7.74	4.40
	LUVLi	3.12	15.95	3.18	**2.15**	3.40	6.39	**3.23**
	AWing	2.84	13.50	2.23	2.58	2.91	5.98	3.75
	SDFL*	**2.72**	12.88	**1.59**	2.58	2.43	**5.71**	3.62
	SDL*	3.04	15.95	2.86	2.72	**1.45**	**5.29**	4.01
	HIH	2.96	15.03	**1.59**	2.58	**1.46**	6.11	**3.49**
	ADNet	**2.72**	**12.72**	**2.15**	**2.44**	1.94	5.79	3.54
	SLPT‡	3.04	15.95	2.86	**1.86**	3.40	6.25	4.01
	SLPT†	**2.76**	**12.27**	2.23	**1.86**	3.40	5.98	3.88
$AUC_{0.1}$ ↑	LAB	0.532	0.235	0.495	0.543	0.539	0.449	0.463
	SAN	0.536	0.236	0.462	0.555	0.522	0.456	0.493
	Coord*	0.549	0.262	0.524	0.559	0.555	0.472	0.491
	DETR†	0.552	0.285	0.520	0.558	0.563	0.471	0.497
	Heatmap*	0.524	0.251	0.510	0.533	0.545	0.459	0.452
	AVS + SAN	0.591	0.311	0.549	**0.609**	0.581	0.516	**0.551**
	LUVLi	0.557	0.310	0.549	0.584	0.588	0.505	0.525
	AWing	0.572	0.312	0.515	0.578	0.572	0.502	0.512
	SDFL*	0.576	0.315	0.550	0.585	0.583	0.504	0.515
	SDL*	0.589	0.315	0.566	0.595	**0.604**	0.524	0.533
	HIH	**0.597**	0.342	**0.590**	**0.606**	**0.604**	**0.527**	**0.549**
	ADNet	**0.602**	**0.344**	0.523	0.580	0.601	**0.530**	0.548
	SLPT‡	0.588	0.327	0.563	0.596	0.595	0.514	0.528
	SLPT†	0.595	**0.348**	**0.574**	0.601	**0.605**	0.515	0.535

Table 1. Performance comparison of the SLPT and the state-of-the-art methods on WFLW and its subsets. The normalization factor is inter-ocular and the threshold for FR is set to 0.1. Key: [**Best**, **Second Best**, *=HRNetW18C, †=HRNetW18C-lite, ‡=ResNet34]

A.3 Convergence curves of SLPT and DETR

The convergence curves of SLPT and DETR is shown in Fig.2. The DETR achieves 4.71% NME at 391 epochs on WFLW test set. The SLPT achieves better performance with around 8× less training epochs. With the increasing of training epochs, the performance of SLPT is improved further, achieving 4.14% NME at 140 epochs.

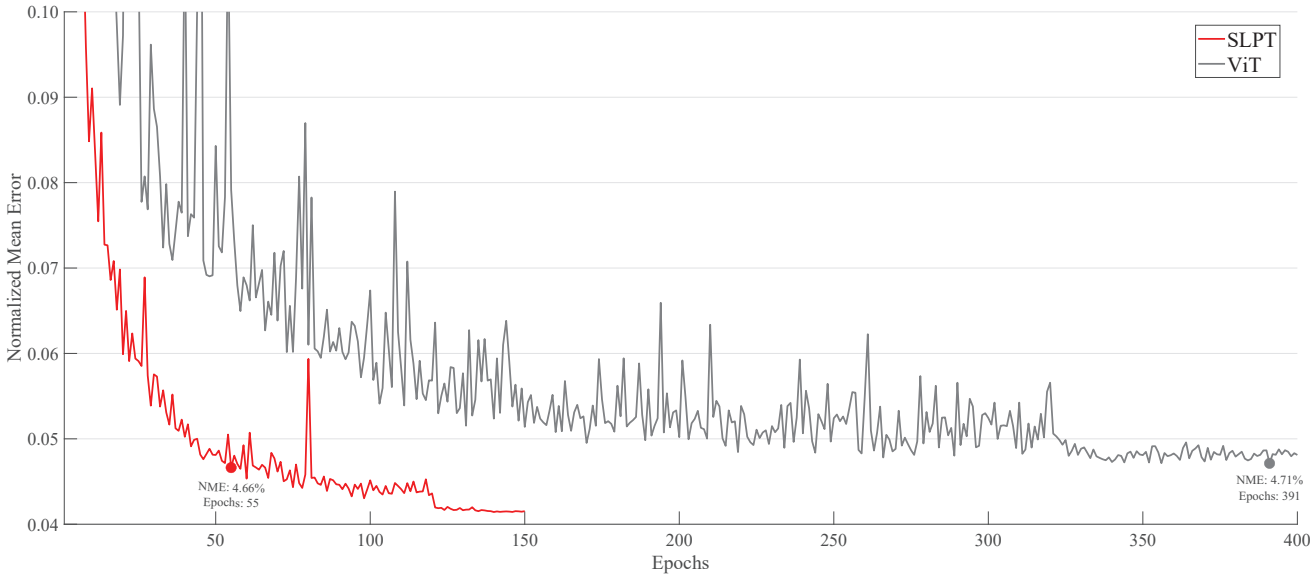

Figure 2. Convergence curves of SLPT and DETR on WFLW test set. The learning rate of SLPT is reduced at 120 and 140 epochs; the learning rate of DETR is reduced at 320 and 360 epochs.

A.4 Evaluation on the input patch size

Each local patch is resized to $K \times K$ and then projected into a vector by a CNN layer with $K \times K$ kernel size. In this section, we explore the influence of the patch size on WFLW test set, as tabulated in Table 2. Compared to 7×7 patches, the 5×5 patches lose more information because of the lower resolution, which leads to degradation of the performance. When the patch size is extended from 7×7 to 9×9, the parameters of the CNN layer is doubled, which leads to the overfitting on the training set. Therefore, we can also observe a slight degradation with 9×9 patch size, from 4.14% to 4.16% in NME.

Patch size	NME(%)	$FR_{0.1}$(%)	$AUC_{0.1}$
5×5	4.17%	**2.76%**	0.593
7×7	**4.14%**	**2.76%**	**0.595**
9×9	4.16%	2.84%	0.594

Table 2. NME(↓), $FR_{0.1}$(↓) and $AUC_{0.1}$(↑) with different patch sizes $K \times K$ on WFLW test set. Key: [**Best**]

A.5 Evaluation on the number of inherent relation layers

Table 3 demonstrates the influence of inherent relation layer number. The performance of SLPT relies on the inherent relation layer heavily. When the number of inherent relation layers increases from 2 to 12, We can observe a significant improvement, from 4.19% to 4.12% in NME. Nevertheless, too many inherent relation layers also increase the parameters and computational complexity dramatically. Considering the real-time capability, we choose the model with 6 inherent relation layers as the optimal model.

Layer number	NME(%)	$FR_{0.1}$(%)	$AUC_{0.1}$
2	4.19%	2.88%	0.592
4	4.17%	2.84%	0.593
6	4.14%	2.76%	0.595
12	**4.12%**	**2.72%**	**0.596**

Table 3. NME(↓), $FR_{0.1}$(↓) and $AUC_{0.1}$(↑) with different patch sizes $K \times K$ on WFLW test set. Key: [**Best**]

A.6 Further example predicted results and inherent relation maps

We visualize the predicted results and adaptive inherent relation maps for the samples of COFW, 300W and WFLW, as shown in Fig.3, Fig.4 and Fig.5 respectively. In the inherent relation maps, we connect each point to the point with highest cross-attention weight. The SLPT tends to utilize the visible landmarks to localize the landmarks with heavy occlusion for robust face alignment. For other landmark, it relies more on its neighboring landmark.

Figure 3. Further example predicted results and attention maps on COFW (random selection)

Figure 4. Further example predicted results and attention maps on 300W (random selection)

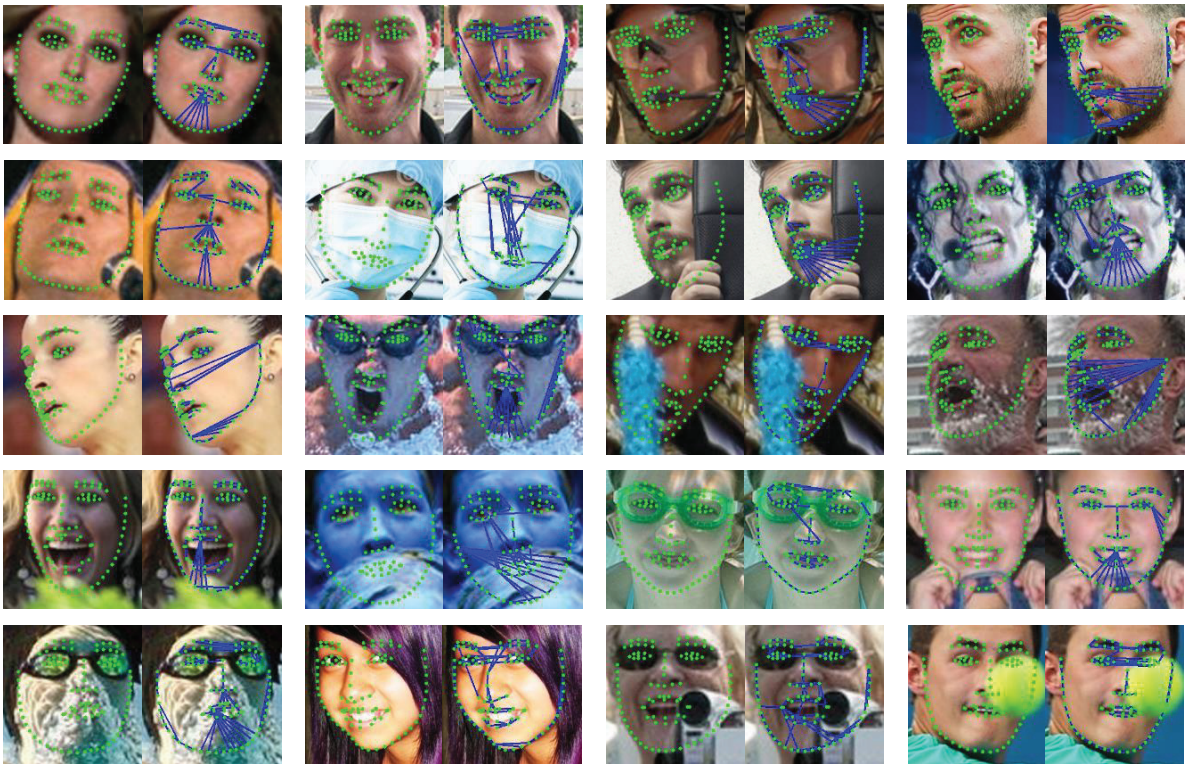

Figure 5. Further example predicted results and attention maps on WFLW (random selection)

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[2] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388, 2018.

[3] Yangyu Huang, Hao Yang, Chong Li, Jongyoo Kim, and Fangyun Wei. Adnet: Leveraging error-bias towards normal direction in face alignment. In *2021 ICCV*, pages 3060–3070, 2021.

[4] Abhinav Kumar, Tim K. Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pages 8233–8243, 2020.

[5] Xing Lan, Qinghao Hu, and Jian Cheng. Revisting quantization error in face alignment. In *2021 ICCVW*, pages 1521–1530, 2021.

[6] Weijian Li, Yuhang Lu, Kang Zheng, Haofu Liao, Chihung Lin, Jiebo Luo, Chi-Tung Cheng, Jing Xiao, Le Lu, Chang-Fu Kuo, and Shun Miao. Structured landmark detection via topology-adapting deep graph learning. In *ECCV 2020*, pages 266–283, Cham, 2020. Springer International Publishing.

[7] Chunze Lin, Beier Zhu, Quan Wang, Renjie Liao, Chen Qian, Jiwen Lu, and Jie Zhou. Structure-coherent deep feature learning for robust face alignment. *IEEE Transactions on Image Processing*, 30:5313–5326, 2021.

[8] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *ICCV*, pages 10152–10162, 2019.

[9] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2021.

[10] Wenyan Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138, 2018.

[11] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.