Supplemental Material for StyleFaceUV: A 3D Face UV Map Generator for View-Consistent Face Image Synthesis

Wei-Chieh Chung*[1]
r08922002@ntu.edu.tw
Jian-Kai Zhu*[1]
r10922033@csie.ntu.edu.tw
I-Chao Shen[2]
ichao.shen@ui.is.s.u-tokyo.ac.jp
Yu-Ting Wu[3]
yutingwu@mail.ntpu.edu.tw
Yung-Yu Chuang[1]
cyy@csie.ntu.edu.tw

[1] National Taiwan University, Taipei, Taiwan
[2] The University of Tokyo, Tokyo, Japan
[3] National Taipei University, New Taipei City, Taiwan

1 Implementation detail

1.1 Step size for the symmetric stylecode

Our method requires a symmetric stylecode for synthesizing the opposite side of the face that is invisible from the original view. As introduced in the main paper, we obtain the symmetric stylecode by shifting the original stylecode. We first find a direction d in the StyleGAN2's latent space so that moving a stylecode along d generates face images from different viewing angles. With the obtained d, we then determine the proper step size for shifting the view angle of the original stylecode, so that the shifted stylecode outputs a face under the desired symmetric view. To predict the step size, we design the "stylecode-to-yaw-angle" mapping network, which is a multi-layer perceptron containing three hidden layers. These three hidden layers are composed of $9 \cdot 512$, $6 \cdot 512$, $3 \cdot 512$ hidden units from the shallow hidden layer to the deep one, with each inter-connected neuron using Tanh as their activation function.

To train the "stylecode-to-yaw-angle" mapping network, we use the "stylecode-to-3DMM coefficients" mapping network described in Section **??** to generate the supervisory signals. To be specific, we first map a stylecode to its 3DMM coefficients using the "stylecode-to-3DMM coefficients" mapping network. After that, we retrieve the rotation part from the 3DMM coefficients and take its negative value as the pseudo ground truth of the symmetric yaw angle. The network is trained by minimizing the L1 distance between the yaw angle retrieved from the predicted 3DMM coefficients corresponding to the shifted stylecode and the

* These authors contributed equally to this work.

pseudo ground truth. To train this mapping network, we use Adam optimizer with learning rate decay from 1×10^{-3} to 1×10^{-6}. The number of training epochs and batch sizes are set to 25 and 16, respectively.

1.2 Generator design

Our network architecture is built on top of StyleGAN2 while adding some modifications to generate 3D data. The major difference is that StyleGAN2 uses a single tRGB output layer to convert per-pixel high-dimensional data to RGB data, while our method uses two separate tRGB layers to output the diffuse map and the generalized displacement map, respectively. Moreover, we add additional constraints from the symmetric view to enforce multi-view consistency. To avoid numerical issues, we apply Tanh activation function on both layers' outputs.

1.3 Texture mapping

Our model generates a diffuse map and a generalized displacement map in UV representation to enhance the appearance and geometry details of a 3D face model. To apply these maps' diffuse color or displacement on vertices, we use 3D-2D texture mapping generated from 3DMMasSTN [1]. The diffuse map stands for the surface color attribute under the original lighting conditions, while the generalized displacement map stands for the surface displacement along XYZ directions.

1.4 Weighted mask

We generate the weighted masks of the original view and symmetric view by utilizing the rendering gradients from the differentiable renderer. To be specific, we first initialize a non-zero texture with the same UV parameterization as the diffuse map. We then use the texture as surface color to render the face using the differentiable renderer. By aggregating the rendered image with summation and back-propagating, we can obtain a pixel-wise mask that records the regions accessed by a view, where the texels have non-zero gradients. Based on the gradient data, we then generate the weighted masks for the original view and symmetric view. For both masks, the texels' values are set to 1.0 if they have non-zero gradients and to 0.0 otherwise. For the symmetric view, we override the texel values of the weighted mask to 0.5 empirically if the texels are also assessed by the original view. Figure 1 showcases some examples of the weighted masks.

2 Experiment

2.1 Ablation study

In this study, we first compare our model trained with only a single angle of StyleGAN2 images, with the one trained with symmetric angles of StyleGAN2 images. We remove the multi-view consistency loss term and train the network using only one angle of the StyleGAN2-generated images. Figure 2 demonstrates the comparison results of the models trained on single-view and multi-view images. The results are similar for both models if the original stylecode produces a front view of the face. On the other hand, if the original

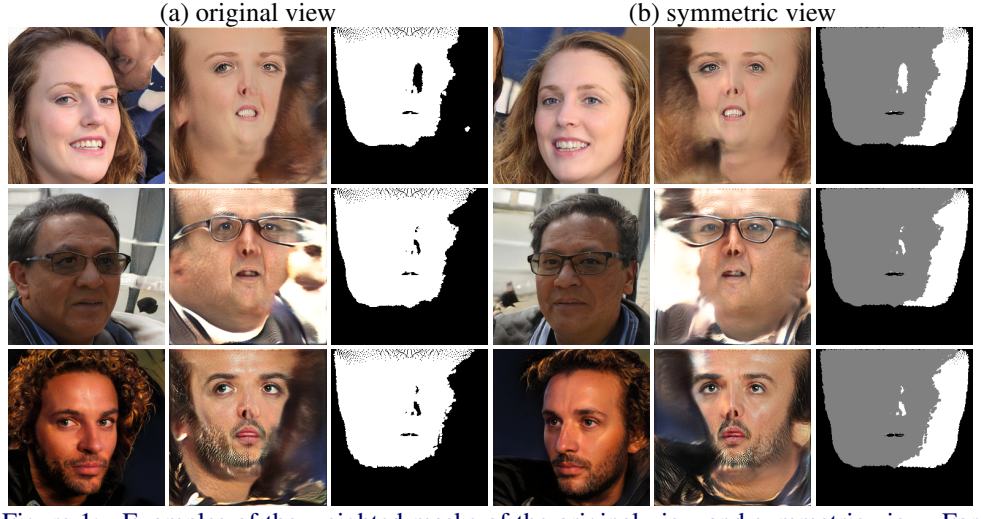


Figure 1: Examples of the weighted masks of the original view and symmetric view. For each example, the first three images are the StyleGAN2's output, the generated diffuse map, and the weighted mask of the original view. The last three images are the StyleGAN2's output, the diffuse map, and the weighted mask of the symmetric view.

stylecode generates a side view, the model trained with multi-view images provides better results since the single-view image does not incorporate the invisible side. We then compare the generated 3D shape with and without displacement maps. As shown in Figure 3, with the displacement map perturbating the vertices on the face, we can represent more geometric details such as wrinkles, frowns, and nasal lines.

2.2 More results

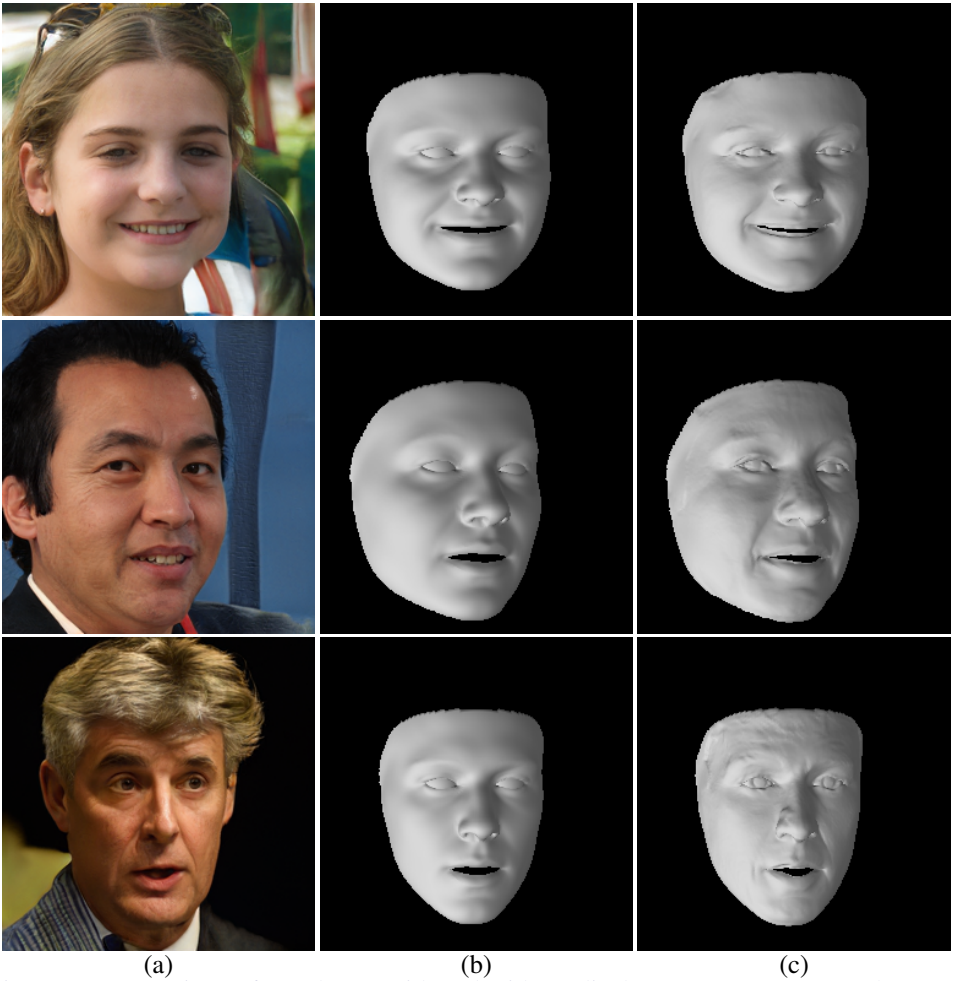
In this subsection, we provide additional results of smooth transition between stylecodes Figure 6) and the controllability of expressions (Figure 4) and lightings (Figure 5).

References

[1] Anil Bas, Patrik Huber, William AP Smith, Muhammad Awais, and Josef Kittler. 3d morphable models as spatial transformer networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 904–912, 2017.

(a) (b) (c)

Figure 2: The synthesized faces from our model trained with and without multi-view loss. (a) Original-view StyleGAN2 images. (b) Symmetric-view StyleGAN2 images. (c) Our synthesized faces. For each example, the upper and lower rows are the synthesized faces from our model trained without and with multi-view loss, respectively.

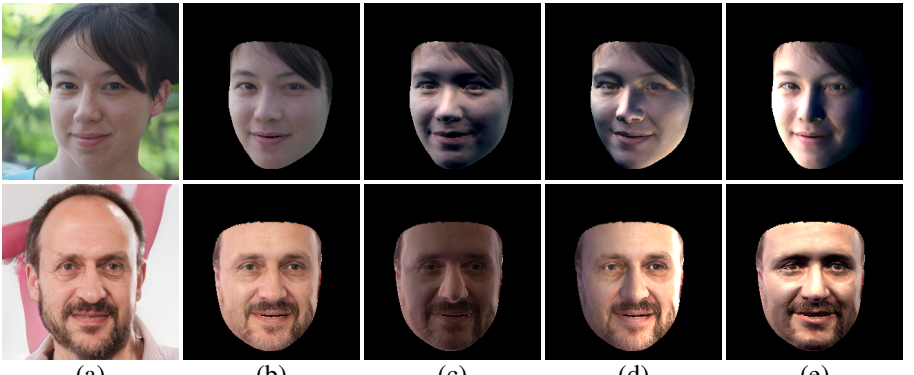


(a) (b) (c)

Figure 3: Comparison of 3D shapes with and without displacement maps. (a) StyleGAN2 images. (b) The shape without the displacement map. (c) The shape with the displacement map.

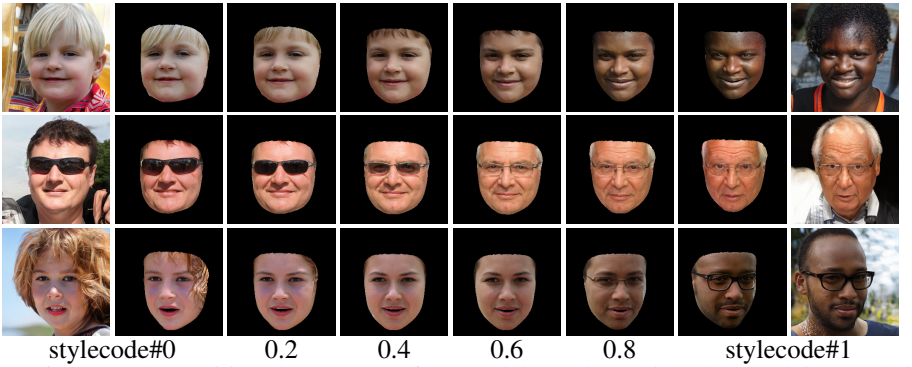
(a) (b) (c) (d) (e)

Figure 4: The controllability of expressions. In each example, we show (a) StyleGAN2 image. (b) the rendered image using our method. From (c)-(e), we show the generated 3D face with manipulated expressions.



(a) (b) (c) (d) (e)

Figure 5: The controllability of lightings. In each example, we show (a) StyleGAN2 image. (b) the rendered image using our method. From (c)-(e), we show the generated 3D face with manipulated lighting.



stylecode#0 0.2 0.4 0.6 0.8 stylecode#1

Figure 6: Smooth transitions between 3D face models can be achieved through interpolating stylecodes.