

# Supplementary Materials: DisControlFace: Adding Disentangled Control to Diffusion Autoencoder for One-shot Explicit Facial Image Editing

Anonymous Authors

## 1 OVERVIEW

In this supplement, we present:

- Section 2: Implementation details.
- Section 3: Additional experiments and analyses.
- Section 4: Model architectures.
- Section 5: More discussions about limitations and future work.

## 2 IMPLEMENTATION DETAILS

The complete configurations of the training of Exp-FaceNet and one-shot fine-tuning are shown in Table 1. We utilized the official codes to generate the results of the compared baselines HeadNeRF<sup>1</sup> [4], GIF<sup>2</sup> [1], and DiffusionRig<sup>3</sup> [3]. In inference, we generate the initial noise map  $X_T$  by sampling from  $\mathcal{N}(0, I)$  rather than computing through a reverse deterministic generative process proposed in the original Diff-AE [5]. The further analysis of this setting is provided in Section ??.

## 3 ADDITIONAL EXPERIMENTS AND ANALYSES

### 3.1 Semantic Conditioning of Diff-AE

In original Diff-AE [5], the initial noise map  $X_T$  in inference is computed through a reverse deterministic generative process, which has a capacity for capturing stochastic details. In Figure 1, we compare the reconstruction and editing results of separately using the reverse deterministic noise and the randomly sampled noise as the initial denoising map in Diff-AE. It can be observed that using a deterministic initial noise map can achieve a better reconstruction of the input image such as the background, however generating the editing images with less accurate explicit control and incoherent facial appearance. This might be because that the reverse deterministic noise computing can encode the stochastic details of the input image, which is crucial for a near-exact reconstruction. However, this process inherently conflicts with the explicit face editing where some details of the input image should be changed based on the modifications of the explicit parameters. Also it is inconsistent with the initialization in the training of Exp-FaceNet. Given this, in this work, we choose to generate the initial noise map  $X_T$  by sampling from  $\mathcal{N}(0, I)$  in inference. Moreover, we can find that using a simple one-shot fine-tuning can effectively enhance the

<sup>1</sup><https://github.com/CrisHY1995/headnerf>

<sup>2</sup><https://github.com/ParthaEth/GIF>

<sup>3</sup><https://github.com/adobe-research/diffusion-rig>

	Exp-FaceNet training	One-shot fine-tuning
Image size	256	
Patch size	16	
Image normalization	[-1, 1]	
Masking ratio	sample from $U(0.25, 0.75)$	
Optimizer	AdamW (no weight decay)	
Diffusion loss	MSE loss	
EMA decay	0.9999	\
Learning rate	1e-4	1e-5
Batch size	32	4
Denoising steps	1000	
Iterations	437500	1500
Device	8 V100s	1 V100
Training time	3 days	4minutes

Table 1: The complete configurations of the training of Exp-FaceNet and one-shot fine-tuning.

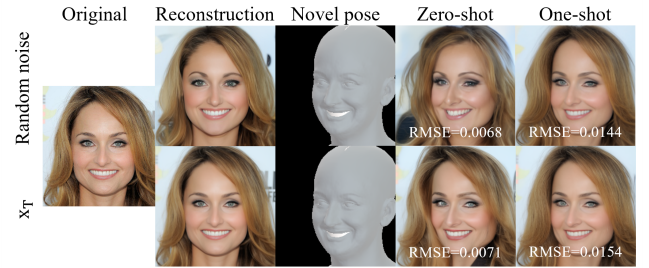


Figure 1: The reconstruction and editing results of using different initial noise map computing strategy in Diff-AE.  $X_T$  indicates using the original reverse deterministic computing to generate the initial noise map of Diff-AE.

semantics preservation in the editing result, e.g., more faithful hair style.

### 3.2 Additional Ablation Studies

**Fine-grained face geometry.** In EMOCA [2], a person-specific detail vector  $\delta$  is specially estimated and can be further combined with pose parameter  $\theta$  and expression parameter  $\phi$  to generate the expression dependent displacement map for refining the face geometry with animatable wrinkle details. Here we compare the editing results between using estimated FLAME parameters to calculate the 3D face mesh and additionally introducing the detail vector  $\delta$  in the mesh calculation. Figure 2 demonstrates that using  $\delta$  can generate a rendered snapshot with refined expression-dependent face geometry, which thereby helps to recover the detailed expressions



**Figure 2: Ablation study on fine-grained face parameters. Adopting the detail vector  $\theta$  estimated by EMOCA [2] can help to generate the control condition with more fine-grained face geometry, which allows faithful facial details preservation.**

(e.g., wrinkles) of the edited face.

**Different masking strategies in inference.** Here we explore how the masking strategy used in inference affects the editing performance. The comparison is shown in Figure 3, we can observe that when setting masking ratio to 0% for all inference steps, the edited image can not match the control signal very well which can be attributed to strong deterministic reconstruction in this setting. Meanwhile, setting masking ratio to 75% and 25% for all inference steps slightly harm the semantics recovering and explicit control, respectively. Besides, we can see the other three masking strategies can achieve better editing results where the proposed linear masking ratio can perform overall best editing with accurate face control and good preservation of facial semantics.

## 4 MODEL ARCHITECTURES

The structural details of the proposed DisControlFace is presented in Figure 4. The detailed architecture of Diff-AE is provided in the published paper [5] and released code<sup>4</sup>. The proposed Exp-FaceNet mirrors the structure of the Conditional DDIM (i.e., U-Net) in Diff-AE, which however, customizes the input layer by setting the input channel number to 6, allowing it to take the concatenated snapshots as input. On this basis, we fuse the 2D feature maps outputted by the input layers of Conditional DDIM and Exp-FaceNet by pixel-wise summation, then feeding the fused feature maps into the subsequent layers of Exp-FaceNet for generating spatial-wise explicit conditioning features. To further provide fine-grained conditioning for the diffusion generation process, we add multi-scale features outputted by the decoder blocks of the Exp-FaceNet ( $f_A$  to  $f_M$  in Figure 4) back to the corresponding blocks of the Diff-AE backbone.

## 5 ADDITIONAL LIMITATIONS AND FUTURE WORK

Our DisControlFace has a separate editing control network besides the U-Net noise predictor and performs denoising diffusion process in image space, which results in the model being able to generate images with limited resolutions. Potential future improvements includes introducing a light-weight super-resolution network to the model or extending the model to a latent diffusion version.

<sup>4</sup><https://github.com/phizaz/diffae>

In the proposed Masked Diff-AE training, we randomly mask some patches of the input image as the input of the semantic encoder of the Diff-AE backbone, which enables an effective training of Exp-FaceNet in a disentangled setting. Meanwhile, only performing random masking on face-related patches is expected to further improve the consistency of the background region of the input portrait with different editing applied, which however might slightly increase the training and inference time. Corresponding explorations and experiments can serve as another potential valuable future work.



**Figure 3: Ablation study on different masking strategies in inference.** We set the inference denoising steps to 20 for all masking strategies. Strategy A: the masing ratio is set to 0% for all 20 steps; Strategy B: the masking ratio is set to 75% for all 20 steps; Strategy C: the masking ratio is set to 25% for all 20 steps; Strategy D: the masking ratio is set to 25% and 75% for the first 10 steps and last 10 steps; Strategy E: the masking ratio is set to 75% and 25% for the first 10 steps and last 10 steps; Strategy F: the linear masking ratio introduced in the main paper.

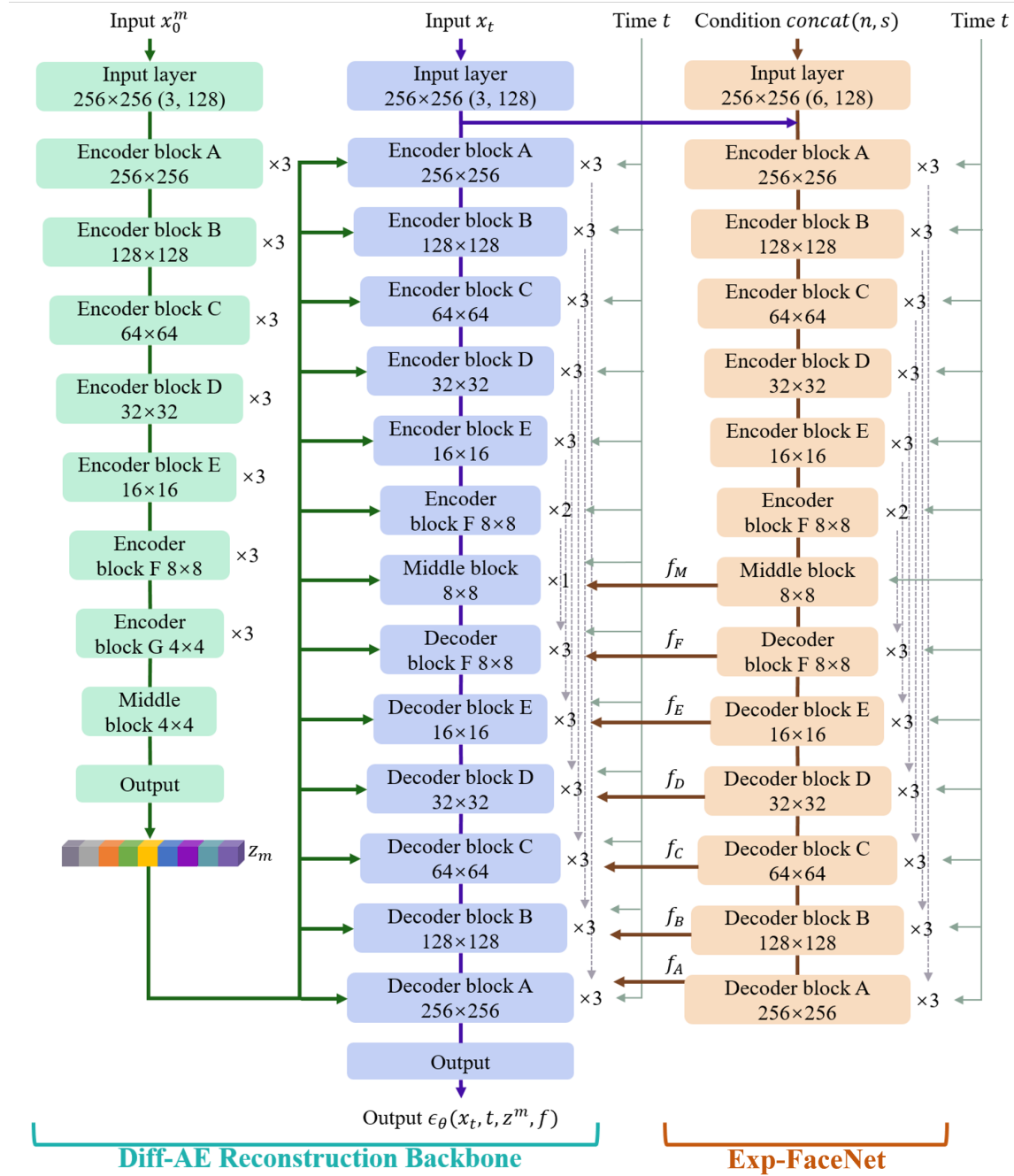


Figure 4: Detailed architecture of the proposed DisControlFace. Please zoom in to see details.

REFERENCES

[1] Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. 2022. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 19900–19916.

[2] Radek Daněček, Michael J Black, and Timo Bolkart. 2022. EMOCA: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20311–20322.

[3] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. 2023. DiffusionRig: Learning Personalized Priors for Facial Appearance Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12736–12746.

[4] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20374–20384.

[5] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. 2022. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10619–10629.