3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation

Supplementary Material

1. More Analysis about PRDL

Ablation Study for f_{min}, f_{max}, and f_{ave}. In the main paper, we have extensively analyzed the gradient of PRDL in the case of f_{min}. In the supplementary material, we leverage the image-fitting framework [1] to further elucidate the roles of f_{min}, f_{max} and f_{ave} based on Part IoU benchmark. As depicted in Fig. 1, individually applying f_{min}, f_{max}, or f_{ave} yields satisfactory results, and their combined application leads to a significant improvement (63.61% in average). It should be noted that all results in Fig. 1 do not include $\mathcal{L}_{lmk}$, $\mathcal{L}_{pho}$, and $\mathcal{L}_{per}$.

More Gradient Analysis about PRDL. The above results indicate that the adoption of various distance measures is beneficial, which are also demonstrated in Fig. 2. We select a subset of v_n for gradient analysis. The effect of f_{max} is similar to that of f_{min}, with the only difference being the selection of points. f_{ave} influences the entire $V_{2d}^p(\boldsymbol{\alpha})$. As indicated by the green box in Fig. 2(b) and Fig. 2(c), the gradient of $\boldsymbol{\mathcal{F}} = \{f_{min}, f_{max}, f_{ave}\}$ is the most effective as it can correct some errors that arise when f_{ave} acts alone in certain scenarios. Aqua-colored arrows are used to indicate the direction of the gradient, with the length serving merely as an illustration. Reshape $\boldsymbol{\Gamma}_p^*$ ($\mathbb{R}^{|\boldsymbol{A}|\times|\boldsymbol{\mathcal{F}}|} \rightarrow \mathbb{R}^{H\times W\times|\boldsymbol{\mathcal{F}}|}$) to visualize its three channels ($f_{min}, f_{max}, f_{ave}$) separately in Fig. 2(d), Fig. 2(e), and Fig. 2(f). Compared to the regression target of the renderer-based Loss, $\boldsymbol{\Gamma}_p^*$ in PRDL is more informative and more conducive for fitting.

2. More Implementation Details

Transforming Segmentation to 2D Points. Two widely recognized definitions of 2D face segmentation regions are Helen [7] or iBugMask [10] and CelebAMask-HQ [8], which divide the face and related areas into 11 parts and 19 parts, respectively. As shown in Fig. 3, we employ the state-of-the-art method DML-CSR [15] for face segmentation. The results of the above two segmentation definitions are shown in Fig. 3(b) and Fig. 3(c), respectively. Through practical experimentation, we find that the 11-part method yields more accurate results. However, the segmentation of the ear regions from this method does not align well with the face model and needs to be removed. Consequently, we remove the corresponding ear regions from Fig. 3(b) based on Fig. 3(c), resulting in Fig. 3(d). Typically, Fig. 3(d) contains noise as indicated by the white dashed circle. To handle this, we identify the noise [2] and eliminate these isolated regions, yielding the outcome depicted in Fig. 3(e). To mitigate the impact of the region above the eyebrows, which is often obscured by hair, we transformed the eyebrows into 2D coordinates, identified their tangents (represented by white dashed lines in Fig. 3(e)), and dynamically removed the area above the eyebrows. The final result is presented in Fig. 3(f).

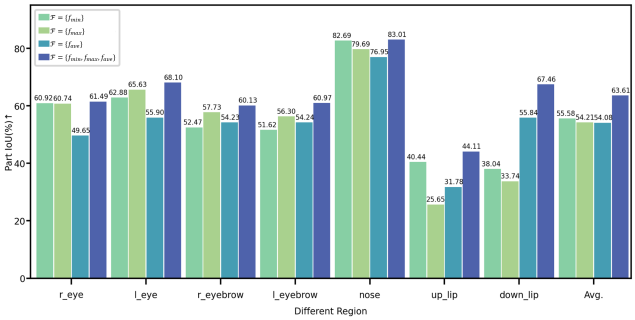

Figure 1. Quantitative comparison on Part IoU benchmark for f_{min}, f_{max}, and f_{ave}.

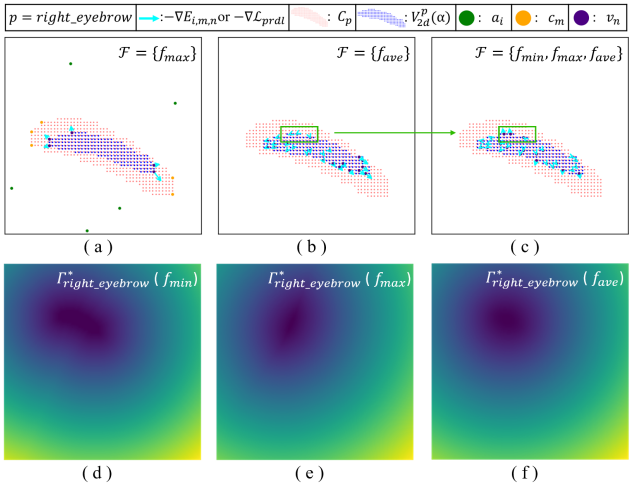

Figure 2. More analysis about PRDL when p = right_eyebrow. (a) Visualization of $-\nabla E_{i,m,n}$ when $\boldsymbol{\mathcal{F}} = \{f_{max}\}$. (b) and (c) depict the visualizations of $-\nabla \mathcal{L}_{prdl}$ when $\boldsymbol{\mathcal{F}} = \{f_{ave}\}$ and $\boldsymbol{\mathcal{F}} = \{f_{min}, f_{max}, f_{ave}\}$, respectively. (d), (e), and (f) visualize $\boldsymbol{\Gamma}_p^*$ in three channels (f_{min}, f_{max}, and f_{ave}).

3D Mesh Part Annotation. As shown in the Fig. 4, the objective of $\{Ind_p\}$ is to partition the specific face model to obtain $\{V_{2d}^p(\boldsymbol{\alpha})\}$ that are consistent with the region semantics of 2D segmentation. When $i \in Ind_p$, it means that the i-th vertex in the mesh belongs to part p.

Our 2D to 3D part mesh annotation method is described in Algorithm 1 with the following settings: $Render(\cdot)$

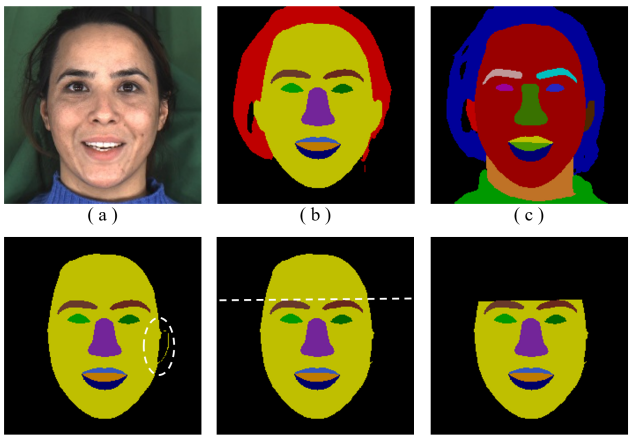

Figure 3. Remove the ear, filter noise and dynamically remove the forehead region according to the position of the eyebrows.

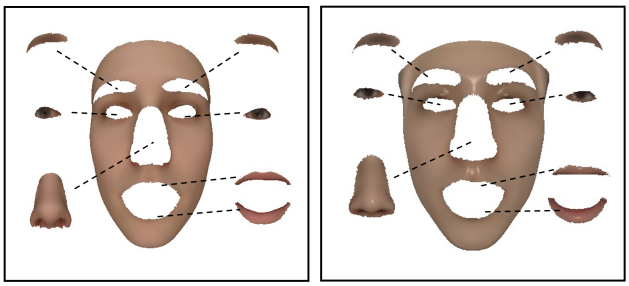

Figure 4. We provide 3D Mesh part annotations for the BFM [11] and FaceVerse [14] face models, which are well-aligned with the widely recognized 2D face segmentation definitions.

renders an image by employing texture on the mesh, and $Seg(\cdot)$ is responsible for segmenting the rendered result. Under the constraint of topological consistency within the same face model, ${V_{3d}}^{all}$ contains 3D face data with distinct poses and expressions, while Tex^{all} comprises diverse texture data. $\boldsymbol{P} = \{$left_eye, right_eye, left_eyebrow, right_eyebrow, up_lip,down_lip, nose, skin$\}$. In practice, if the segmentation resolution of the face parsing method is large enough, k could be equal to 1 in Algorithm 1. The few errant vertex indices in $\{Ind_p\}$ should be manually correct. The proposed algorithm 1 can also be applied to 2D to 3D landmark marching. To ensure consistency with the ground truth $\boldsymbol{C}_p$, the upper forehead region above the eyebrows is dynamically excluded, and the points obstructed by hair are also removed, as illustrated in Fig. 5.

Test Images for Part IoU. Multi-view Emotional Audio-visual Dataset (MEAD) [13] is a talking-face dataset corpus featuring 60 actors talking with 8 different emotions at three different intensity levels, which can provide high-quality details of facial expressions. We select 10 identities from MEAD, containing diversity across genders and ethnicity. We randomly select 50 different frontal images from each identity to constitute the Part IoU testing set. Fig. 6 shows a subset of these images.

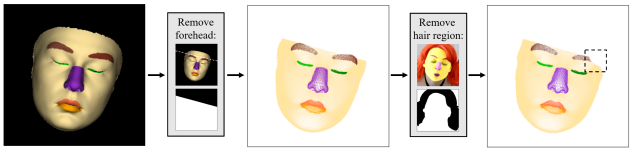

Figure 5. Remove the forehead region and the points obstructed by hair to ensure consistency with the ground truth $\{\boldsymbol{C}_p\}$.

Algorithm 1: Identify part indices $\{Ind_p\}$ of the mesh.

Input: $Render(\cdot)$, $Seg(\cdot)$, ${V_{3d}}^{all}$,Tex^{all}, $\boldsymbol{P}$
Init: $Ind_p = \varnothing$, k (k-nearest-neighbor)
for $\forall\, V_{3d} \in {V_{3d}}^{all}$ *and* $\forall\, Tex \in Tex^{all}$ **do**
  // Get the segmentation,
  $I^{seg} = Seg(Render(V_{3d}, Tex))$,
  // Transform I^{seg} to coordinates,
  $\{\boldsymbol{C}_p | p \in \boldsymbol{P}\} \leftarrow I^{seg}$,
  // Project V_{3d} to the image plane,
  $V_{2d} = Project(V_{3d})$
  for $p \in \boldsymbol{P}$ **do**
    for $\boldsymbol{c} \in \boldsymbol{C}_p$ **do**
      // $\boldsymbol{c}$ is a 2D coordinate,
      Find the first k vertices in V_{2d} that are closest to $\boldsymbol{c}$, and these k vertices should be visible, append their corresponding indices to Ind_p.
    end
  end
end
Output: $\{Ind_p\}$

3. More Comparison with the Other Methods

Fig. 9 depicts a more comparison between our results and the other state-of-the-art methods, *i.e.* PRNet [4], MGCNet [12], Deep3D [3], 3DDFA-V2 [6], HRN [9], and DECA [5]. Leveraging the advancements brought by PRDL, our method excel in capturing extreme facial expressions. Part IoU measures the overlap performance between each part of the reconstruction and the ground truth. The visualization of Part IoU for every method can be found in Fig. 7, which shows that PRDL enhances the alignment of reconstructed facial features with the original image.

4. More Results about Synthetic Data

Fig. 10 illustrates more results about our synthetic emotional expression dataset. The dataset currently consists of over $200K$ images, including synthetic expressions such as closed-eye, open-mouth, and frown. This dataset will be publicly available to facilitate the related research.

Figure 6. A subset of test images for Part IoU.

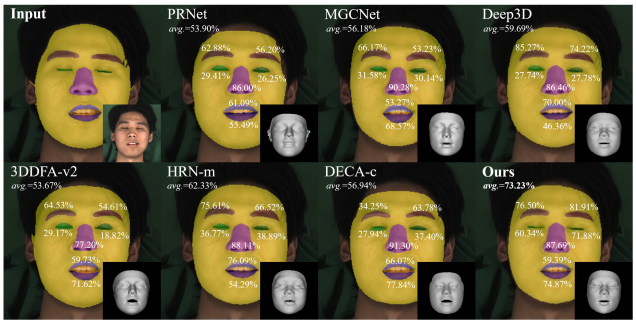

Figure 7. Comparison on Part IoU. The IoU value and visualizations for each reconstructed part are annotated, and the bottom right corner of each image is the corresponding 3D reconstruction.

5. Limitations

We summarize two limitations of our approach. Firstly, while Fig. 9 has demonstrated the excellent performance of our method on extreme facial expressions, it is constrained by the limited linear space of the 3DMM, resulting in some imperfections in reconstructing particularly challenging expressions. Secondly, although our method can handle occluded faces, it may struggle with severe occlusions, as illustrated in Fig. 8. In the future, we will extend our method to fine-grained face reconstruction and multi-view face reconstruction to address these limitations.

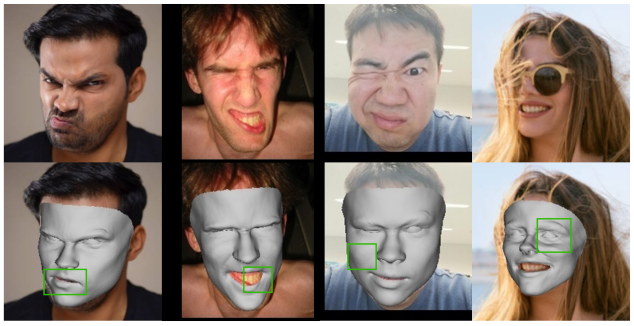

Figure 8. Limitations of our method. In cases of extremely challenging facial expressions or heavily occluded faces, our reconstructions may exhibit some minor errors.

References

[1] 3dmm model fitting using pytorch. https://github.com/ascust/3DMM-Fitting-Pytorch, 2021. 1

[2] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 1

[3] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2, 5

[4] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018. 2, 5

[5] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. 2021. 2, 5

[6] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. pages 152–168, 2020. 2, 5

[7] Vuong Le, Jonathan Brandt, Zhe L. Lin, Lubomir D. Bourdev, and Thomas S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, 2012. 1

[8] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[9] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 394–403, 2023. 2, 5

[10] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. *Image and Vision Computing*, 112:104190, 2021. 1

[11] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 sixth IEEE*

international conference on advanced video and signal based surveillance, pages 296–301. Ieee, 2009. 2

[12] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, pages 53–70. Springer, 2020. 2, 5

[13] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 2

[14] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20333–20342, 2022. 2

[15] Qi Zheng, Jiankang Deng, Zheng Zhu, Ying Li, and Stefanos Zafeiriou. Decoupled multi-task learning with cyclical self-regulation for face parsing. In *Computer Vision and Pattern Recognition*, 2022. 1

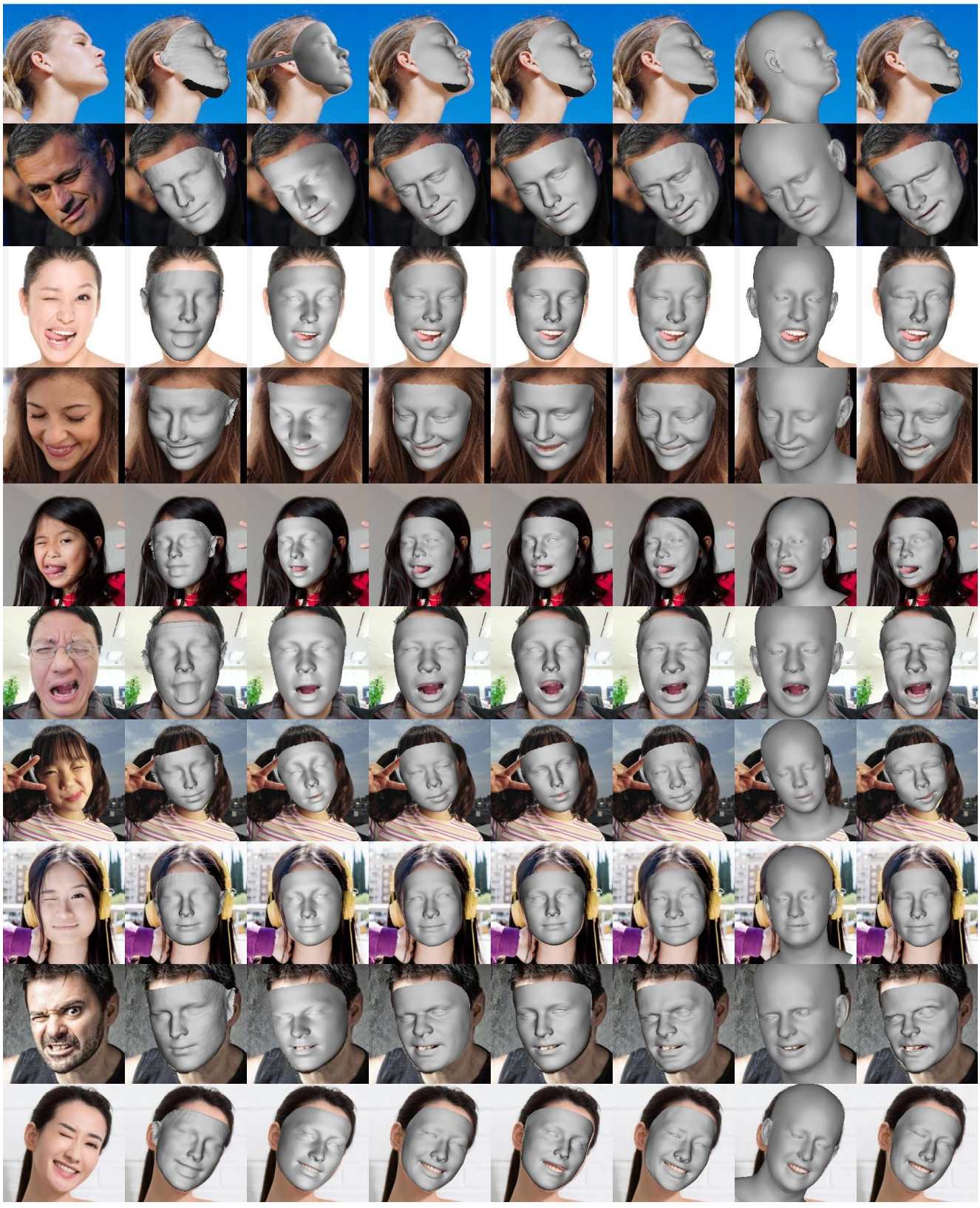

Figure 9. More comparison with the other methods. From left to right: Input image, PRNet [4], MGCNet [12], Deep3D [3], 3DDFA-V2 [6], HRN [9], DECA [5], and Ours.

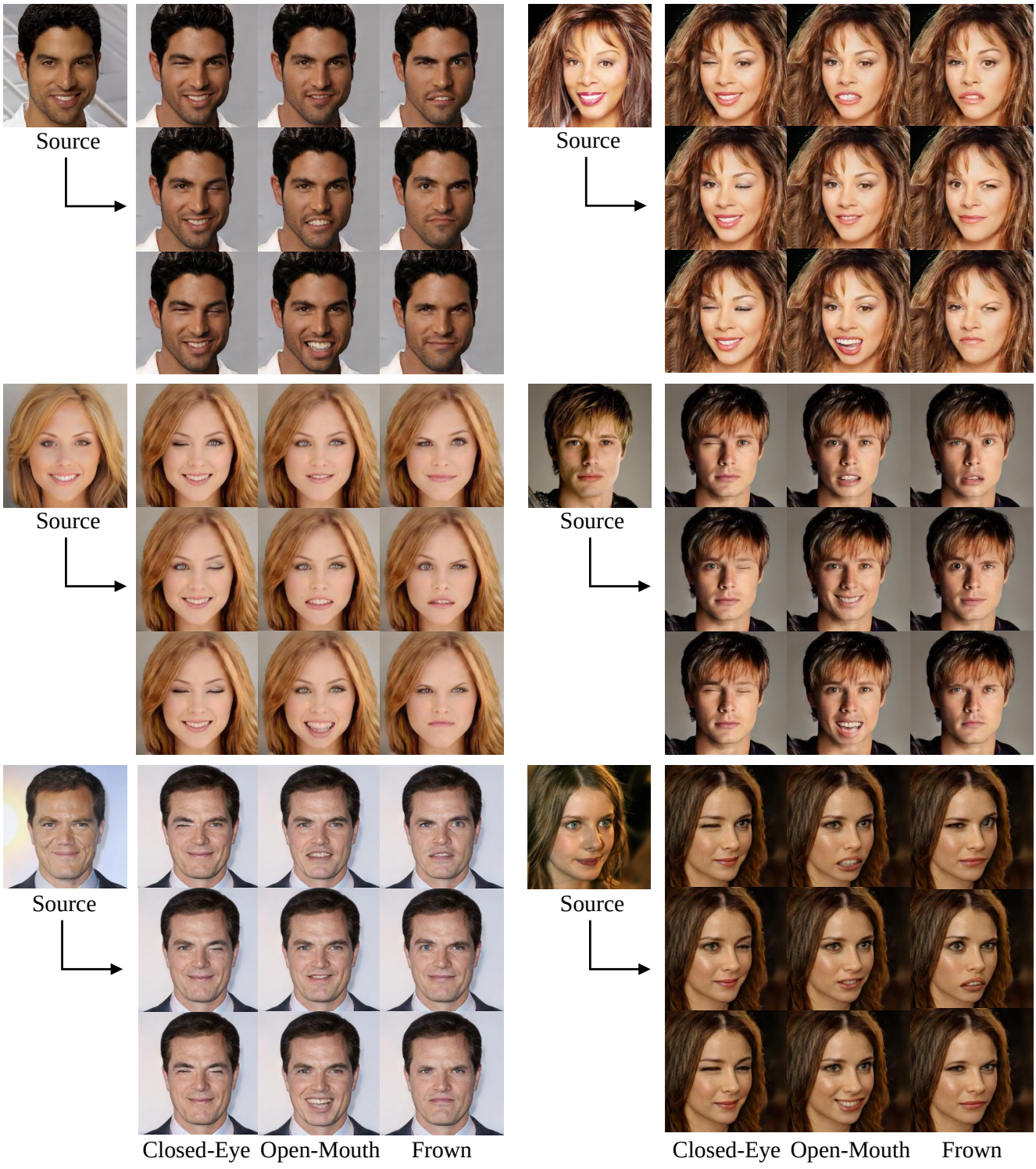

Figure 10. Examples of our synthetic face dataset.