Supplementary Material of HP-Capsule: Unsupervised Face Part Discovery by Hierarchical Parsing Capsule Network

Chang Yu[1,2], Xiangyu Zhu[1,2], Xiaomei Zhang[1,2], Zidu Wang[1,2], Zhaoxiang Zhang[1,2,3], Zhen Lei[1,2,3*]
[1]NLPR, Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] Centre for Artificial Intelligence and Robotics, Hong Kong Institute of Science & Innovation,
Chinese Academy of Sciences
{chang.yu, xiangyu.zhu, zlei}@nlpr.ia.ac.cn
{zhangxiaomei2016, wangzidu2022, zhaoxiang.zhang}@ia.ac.cn

A. More Implementation Details

A.1. Training Details

We set the number of subpart capsules $K = 75$, the number of part capsules $M = 5$, $\gamma = 0.5$ for VAF, $\tau = 16$ for subpart presence sparsity loss. $\lambda_{cen}, \lambda_{cls}, \lambda_{silh}$ for loss combination are set to $0.5, 10^2, 10^3$, and other hyper-parameters are set to 1. The input images are resized to 128×128. The subpart templates are set to 40×40 and the part templates are set to 128×128.

A.2. Details of Template Transformation

During the subpart discovery, our Hierarchical Parsing Capsule Network (HP-Capsule) performs affine transformation with pose θ^s to transform each subpart template into the image space. In our implementation, $\theta^s = (s^s, h^s, a_x^s, a_y^s, t_x^s, t_y^s)$ is a 6-tuple, including s^s for scaling, h^s for shearing, (a_x^s, a_y^s) for rotation, and (t_x^s, t_x^s) for translation. The transformation matrix can be formulated as:

$$A = \begin{bmatrix} s^s cosa & -s^s sina + s^s h^s cosa & t_x^s \\ s^s sina & s^s cosa + s^s h^s sina & t_y^s \\ 0 & 0 & 1 \end{bmatrix} \quad (1)$$

where $(cosa, sina) = (a_x^s, a_y^s)/\|(a_x^s, a_y^s)\|_2$. We use a_x^s and a_y^s to estimate the rotation angle to avoid the continuity issue [2, 9].

A.3. More Details about $\mathrm{NME_{DL}}$

We propose a new evaluation metric $\mathrm{NME_{DL}}$ in the main paper to evaluate the unsupervised segmentation results on a detailed level, which uses a very shallow network to directly predict landmarks from the segmentation maps. The shallow network is trained using Adam with 10^{-4} learning rate for 50 epochs and all the input segmentation maps are resized to 32×32. During the evaluation, we choose three different shallow networks for $\mathrm{NME_{DL}}$, including two convolution layers, one Residual Block [4], and two Residual Blocks. Each of them is followed by a linear layer. Table 1 shows the sophisticated evaluation by $\mathrm{NME_{DL}}$ with different architectures on BP4D. It can be seen that our method surpasses other methods with better semantic consistency.

Table 1. The quantitative comparison of unsupervised face segmentation on BP4D. $\mathrm{NME_{DL}}(\%)$ is implemented with different architectures to evaluate the semantic consistency of landmarks.

Method	2 Conv	1 ResBlock	2 ResBlocks
DFF [1]	14.53	12.26	12.28
SCOPS [5]	7.91	6.74	6.94
HP-Capsule	**6.83**	**6.10**	**6.26**

B. More Analysis on the Exploited Face Hierarchy

To further explore the visual perception mechanism of neural networks, we visualize the learning procedure of HP-Capsule in Figure 1. It can be seen that the network captures the facial features in the sequences of face contour, mouth, nose, and finally eyes. One reasonable guess is that nose and eyes contain more identity-related information, making them more difficult to be reconstructed. This conclusion is also consistent with the work of Williford *et al.* [7] that shows nose and eyes contain more discriminative features

[*] Corresponding author.

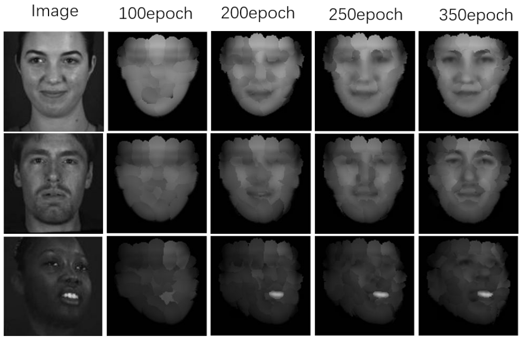

Figure 1. The learning procedure of HP-Capsule. The network captures the facial features in the sequence of face contour, mouth, nose, and finally eyes, which indicates the nose and eyes might contain more identity-specific information.

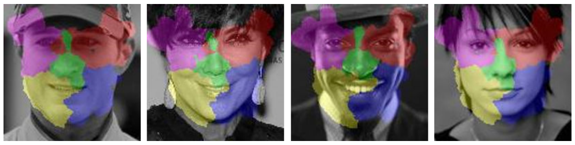

Figure 2. The segmentation results of HP-Capsule on CelebA.

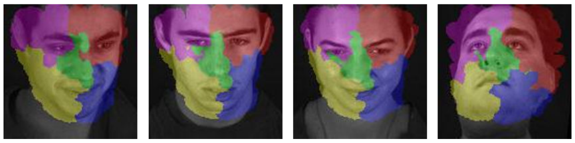

Figure 3. Some failed results of unsupervised face segmentation on BP4D. The parts of faces in large poses may still have flaws after refinement.

for face recognition.

C. More Visualization Results

To validate the potential of HP-Capsule under the in-the-wild scenarios, we evaluate our method on CelebA [6]. As shown in Figure 2, our method can also keep semantic consistency among the in-the-wild images.

We provide more visualization results of our HP-Capsule on BP4D [8, 10] and Multi-PIE [3]. Figure 4 shows the hierarchical face parts discovered by HP-Capsule. Figure 5 and Figure 6 show the unsupervised segmentation results on BP4D and Multi-PIE. It can be seen that our method can discover the face hierarchy directly from the unlabeled images and keep semantic consistency across different samples.

We also show some typical failure cases in Figure 3. Although the semantics of parts have been improved after introducing the Transformer-based Parsing Module for refinement, there may still be flaws on some faces in large poses.

References

[1] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–352, 2018. 1

[2] Qinzhe Gao, Bin Wang, Libin Liu, and Baoquan Chen. Unsupervised co-part segmentation through assembly. In *International Conference on Machine Learning*, 2021. 1

[3] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878, 2019. 1

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2

[7] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. Explainable face recognition. In *European Conference on Computer Vision*, pages 248–263. Springer, 2020. 1

[8] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. 2

[9] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 1

[10] Xiangyu Zhu, Fan Yang, Di Huang, Chang Yu, Hao Wang, Jianzhu Guo, Zhen Lei, and Stan Z Li. Beyond 3dmm space: Towards fine-grained 3d face reconstruction. In *European Conference on Computer Vision*, pages 343–358. Springer, 2020. 2

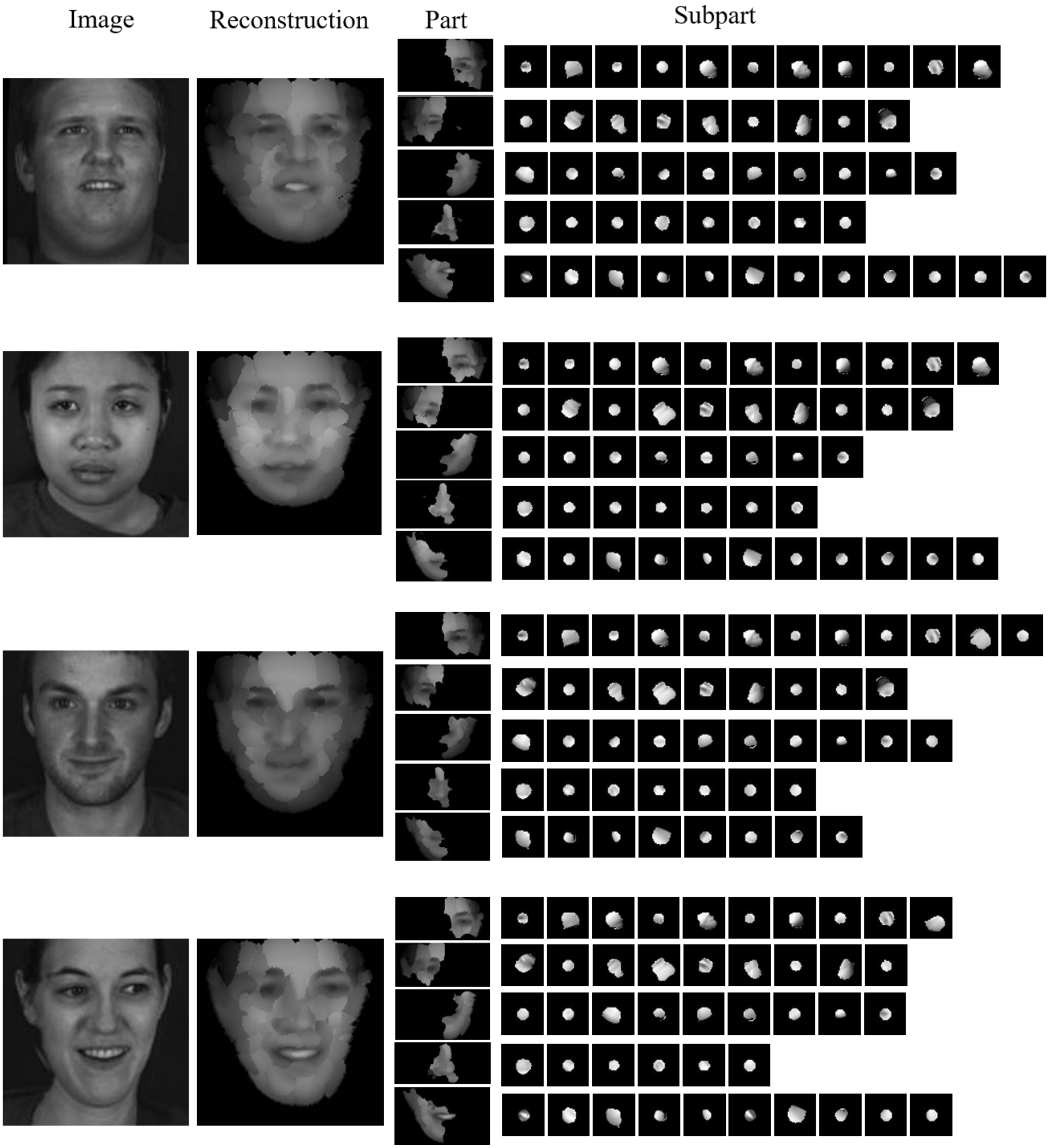

Figure 4. The hierarchical face parts discovered by HP-Capsule. For each input, HP-Capsule automatically selects a set of subparts to describe the current object and aggregates them to get parts with more prominent semantics.

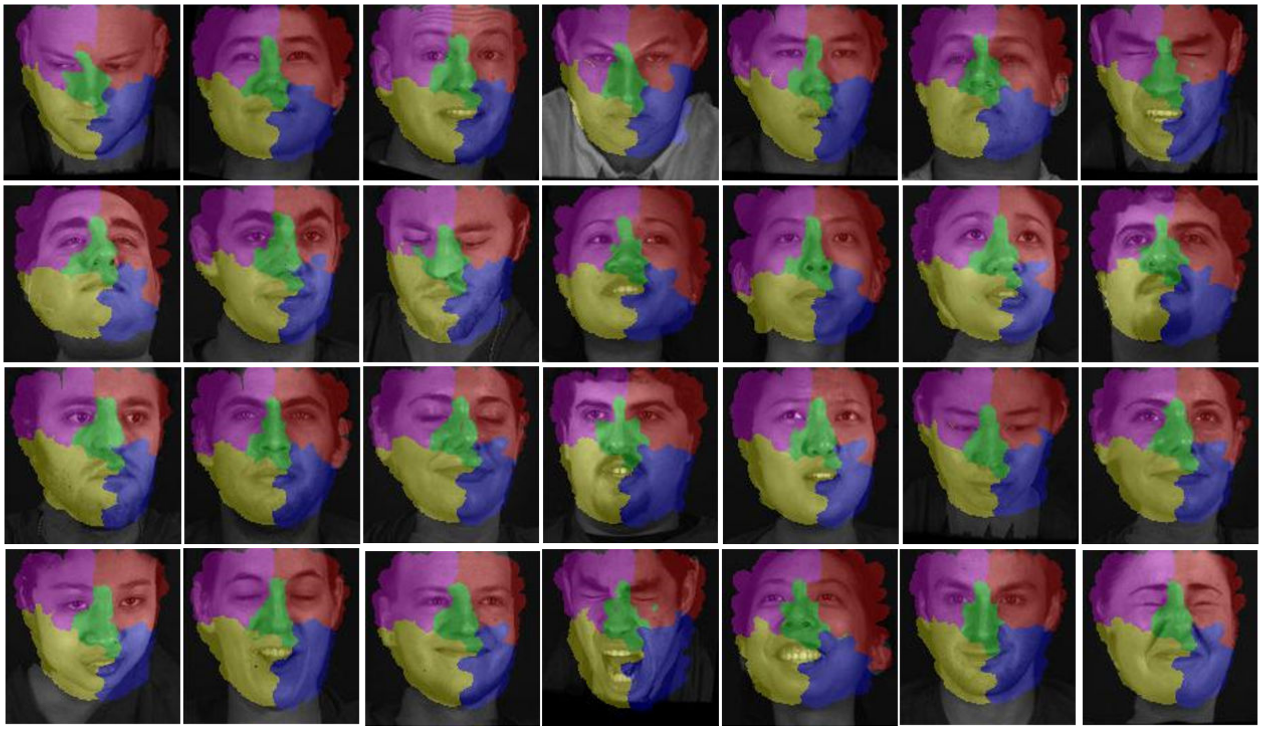

Figure 5. Visualization results of unsupervised face segmentation on BP4D.

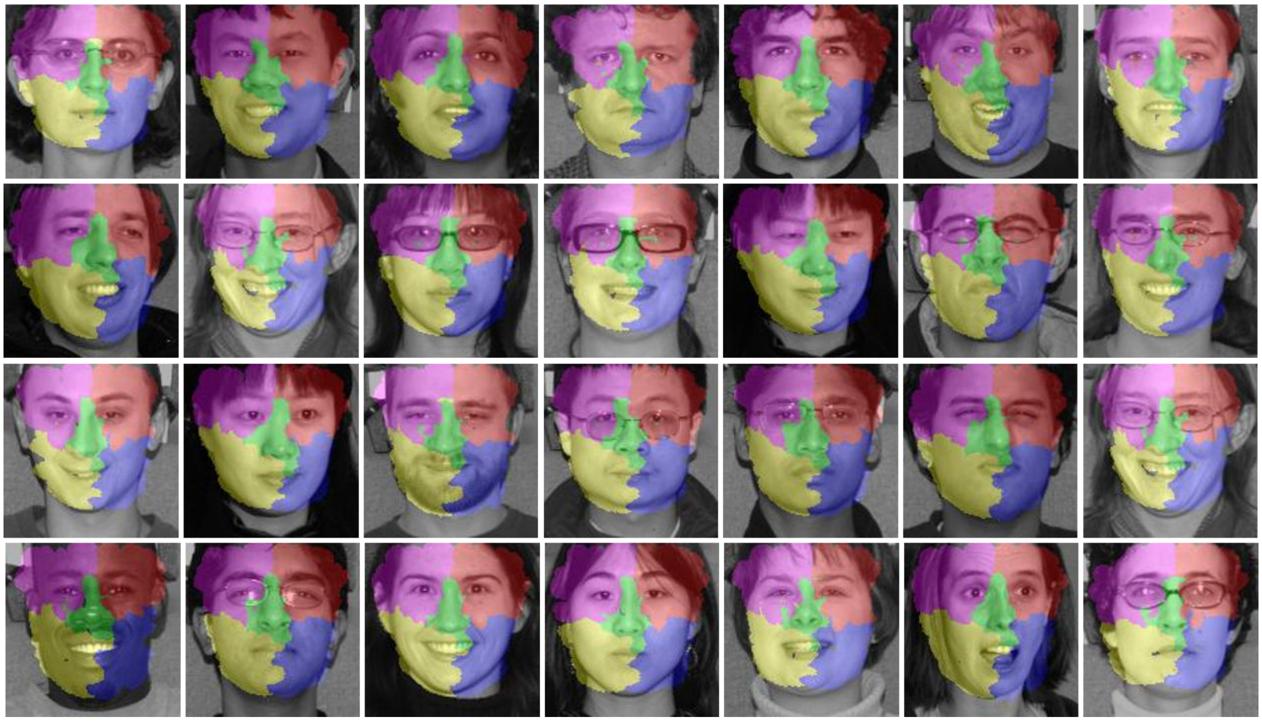

Figure 6. Visualization results of unsupervised face segmentation on Multi-PIE.