Supplementary Materials for Learning to Generate Conditional Tri-plane for 3D-aware Expression Controllable Portrait Animation

Taekyung Ki[1], Dongchan Min[2], and Gyeongsu Chae[1]

[1] DeepBrainAI Inc., South Korea
[2] Graduate School of AI, KAIST, South Korea
taek@deepbrain.io alsehdcks95@kaist.ac.kr gc@deepbrain.io
https://deepbrainai-research.github.io/export3d

1 Supplementary Material

1.1 3D Morphable Models (3DMM).

3D Morphable Models (3DMM) [2] are statistical models of 3D shape and their corresponding texture. In this paper, we only consider the shape representation of 3DMM. To be specific, a face shape $\mathbf{S}$ is initialized with the average shape $\bar{\mathbf{S}}$ and further shaped by a linear combination of expression and identity as follows:

$$\mathbf{S} = \bar{\mathbf{S}} + \alpha \mathbf{U}_{id} + \beta \mathbf{U}_{exp}, \tag{1}$$

where $\mathbf{U}_{id} \in \mathbb{R}^{80 \times d_{3dmm}}$, $\mathbf{U}_{exp} \in \mathbb{R}^{68 \times d_{3dmm}}$ are the pre-defined bases of identity and expression subspaces of 3D face space, respectively. d_{3dmm} is the dimension of the 3D face space. The coefficients $\alpha \in \mathbb{R}^{80}$ and $\beta \in \mathbb{R}^{64}$ determine the facial identity and expression for the face geometry reconstruction by scaling each basis vector [1].

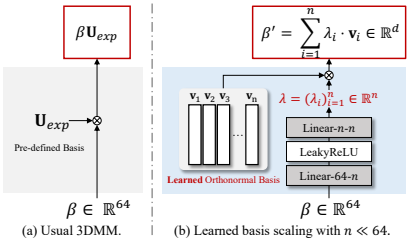


Fig. S1: 3DMM [2] vs. Leaned Basis Scaling (LeBS). 3DMM based method reconstructs 3D facial geometry by scaling the the pre-defined basis $\mathbf{U}_{exp}$ with expression parameters $\beta \in \mathbb{R}^{64}$. LeBS, on the other hand, uses the learned basis $V = \{v_i\}_{i=1}^{n} \subseteq \mathbb{R}^d$ which is scaled by the low-dimensional coefficients $\lambda = (\lambda_i)_{i=1}^{n} \in \mathbb{R}^n$ $(n \ll 64)$.

In this paper, we term **appearance** as the set of geometric features that determine the facial identity of a given face, such as head size, face contour, face proportion, eyebrows, eye shape, mouth shape, jaw shape, etc., and **expression** as the motion of these appearance features, such as mouth opening (closing), eye blinking, etc.

1.2 Detailed Model Architectures.

Our model consists of four parts: Learned Basis Scaling (**LeBS**), Hybrid Triplane Generator **G**, Light-weight MLP decoder (**MLP**) for color and density

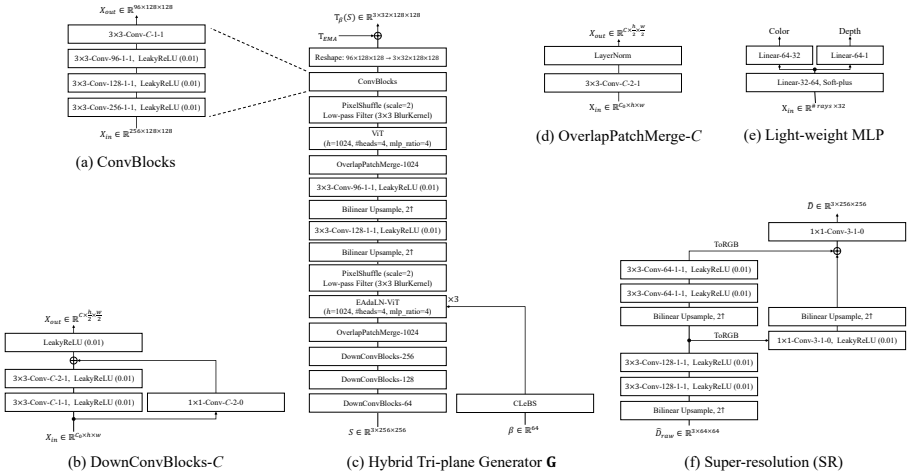


Fig. S2: The detailed model architectures. $k \times k$-Conv-C-s-p is the convolution operator with the kernel size $k \times k$, output channel size C, stride step s, and padding size p. Linear-C_0-C_1 is the fully-connected layer of the input channel size C_0 and the output channel size C_1.

prediction used in the differentiable volume rendering [17], and Super-resolution (**SR**) module. The detailed model architectures are shown in Fig. S1 and Fig. S2.

LeBS consists of two fully-connected layers along with the learned orthonormal basis $V \subseteq \mathbb{R}^d$. We apply QR-decomposition [21] to a learnable weight in $\mathbb{R}^{d \times n}$ to explicitly compute $V \subseteq \mathbb{R}^{n \times d}$. We set the dimension of the expression space $d = \frac{h}{4}$ to be same as the dimension of the visual tokens where $h = 1024$ is the size of the hidden state in the EAdaLN-ViT blocks. We experimentally choose $n = 10$ for the number of basis vectors. We observe that increasing n produces duplicated expression directions. For the contrastive pre-training of LeBS, we employ ResNetSE18 feature extractor [8] followed by a single fully-connected layer to output the d-dimensional vector, serving as the image encoder $f_I(\cdot)$. Notably, we do not introduce an orthonormal basis to $f_I(\cdot)$.

Inspired by [20], we incorporate ViT blocks [7] into our generator $\mathbf{G}$, specifically utilizing those from SegFormer [22] and DiT [18]. In both EAdaLN-ViT and ViT, we employ four heads with 1024 hidden dimensions for the multi-head self-attention. It is worth mentioning that the architectures of EAdaLN-ViT and ViT illustrated in Fig. S2 are the same, with the exception of EAdaLN integration for expression transfer. We employ the exponential moving average (EMA) on the tri-planes for stabilizing the training. More precisely, in the j-th gradient step, we calculate and update the EMA T^j_{EMA} and the current tri-plane T^j as follows:

$$\mathrm{T}^j_{EMA} \leftarrow \delta \cdot \mathrm{T}^{j-1}_{EMA} + (1-\delta) \cdot \bar{\mathrm{T}}^j \quad \text{and} \quad \mathrm{T}^j \leftarrow \mathrm{T}^j + \mathrm{T}^{j-1}_{EMA} \tag{2}$$

where $\bar{\mathrm{T}}^j$ is the average tri-plane calculated within the j-th batch and T^0_{EMA} is initialized by $\mathbf{0} \in \mathbb{R}^{3\times32\times128\times128}$. We set $\delta = 0.998$ as the weight for the moving average.

MLP for color and density prediction consists of a stack of fully-connected layers with soft-plus activation. In contrast to [4], we use two fully-connected layers to separately predict them.

For **SR**, we follow the super-resolution module used in [4, 10] except for the style modulated convolutions.

1.3 Training Objectives

Our model is trained with reconstruction manner that reconstruct a driving frame D from a source frame S with the driving expression parameters β_D and camera parameters p_D where these frames are randomly sampled from the same video clip. The training consists of two stages. In the first phase, we employ MSE loss $\mathcal{L}_2$ and VGG16 [19] multi-scale perceptual loss $\mathcal{L}_{lpips}$ [23] to minimize the perceptual distance between the generated frame $\hat{D}$ and the driving frame D. We also minimize the distance between the raw rendered image $\hat{D}_{raw}$ and raw driving image D_{raw} using the same loss functions, denoted by $\mathcal{L}_2^{raw}$ and $\mathcal{L}_{lpips}^{raw}$, respectively:

$$\mathcal{L}_{rec} = \mathcal{L}_2^{raw} + \mathcal{L}_2 + \mathcal{L}_{lpips}^{raw} + \mathcal{L}_{lpips}. \tag{3}$$

In the second phase, we integrate the conditional discriminator used in [9], using the camera parameter as additional condition and employing binary cross-entropy loss to compute adversarial loss $\mathcal{L}_{adv}$. The total loss function $\mathcal{L}_{total}$ is

$$\mathcal{L}_{total} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv}, \tag{4}$$

where λ_{rec} and λ_{adv} are balancing coefficients.

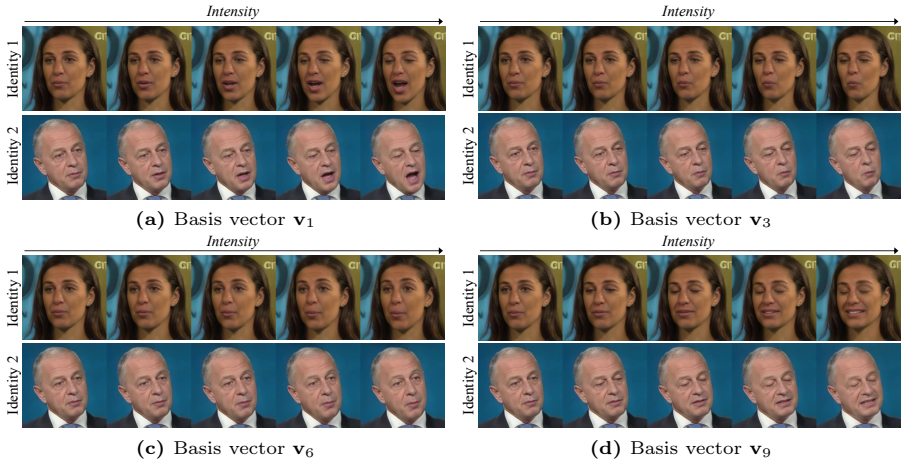
1.4 More Implementation Details.

Training. Since our model does not rely on pre-trained EG3D [4, 20], it is trained end-to-end, except for CLeBS. For the contrastive pre-training of LeBS, we draw 32 negative samples for each positive sample, set the temperature τ to 0.07, and train it for 60,000 steps. Longer pre-training does not lead to significant performance improvements.

We empirically set the balancing coefficients in Eq. (4) by $\lambda_{rec} = 1$, and $\lambda_{adv} = 0.01$. We train our model for 300,000 steps with the reconstruction loss Eq. (3) and then incorporate the adversarial loss Eq. (4) for 10,000 steps to slightly improve the visual quality. For all training, we use Adam [13] optimizer with the learning rate 10^{-4} for Export3D, 10^{-4} for CLeBS, and 10^{-5} for the discriminator, respectively. Overall training conducts on a single A100 GPU about 5 days with batch size 8. In the inference phase, we use randomly sampled frontal frame as the source frame.

Table S1: Quantitative comparison of on VFHQ with "background".

Method	Same-identity						Cross-identity		
	PSNR↑	**SSIM↑**	**AKD↓**	**CSIM↑**	**AED↓**	**APD↓**	**CSIM↑**	**AED↓**	**APD↓**
ROME	8.309	0.400	11.179	0.592	0.123	0.173	0.495	0.236	0.201
OTAvatar	10.667	0.457	15.236	0.492	0.181	0.182	0.492	0.288	0.237
HiDeNeRF	12.254	0.345	22.136	0.354	0.135	0.252	0.408	0.259	0.230
Ours	**23.555**	**0.704**	**3.453**	**0.811**	**0.082**	**0.030**	**0.694**	**0.208**	**0.080**



(a) Basis vector $\mathbf{v}_1$ **(b)** Basis vector $\mathbf{v}_3$

(c) Basis vector $\mathbf{v}_6$ **(d)** Basis vector $\mathbf{v}_9$

Fig. S3: Linear scaling along the different basis vectors of CLeBS.

1.5 Evaluation.

Evaluation metrics. We provide additional explanations of the evaluation metrics. Average key-point distance (**AKD**) is the L1 distance of 68 facial key-points between the generated image and the driving image, which measures the facial structure similarity based on the key-points. We use the face-alignment [3] to extract the key-points. Cosine similarity of identity embedding (**CSIM**) is the cosine similarity between the identity embeddings of the source image and the generated image where the embeddings are extracted from ArcFace [5]. Average expression distance (**AED**) and average pose distance (**APD**) are the L1 distance between the expression parameters (64 dimensions) and the pose parameters (6 dimensions), respectively extracted from the generated image and the driving image. We use the 3DMM extractor [6] to extract those parameters.

1.6 Additional Results.

Further comparison without removing the background. In Tab. S1, we provide additional quantitative comparison with ROME [12], OTAvatar [16], and

Fig. S4: Novel-view synthesis results with expression transfer.

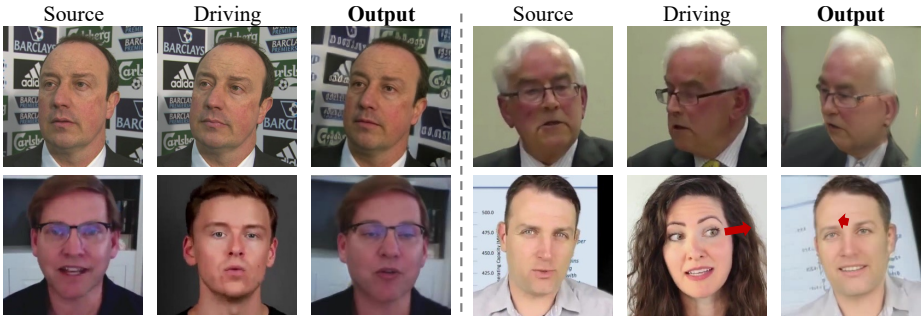


Fig. S5: Limitation cases of Export3D. The red arrows indicate the directions of eye gaze.

HiDe-NeRF [14] to verify that these models have advantage on the evaluation metrics without background.

Linear scaling along the orthonormal basis. In Fig. S3, we show additional results of linear scaling along the different basis vectors [21]. For $\mathbf{v}_1$, we scale λ_1 from 1 to -7, showing mouth opening and eye closing. For $\mathbf{v}_3$, we scale λ_3 from 1 to 20, showing eye closing and lip pursing. For $\mathbf{v}_6$, we scale λ_6 from 1 to -7, showing eyebrow moving. For $\mathbf{v}_9$, we scale λ_9 1 from to -10, showing eye closing and smiling. Since our method does not constrain the range of the coefficients $\lambda = (\lambda_i)_{i=1}^{10}$, the manipulation can be realized along the negative scaling. Please refer to video results.

Additional comparison with HiDe-NeRF. In Fig. S4, we exhibit additional comparison results with HiDe-NeRF [14] for novel-view synthesis with expression transfer. Please refer to the video results for further details.

1.7 Limitations and Future Work.

We exhibit the limitation cases of Export3D in Fig. S5. Since the tri-plane represents [4] the foreground and the background as a whole, our model jointly renders

them, resulting in head pose-aligned distortion. Several prior works [12, 14–16] address this issue by removing the complex background and providing the volume rendering with a uniform background. However, they heavily rely on the performance of the background segmentation model [11], exhibiting the temporal jitters in the generated videos. Additionally, our model cannot control eye gazing since the 3DMM parameters do not model eye movement. We leave these limitations for future research.

References

1. Amberg, B., Knothe, R., Vetter, T.: Expression invariant 3d face recognition with a morphable model. In: 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. pp. 1–6 (2008)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1021–1030 (2017)
4. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3d generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16123–16133 (2022)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4690–4699 (2019)
6. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: 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 (CVPRW). pp. 0–0 (2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7132–7141 (2018)
9. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in Neural Information Processing Systems (NeurIPS) **33**, 12104–12114 (2020)
10. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8110–8119 (2020)
11. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1140–1147 (2022)

12. Khakhulin, T., Sklyarova, V., Lempitsky, V., Zakharov, E.: Realistic one-shot mesh-based head avatars. In: European Conference on Computer Vision (ECCV). pp. 345–362 (2022)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Li, W., Zhang, L., Wang, D., Zhao, B., Wang, Z., Chen, M., Zhang, B., Wang, Z., Bo, L., Li, X.: One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17969–17978 (2023)
15. Li, X., De Mello, S., Liu, S., Nagano, K., Iqbal, U., Kautz, J.: Generalizable one-shot 3d neural head avatar. Advances in Neural Information Processing Systems (NeurIPS) **36** (2024)
16. Ma, Z., Zhu, X., Qi, G.J., Lei, Z., Zhang, L.: Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16901–16910 (2023)
17. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
18. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748 (2022)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
20. Trevithick, A., Chan, M., Stengel, M., Chan, E.R., Liu, C., Yu, Z., Khamis, S., Chandraker, M., Ramamoorthi, R., Nagano, K.: Real-time radiance fields for single-image portrait view synthesis. arXiv preprint arXiv:2305.02310 (2023)
21. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. arXiv preprint arXiv:2203.09043 (2022)
22. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems (NeurIPS) **34**, 12077–12090 (2021)
23. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)