LEVERAGING 3D GAUSSIAN SPLATTING TO ENHANCE FACE PARSING

Paper ID: 1569

Supplementary Material

This supplementary material provides additional information and experimental details that were not included in the main paper.

0.1. Dataset

We use the Facescape [1, 2] dataset for training and evaluating our approach. Among all available identities, only ten are included in the publicly accessible list. Out of them, eight have multi-view data suitable for constructing the 3D Gaussian Splatting [3] models. Six identities are used for training and two are held out for testing. To make the face-parsing model robust, we use a different expression for each identity. The identities used for training and testing are summarized in the following table:

Training Identities	**Expression**
122	$neutral$
340	$sadness$
344	$anger$
393	jaw_right
421	$lip\_puckerer$
527	$brow\_lower$
Testing Identities	**Expression**
395	$neutral$
610	$brow\_raiser$

Table 1. Training and testing identities with corresponding expressions.

We showcase all the training identities with few pose variations in Fig. 1. All the test identity images with their face-parsing results from fine-tuned model using our method are shown in Fig. 4.

0.2. Experimental setup

The multi-view images are used to generate $3DGS_{RGB}$ for each identity. These images are also processed using our baseline face-parsing model BiSeNet [4] to produce the baseline segmentation labels. The baseline labels, along with $3DGS_{RGB}$, are then utilized to create $3DGS_{SEG}$ for each identity.

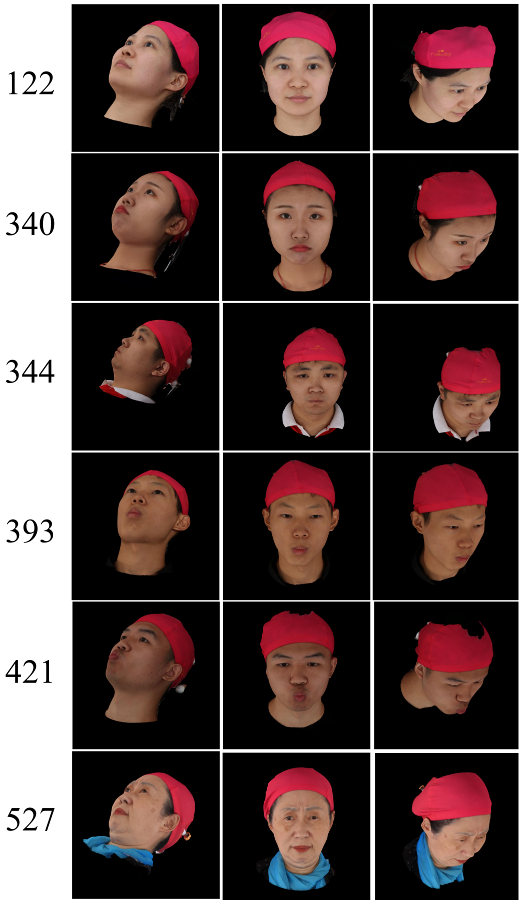

Fig. 1. Training identities displayed with selected pose variations.

For constructing $3DGS_{RGB}$, we employ the default settings from the 3D Gaussian Splatting implementation[1]. The following modifications are made to generate $3DGS_{SEG}$:

- Load $3DGS_{RGB}$ using the $start\_checkpoint$ command.
- Freeze all parameters except for color ($\_features\_dc$ as seen in official repository[1]) by setting them as non-trainable during initialization.
- Disable the densification step, as it is unnecessary when optimizing only for color.

[1] https://github.com/graphdeco-inria/gaussian-splatting

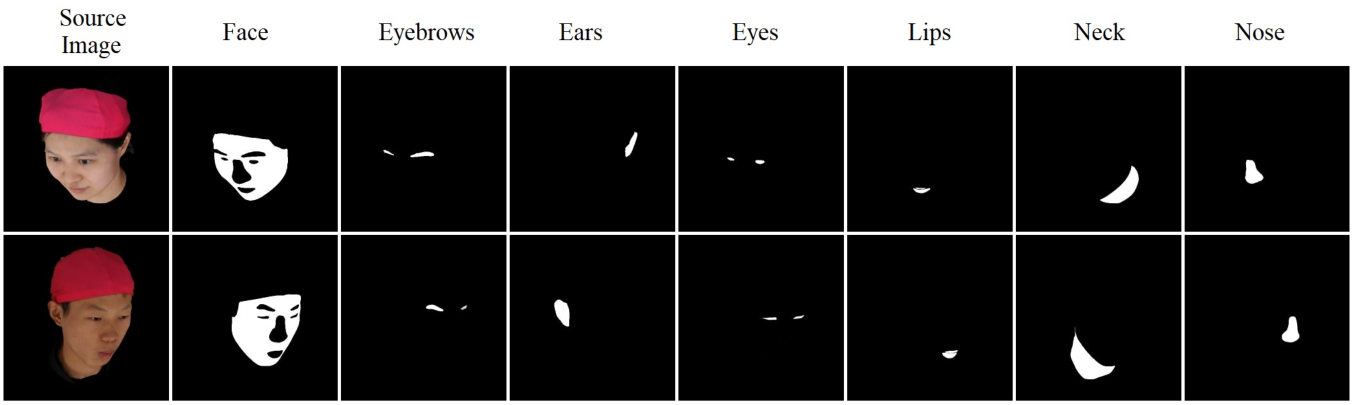

Fig. 2. Sample segmentation masks used for fine-tuning the baseline model on key facial features.

For sampling of the images, we utilized the SuperSplat[2] editor. Both $3DGS_{RGB}$ and $3DGS_{SEG}$ were sampled from the following viewpoints:

x	y	z
0	45	0
0	25	0
180	-65	180
180	-45	180
45	40	0
45	10	0
35	-10	0
30	-30	0
40	-45	0
30	-70	0
-170	-65	180
-150	-50	180
-45	30	0
-65	5	0
-45	-30	0
-45	-90	0
135	-65	180
135	-40	180
0	-45	0

Table 2. Sampling angles (in degrees) used for $3DGS_{RGB}$ and $3DGS_{SEG}$.

The baseline model is fine-tuned over 3000 iterations with a learning rate of $1e-4$. While fine-tuning, we only focus on the following features: [face, eyebrows, eyes, ears, nose, lips, neck]. These features are clearly visible in all the identities and simplify the training set. We showcase a few of these segmentation masks used for our training in Fig. 2.
Fig.3 compares segmentation masks from the baseline model, 3DGS, and manually refined versions. The 3DGS masks are noticeably closer to the refined masks than the baseline outputs, demonstrating the reduced manual effort required for refinement.

0.3. Qualitative results

As mentioned Fig. 4 demonstrates highly precise face-parsing results on held-out test dataset. To assess the robustness of our fine-tuned model on an out-of-distribution dataset, we evaluate it using the NeRSemble [5] dataset. Fig. 5 illustrates the improved results by our fine-tuned model which are otherwise inaccurately predicted by state-of-the-art face-parsing models.

1. REFERENCES

[1] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao, "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[2] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Menghua Wu, Qiu Shen, Ruigang Yang, and Xun Cao, "Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis, "3d gaussian splatting for real-time radiance field rendering.," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[4] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in

[2] https://playcanvas.com/supersplat/editor

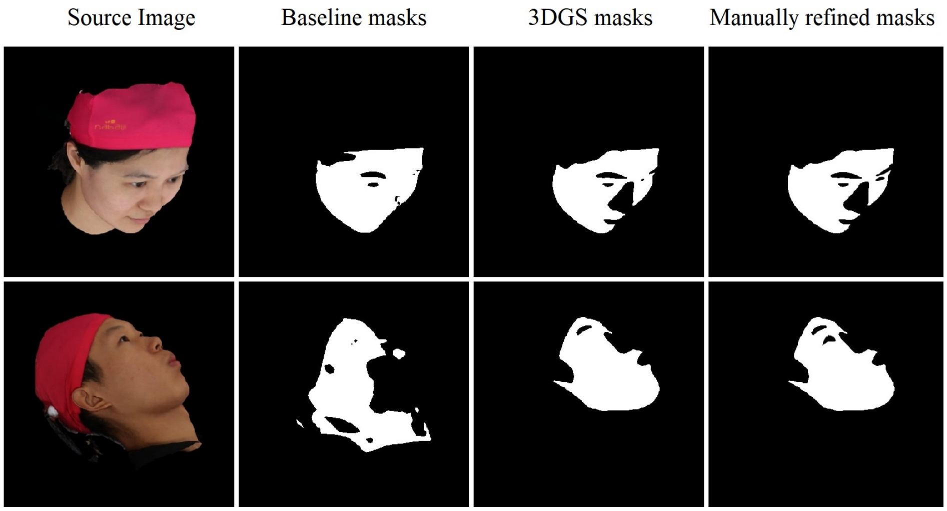

Fig. 3. Comparison of segmentation masks from the baseline model, 3DGS, and manually refined versions. The 3DGS masks require minimal manual corrections—only the right eye and right eyebrow in the top row and the left eye in the bottom row—whereas the baseline masks would need extensive refinements to match the final version. This demonstrates the reduced manual effort needed with our proposed method using 3DGS.

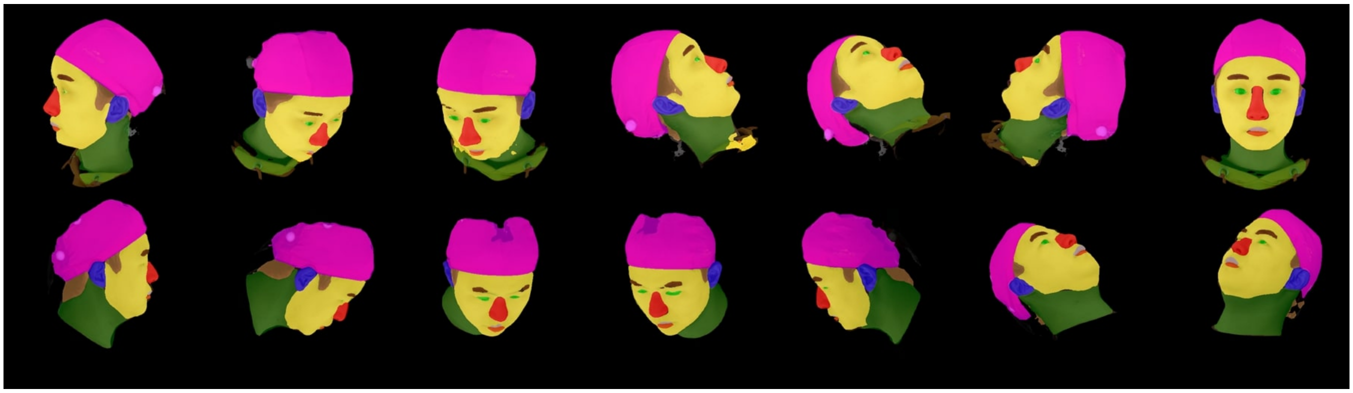

Fig. 4. Predicted segmentation masks on test images of two test subjects *(top and bottom)*

Proceedings of the European conference on computer vision (ECCV), 2018, pp. 325–341.

[5] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner, “Nersemble: Multiview radiance field reconstruction of human heads,” *ACM Trans. Graph.*, vol. 42, no. 4, jul 2023.

[6] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *Advances in neural information processing systems*, vol. 34, pp. 12077–12090, 2021.

[7] Kartik Narayan, Vibashan VS, Rama Chellappa, and Vishal M Patel, “Facexformer: A unified transformer for facial analysis,” *arXiv preprint arXiv:2403.12960*, 2024.

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

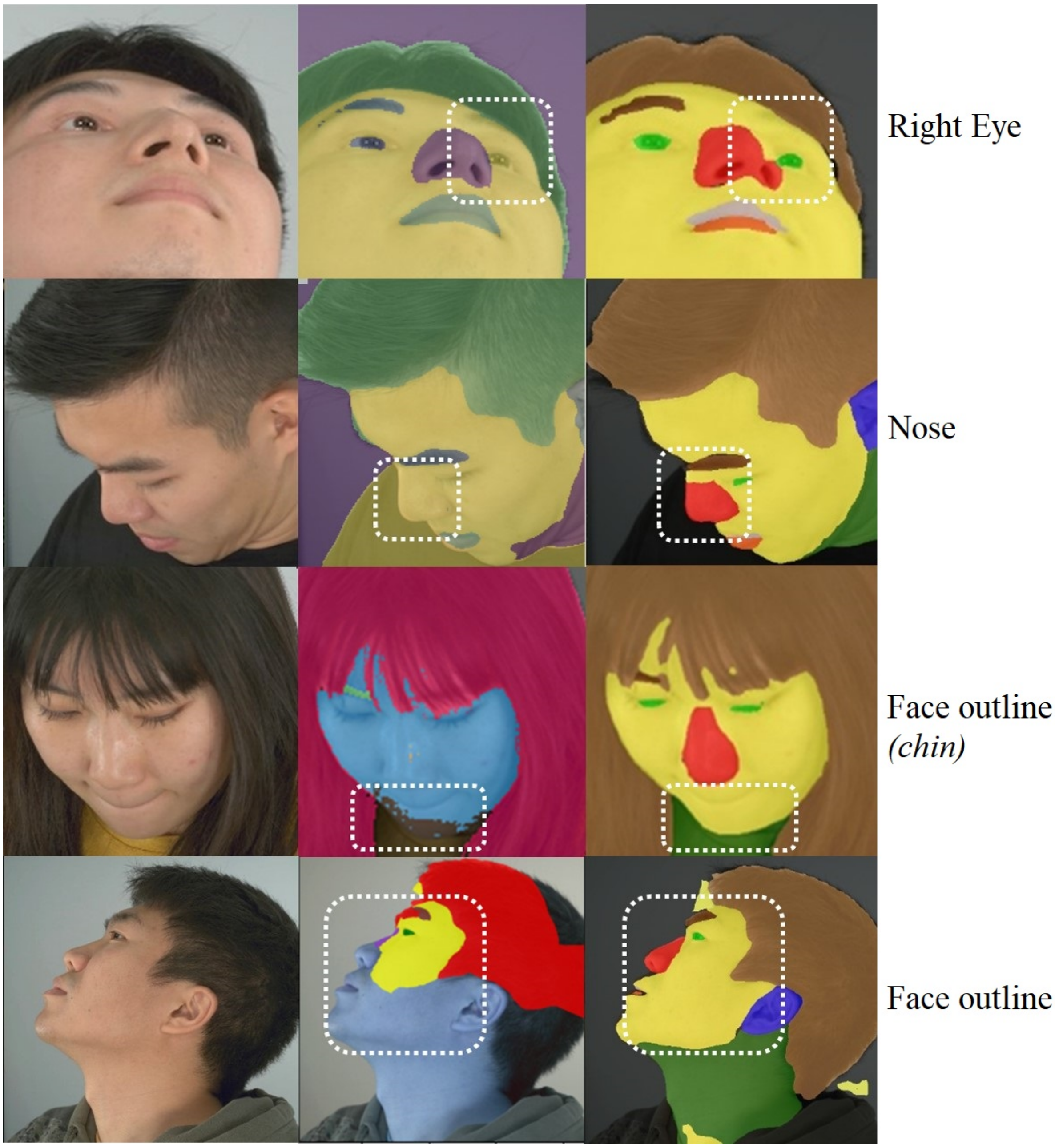

Fig. 5. Qualitative comparison of face-parsing results on an out-of-distribution dataset. The left column shows source images, the second column presents predictions from state-of-the-art models (1st and 2nd row: SegFormer [6], 3rd row: FaceXFormer [7], 4th row: RoI Tanh-polar Transformer [8]), the third column displays results from our fine-tuned BiSeNet, and the fourth column highlights the specific limitations in each prediction.