Supplementary Material of
Graphics Capsule: Learning Hierarchical 3D Face Representations from 2D Images

Chang Yu[1,2], Xiangyu Zhu[1,2]*, Xiaomei Zhang[1,2], Zhaoxiang Zhang[1,2,3], Zhen Lei[1,2,3]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems,
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, zhaoxiang.zhang}@ia.ac.cn

A. More Implementation Details

A.1. Training Details

Our IGC-Net is composed of an image encoder, a capsule decoder, and a lighting module. During training, all of the modules are trained together on the CelebA [2] dataset with the batch size $B = 64$. The input images are resized to 64×64. For background separation, the γ is set to be $0.25(D_{max} - D_{min}) + D_{min}$, where D_{max} is the maximum depth value of current face and D_{min} is the minimum depth value of the current face. With a GeForce RTX3090, the training procedure takes about 1.5 days.

A.2. Evaluation Details

For the analysis of hierarchical 3D face representations, we freeze the networks to extract the corresponding features and train a linear classifier for recognition. In Multi-PIE [1], 7 views ($0°, \pm15°, \pm45°, \pm60°$) of images are used for training and 2 views ($\pm30°$) of images are used for testing.

B. Comparison in Large Poses

Similar to our method, HP-Capsule [6] also aims to incorporate the capsule network to discover semantic facial parts from unlabeled images. However, their capsule parameters are defined in the 2D space, limiting their capacity to tackle the faces under large poses. In this section, we provide more unsupervised segmentation results of HP-capsule and our method, shown in Figure 1. It can be seen that HP-Capsule deteriorates seriously when faces rotate in yaw angles. On the contrary, our method shows semantic consistency across large poses due to the explicit 3D representations embedded in the capsules.

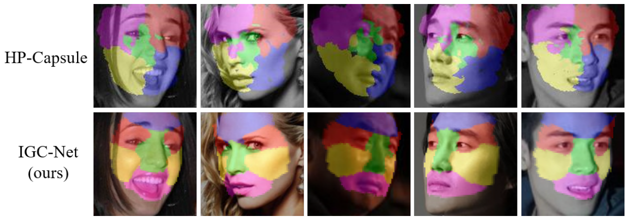

Figure 1. The comparison of HP-Capsule (2D) and our method under large poses (3D). It can be seen that the segment results of our method are with better semantic consistency under large poses due to the superiority of the 3D descriptions.

C. More Discovered Facial Parts

In IGC-Net, the hyper-parameter M controls the number of the part capsules. Specifically, one part capsule is used for background modeling and the $M - 1$ part capsules are used for facial parts. Figure 2 shows the learned parsing manners with different M. When $M = 4$, the discovered facial parts include the forehead and eyes, the nose and cheek, and the jaw. When $M = 5$, the faces are decomposed as: the forehead and nose, the eyes, the cheek, and the mouth. Each of them keeps semantic consistency across different samples.

*Corresponding author.

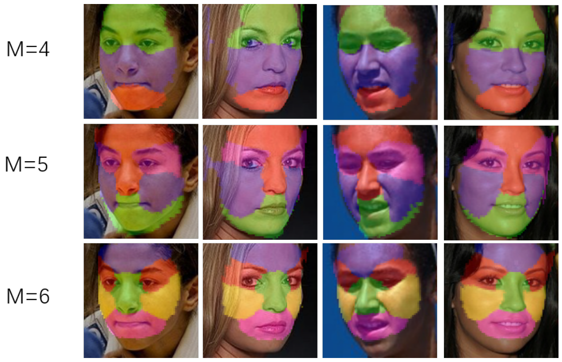

Figure 2. The discovered parts with different M. M is the hyper-parameter that controls the number of parts. It can be seen that, for different parsing manners, the eyes and nose are always inducted into different parts. Besides, our method performs satisfying semantic consistency across different samples.

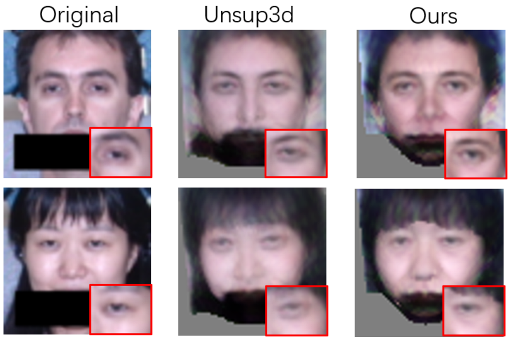

Figure 3. The reconstruction results of the mouth-occluded faces. It can be seen that the visible regions of our method are clearer and more identical to the original images, which validates that the hierarchical representations are more robust to the part-level disturbance than the global representations.

D. More Analysis about Hierarchical 3D Face Representation

Hierarchical vs. Global. Another suggestion for the vision mechanism is that the perception is hierarchical [3, 4]. The recognition procedure is related to the modular parts and their relationships so that the vision system is able to identify the parts and further recognize the objects when they are partially obscured. Under this circumstance, when the objects are partially occluded, the representations of these obscured parts are allowed to change, but the representations of the remained parts should keep their characteristics [4]. To evaluate the performance of our hierarchical representations in this situation, we compare our hierarchical representations with the global representations extracted by Unsup3d [5], which is an unsupervised method for 3D reconstruction.

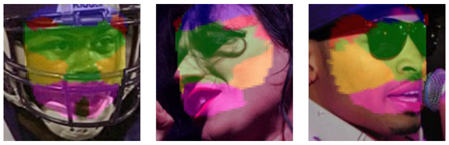

Figure 4. Failure cases. When the samples are large-area occluded or partially occluded but with large poses, the proposed method fails to decompose the faces properly.

We randomly covered the mouth regions (covering eyes makes both models collapse) of the faces in Multi-PIE and use the models trained on CelebA to recover the input images, shown in Figure 3. Firstly, we can see from the second column that parts are not independently encoded in the global representations. The reconstruction results of the eyes are disturbed by the occluded mouth, indicating that the injury of one part affects the other parts. Secondly, the third column shows better reconstruction results of our hierarchical representations, where the unobscured regions are clearer and more identical to the original images. The superiority comes from the disentangled descriptions of different parts, which enables each part to be free from the injury of other parts.

E. Failure Cases

Although the proposed IGC-Net can tackle the images with varied poses, there still exists some failure cases. As shown in Figure 4, when the samples are large-area occluded, or partially occluded but with large poses, IGC-Net fails to decompose them properly.

References

[1] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 1

[2] 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. 1

[3] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT press, 2010. 2

[4] Edmund T Rolls. Memory, attention, and decision-making: a unifying computational neuroscience approach. 2007. 2

[5] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2020. 2

[6] Chang Yu, Xiangyu Zhu, Xiaomei Zhang, Zidu Wang, Zhaoxiang Zhang, and Zhen Lei. Hp-capsule: Unsupervised face part discovery by hierarchical parsing capsule network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4032–4041, 2022. 1