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

Anonymous Authors

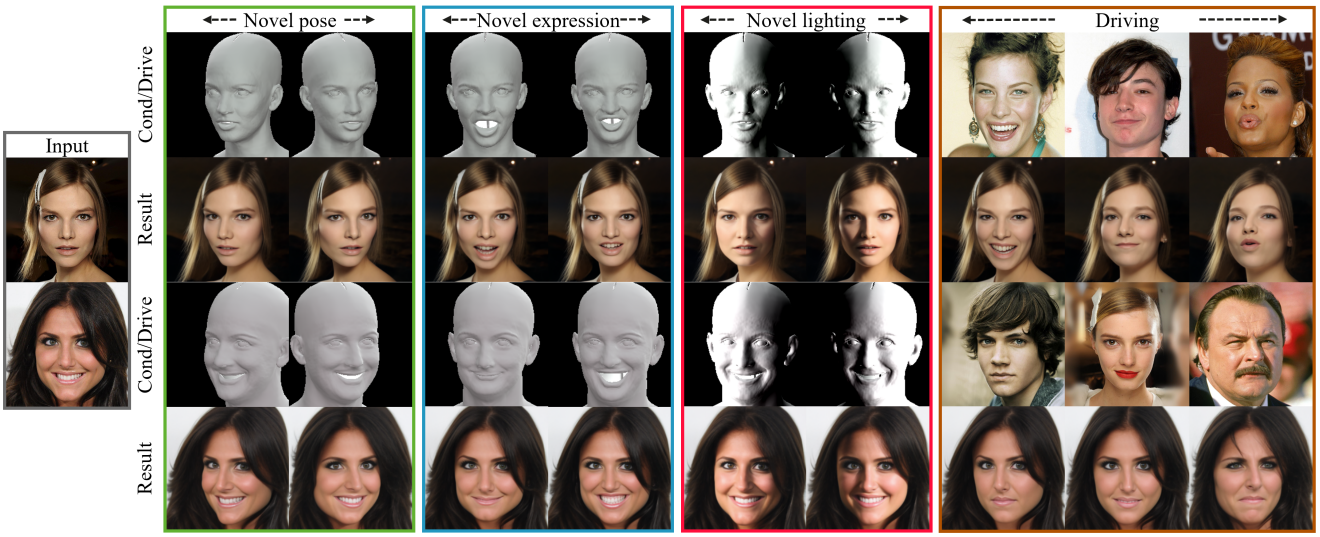

Figure 1: Our DiscontrolFace can edit the input face image based on explicit parametric control and faithfully preserve the facial semantic appearance under one-shot scenario. Our model can generate realistic and faithful facial image corresponding to diverse pose, expression and lighting conditions and also supports cross-identity face driving.

ABSTRACT

In this work, we focus on exploring explicit fine-grained control of generative facial image editing, all while generating faithful facial appearances and consistent semantic details, which however, is quite challenging and has not been extensively explored, especially under an one-shot scenario. We identify the key challenge as the exploration of disentangled conditional control between high-level semantics and explicit parameters (*e.g.,* 3DMM) in the generation process, and accordingly propose a novel diffusion-based editing framework, named DisControlFace. Specifically, we leverage a Diffusion Autoencoder (Diff-AE) as the semantic reconstruction backbone. To enable explicit face editing, we construct an Exp-FaceNet that is compatible with Diff-AE to generate spatial-wise explicit control conditions based on estimated 3DMM parameters. Different from current diffusion-based editing methods that train the whole conditional generative model from scratch, we freeze the pre-trained weights of the Diff-AE to maintain its semantically deterministic conditioning capability and accordingly propose a random semantic masking (RSM) strategy to effectively achieve an independent training of Exp-FaceNet. This setting endows the model with disentangled face control meanwhile reducing semantic information shift in editing. Our model can be trained using 2D in-the-wild portrait images without requiring 3D or video data and perform robust editing on any new facial image through a simple one-shot fine-tuning. Comprehensive experiments demonstrate that DisControlFace can generate realistic facial images with better editing accuracy and identity preservation over state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision problems**.

KEYWORDS

Facial image editing, Explicit parametric control, Conditional diffusion model

Unpublished working draft. Not for distribution.

ACM MM, 2024, Melbourne, Australia

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Facial image editing has long been a hot research topic in the fields of computer vision and computer graphics, where the key challenge is to effectively achieve fine-grained controllable generation of realistic facial images while preserving semantic face priors.

3D Morphable Models (3DMMs) [3, 16] have been widely employed to represent variations in facial shape and texture [4, 16,

19, 29, 56–59], whereas their ability to capture personalized facial features is limited and the performance highly depends on the quality and diversity of the 3D face training data. On top of this, subsequent learning-based explicit face models [7, 11, 13, 28, 50, 56–59, 63] achieve controllable generations of dynamic and expressive facial animations by capturing the nuances of facial features under different expressions, poses, and lighting conditions, nevertheless, can neither generate realistic facial appearances that correspond to the animated 3D face geometries nor model refined geometric details in non-facial regions, *e.g.,* hair, eyes, and mouth. Follow-up efforts integrate explicit facial modeling with implicit 3D-aware representations like Neural Radiance Fields (NeRFs) to reconstruct animated realistic head avatars [14, 23, 30, 33, 34, 39–42, 51, 52, 66], which however, heavily rely on 3D consistent data such as monocular portrait videos and tend to exhibit limited generalization.

In contrast, generative face models enable single image reconstruction and editing due to the superior capability in learning rich face priors from in-the-wild portraits. Recent GAN-based approaches [2, 5, 6, 10, 18, 22, 37, 55, 62] incorporate explicit 3D facial priors and implicit neural representations to achieve directed generations of high-resolution, realistic, and view-consistent facial images without the need of 3D face scans or portrait videos. However, those methods mainly provide implicit or limited explicit controls of face generations.

More recently, the diffusion-based framework [21] has emerged as the predominant choice for various generation tasks, owing to its impressive performance and diverse conditioning options. Some approaches have also shown promising enhancements in face reconstruction and various face editing tasks, such as face relighting [43], semantic attributes manipulation [44], and explicit appearance control [12]. Unfortunately, it can be seen that when editing and modifying some specific facial attributes, other facial attributes or editing-irrelevant details often occur unexpected and uncontrollable changes, leading to an incoherent and identity-altered generated face. This prevalent issue can be attributed to that these generative face models struggle to effectively perform disentangled control in the generation process.

In this work, we propose a novel diffusion-based generative framework, namely DisControlFace to achieve one-shot editing of facial images. To generate a photo-realistic, high-fidelity facial appearance corresponding to specific explicit parameters (*e.g.,* 3DMM) while faithfully preserving high-level semantic priors, we particularly focus on enhancing the diffusion model with disentangled conditional control. Specifically, we adopt a Diffusion Autoencoder (Diff-AE) [44] as the generative backbone, which can enable a deterministic image reconstruction by conditioning Denoising Diffusion Implicit Model (DDIM)[53] on the semantic information of the input image. We then specially construct an Exp-FaceNet compatible with the Diff-AE backbone, which further provides multi-scale, spatial-aware DDIM conditioning corresponding to the facial parameters of shape, pose, expression, and lighting. Moreover, we claim that training different DDIM conditoning together, as with existing methods is not conductive to learning disentangled face control. We therefore freeze the pre-trained weights of Diff-AE and accordingly design a random semantic masking (RSM) strategy to enable the training of Exp-FaceNet, by means of which the model can learn explicit parameteric face control independently without affecting semantically deterministic DDIM conditioning. Also benefiting from this disentangled setting, instead of relying on 3D or video data, we can utilize 2D in-the-wild portrait dataset such as FFHQ [26] to effectively train Exp-FaceNet to learn a robust and generalized capability in explicit face editing. Considering there exists domain gap between the pre-trained face data and the target new face image, which tends to prevent Diff-AE from performing near-exact semantic reconstrcution, we finally introduce an one-shot fine-tuning to Diff-AE so as to restore personal identity and editing-irrelevant details of the input portrait under a subject-agnostic scenario. Our approach not only achieves state-of-the-art (SOTA) qualitative and quantitative results for one-shot explicit facial image editing, but also supports generating realistic and faithful facial appearance of specific individuals in image inpainting, semantic attributes manipulations, and cross-identity face driving (shown in Figure 1).

Our contributions can be summarized as follows:

- We propose a novel diffusion-based generation framework, consisting of a Diffusion Autoencoder (Diff-AE) backbone and an explicit face control network (Exp-FaceNet) for synthesizing photo-realistic, high-fidelity portrait images corresponding to the editing of explicit facial properties only trained with 2D in-the-wild images.
- To the best of our knowledge, we are the first to introduce a weight-frozen pre-trained Diff-AE to explicit face editing pipeline to provide deterministic semantic conditioning, meanwhile designing an effective training strategy to enable the Exp-FaceNet with a disentangled explicit parameteric (*e.g.,* 3DMM) conditioning.
- Our method achieves SOTA generation performance in explicit facial image editing, and also supports various one-shot face editing tasks.

2 RELATED WORK

Generative Face Modeling. Various GAN-based models [6, 26, 27, 35, 38, 49] have been proposed to synthesize realistic facial images by learning the underlying data priors from large-scale in-the-wild images [6, 38, 49]. However, when it comes to precise control and interpretability, those generative face models fall short compared to explicit parametric models. Given this, several approaches [5, 54, 55, 62] go a step further by integrating explicit parameters and GANs, which can simultaneously generate highly realistic and coherent portraits and achieve fine-grained control of facial appearance. Recently, diffusion models [21, 53] have gained recognition for their superior ability to learn data distributions compared to GANs, and thus have been widely adopted to generate realistic and diverse images in various generation tasks, including facial image synthesis [44, 44]. DiffusionRig [12], a closely related method, introduces pixel-aligned physical properties rendered from explicit parameters estimated by DECA [13] to denoising diffusion process to generate photo-realistic facial images corresponding to target pose, expression, and lighting conditions, which however, highly relies on a personalized fine-tuning (around 20 images) to preserve the facial appearance priors of a specific person . Similar problems widely exist in most conditional diffusion face models, as can be observed where identity shifts and unexpected attributes

alterations may occur during face reconstruction and editing. This can be attributed to the lack of disentangled control capabilities when conditioning diffusion models with both facial semantics and physical information. DisControlFace overcomes this challenge by leveraging a weight-frozen pre-trained Diff-AE to provide semantically deterministic DDIM conditioning and independently training a separate Exp-FaceNet to learn disentangled face control based on 3DMM parameters.

Conditional Diffusion Model. Conditional DDIM enables the Denoising Diffusion Probabilistic Model (DDPM) [21] to generate content consistent with specific control signals though various conditioning manners. Most existing models encode various control information into global conditional vectors, which can be text embedding [45] or semantic embedding [43, 44]. To achieve spatial-aware and precise control of the generation, some approaches (*e.g.,* DiffusionRig [12] and SR3 [48]) concatenate various spatial conditions and denoised images together as the input of the U-Net noise predictor in each denoising step. However, this form of conditioning requires the U-Net to have a unique input layer, resulting in the model having limited generalization and making it hard to reuse existing well-trained diffusion models. Besides, some other methods specially construct spatial conditioning branches to extract spatial-aligned conditional features and insert them into the U-Net [43, 61, 65]. ControlNet [65] has been widely employed to add various spatial visual guidance (*e.g.,* edge maps, pose maps, depth maps, *etc.*) to text-to-image generation models such as Stable Diffusion [45]. Whereas, it may not suitable for deterministic reconstruction or editing tasks since there still exists uncertainty and randomness in the generation controlled by visual guidance and text prompts. In contrast, our Exp-FaceNet is specially designed to be compatible with a semantic reconstruction DDIM, Diff-AE [44], aiming to better address facial image editing based on explicit control information.

Learning Specific Facial Priors. Effectively extracting the facial appearance priors of the specific person and injecting this global prior into the generation process is crucial for preserving facial semantics such as identity, accessories, hairstyle, and background information in facial image editing. Most existing generative methods address this issue by fine-tuning the network in various settings [15, 24, 36, 47], or designing special optimization strategies, such as identity penalty, face recognition loss, and latent representation editing [1, 31, 60, 64]. In contrast to previous work like [12, 36] which collect personal albums to learn personalized facial priors, in this work, we focus on the subject-agnostic editing scenario, which is more challenging but practical. On the basis of the inherent and maintained facial semantics capturing capability of Diff-AE, only a fast and yet simple fine-tuning using an out-of-domain new face image is needed to faithfully restore the semantic appearance details of the target image during editing.

3 PRELIMINARIES

3.1 3D Morphable Face Models

FLAME [29] , a popular 3DMM model, can be expressed as $M(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}) : \mathbb{R}^{|\boldsymbol{\beta}|\times|\boldsymbol{\theta}|\times|\boldsymbol{\psi}|} \rightarrow \mathbb{R}^{3N}$, which takes shape $\boldsymbol{\beta}$, pose $\boldsymbol{\theta}$, and expression $\boldsymbol{\psi}$ as inputs and outputs a face mesh with N vertices. On this basis, some off-the-shelf 3D face estimators, such as DECA [13] and EMOCA [7], achieve 3D face reconstruction by regressing individual-specific FLAME parameters from in-the-wild images. We utilize EMOCA to obtain a face mesh corresponding to the input image with enhanced expression consistency, and render it to pixel-aligned explicit conditions. Specifically, given a single image I, the coarse branch of EMOCA estimate its corresponding $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, $\boldsymbol{\psi}$, albedo $\boldsymbol{\alpha}$, spherical harmonic (SH) illumination coefficient $\boldsymbol{l}$, and camera c, which can be expressed as $E_c(I) \rightarrow (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{l}, \boldsymbol{c})$. The detailed branch further outputs a detail vector δ and computes the displacement map D in UV space, which is specifically expressed as $E_d(I) \rightarrow \boldsymbol{\delta}$ and $F_d(\boldsymbol{\delta}, \boldsymbol{\psi}, \boldsymbol{\theta}_{jaw}) \rightarrow D$.

3.2 Diffusion Autoencoders (Diff-AE)

Diff-AE [44] reformulates the traditional diffusion generation model into an autoencoder and captures high-level image semantics for DDIM conditioning, therefore supporting near-exact reconstruction and attribute manipulation of the input image. Specifically, Diff-AE uses a semantic encoder to generate a 512-dimentional latent code z with global semantics of the input image x_0. Then, z can be introduced to the reverse deterministic generative process of DDIM to obtain a noisy map x_T which captures the stochastic variations of x_0. Last, a conditional DDIM model decodes (z, x_T) to achieve a deterministic reconstruction of x_0. By linearly modifying the semantic latent code z, Diff-AE can manipulate diverse global semantic attributes, *e.g.,* age, gender, and hairstyle. We introduce Diff-AE to our DisControlFace as the reconstruction backbone and freeze its pre-trained weights to provide semantically deterministic DDIM conditioning during the training of explicit face control.

4 METHOD

In the pursuit of a robust one-shot explicit facial image editing, we propose a generative framework, namely DisControlFace, providing DDIM conditioning on disentangled face control between high-level semantics and explicit 3DMM parameters (shown in Figure 2). Specifically, we adopt a weight-frozen Diff-AE as a semantic reconstruction backbone and construct an Exp-FaceNet to provide explicit parametric face control (Sec. 4.1). Furthermore, we design a training strategy to effectively enable the training of Exp-FaceNet (Sec. 4.2). Finally, we adopt an one-shot fine-tuning to improve the semantic consistency and faithfulness of the generated facial image under the subject-agnostic editing scenario (Sec. 4.3).

4.1 Exp-FaceNet

On the basis of the adopted Diff-AE reconstruction backbone, an intuitive idea for learning explicit face editing capability is to further build additional DDIM conditioning on 3DMM parameters. Compared to directly adopting the 3DMM parameters as non-spatial control conditions, generating pixel-aligned conditional maps based on those parameters is more conducive and compatible to convolution-based visual representation, which also helps enable fine-grained spatial control for the denoising diffusion process. Given this, we specially construct Exp-FaceNet, an explicit control network compatible with the adopted Diff-AE reconstruction backbone to perform a disentangled spatial conditioning for DDIM-based generation. Here we firs estimate 3DMM parameters from facial images

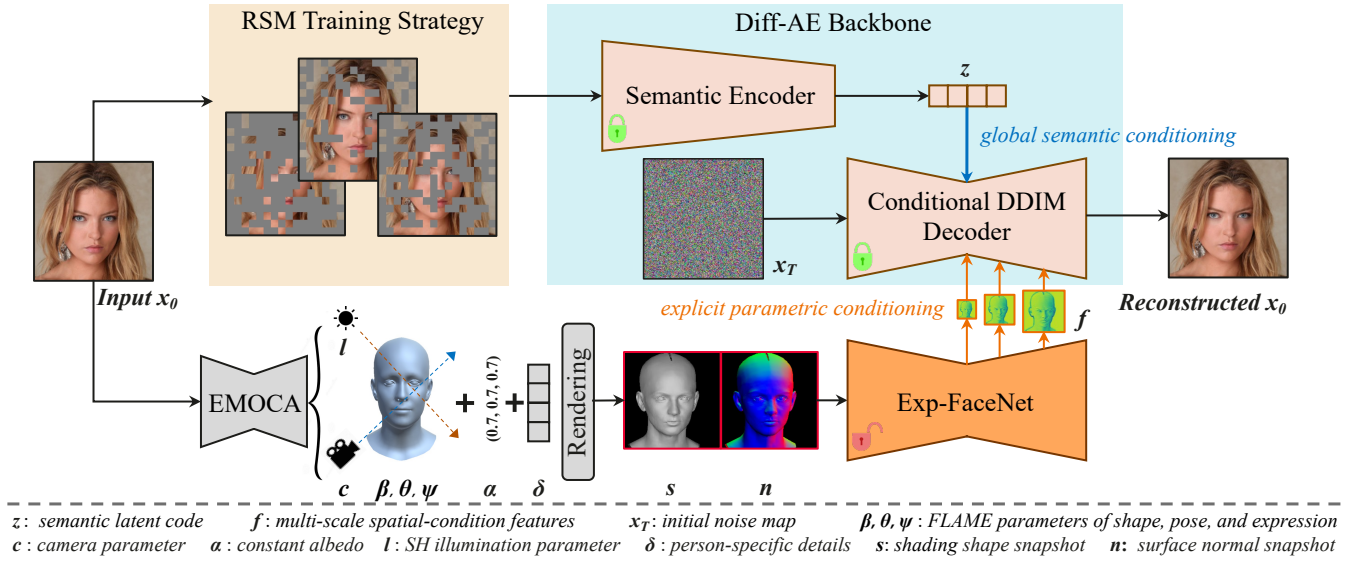

Figure 2: Pipeline overview. Our DisControlNet leverages Diffusion Autoencoder (Diff-AE) as the reconstruction backbone freeze its pre-trained weights to maintain the semantic deterministic conditioning capability, which is effective in reducing semantic information shift during the editing of the input portrait image. Then, an explicit face control network, Exp-FaceNet compatible with the Diff-AE is constructed, which takes pixel-aligned snapshots rendered from estimated explicit parameters as inputs and generates multi-scale control features to condition the DDIM decoder. Moreover, a random semantic masking (RSM) training strategy is accordingly designed to enable a disentangled explicit face control of Exp-FaceNet.

and transfer them to the corresponding visual guidance map. Specifically, we use EMOCA [7] to predict FLAME parameters (including shape β, pose θ, and expression ψ), SH illumination parameter l, and camera parameter c from an input portrait. Different from those previous work [12, 17, 43], here we also adopt the person-specific detail vector δ estimated by EMOCA, which can be combined with θ and ψ to generate the expression-dependent displacement map for refining the face geometry with animatable wrinkle details. To avoid undesired disturbance to facial appearance priors caused by inaccurate and unrealistic appearance estimation, we set the albedo map α to a constant gray value, thereby focusing on controlling the edit of shape, pose, expression, and lighting. We render these explicit parameters into a surface normal snapshot n:

$$n = \mathcal{R}(\mathcal{G}(\mathcal{M}(\beta, \theta, \psi)), c, \delta) \tag{1}$$

where the FLAME model $\mathcal{M}$ is used to calculate the 3D face mesh, $\mathcal{G}$ and $\mathcal{R}$ indicate normal calculation function and the Lambertian reflectance renderer, respectively. By means of this, n can reflect the fine-grained geometry of the input face and is compatible with pose parameter θ and expression parameter ψ. Furthermore, we also generate a shading shape snapshot s to illustrate the lighting conditions associated with the SH illumination parameter l:

$$s = \mathcal{R}(\mathcal{M}(\beta, \theta, \psi), \alpha, l, c, \delta) \tag{2}$$

Considering U-Net [46] excels at extracting spatial features from images for various vision tasks, we construct Exp-FaceNet in a similar U-shape structure, which takes channel-concatenated snapshots n and s as the input visual guidance map and generates multi-scale deep features for spatial-aware conditioning. Then we feed back the spatial-condition features outputted by each stage of the U-Net's decoder back to Diff-AE to provide multi-scale conditional control. Please see the detailed architecture in the supplement.

4.2 RSM Training Strategy

DisControlFace can be regarded as a DDIM model that is simultaneously conditioned by a global semantic code and multi-scale explicit-control feature maps (see Figure 2). As mentioned before, we freeze the pre-trained weights of Diff-AE to preserve global semantics and enable Exp-FaceNet to learn explicit face control in a disentangled way. However, since Diff-AE backbone already allows a deterministic image reconstruction under this setting, only limited gradients can be generated during error back-propagation, which are far from sufficient to effectively train Exp-FaceNet. Consequently, it is infeasible to train Exp-FaceNet in a traditional conditional DDIM generation form. To address this issue, we design a random semantic masking (RSM) strategy, not for the purpose of representation learning like Masked Autoencoders (MAE) [20], but rather to achieve the training of Exp-FaceNet effectively. Concretely, we divide the input image x_0 into regular non-overlapping patches and randomly mask different portions of patches to obtain a masked image x_0^m at different timesteps. By means of this, the semantic latent code z^m encoded by the semantic encoder $\mathcal{E}_\eta$ of Diff-AE only contains fragmented and incomplete content and spatial information of x_0. Meanwhile, the spatial-condition features f generated by Exp-FaceNet $\mathcal{F}_\phi$ comprise the fine-grained face shape as well as the camera parameter and lighting condition of x_0, which can help to restore the masked face regions in x_0^m in each random denoising timestep. The overall training objective can thereby be

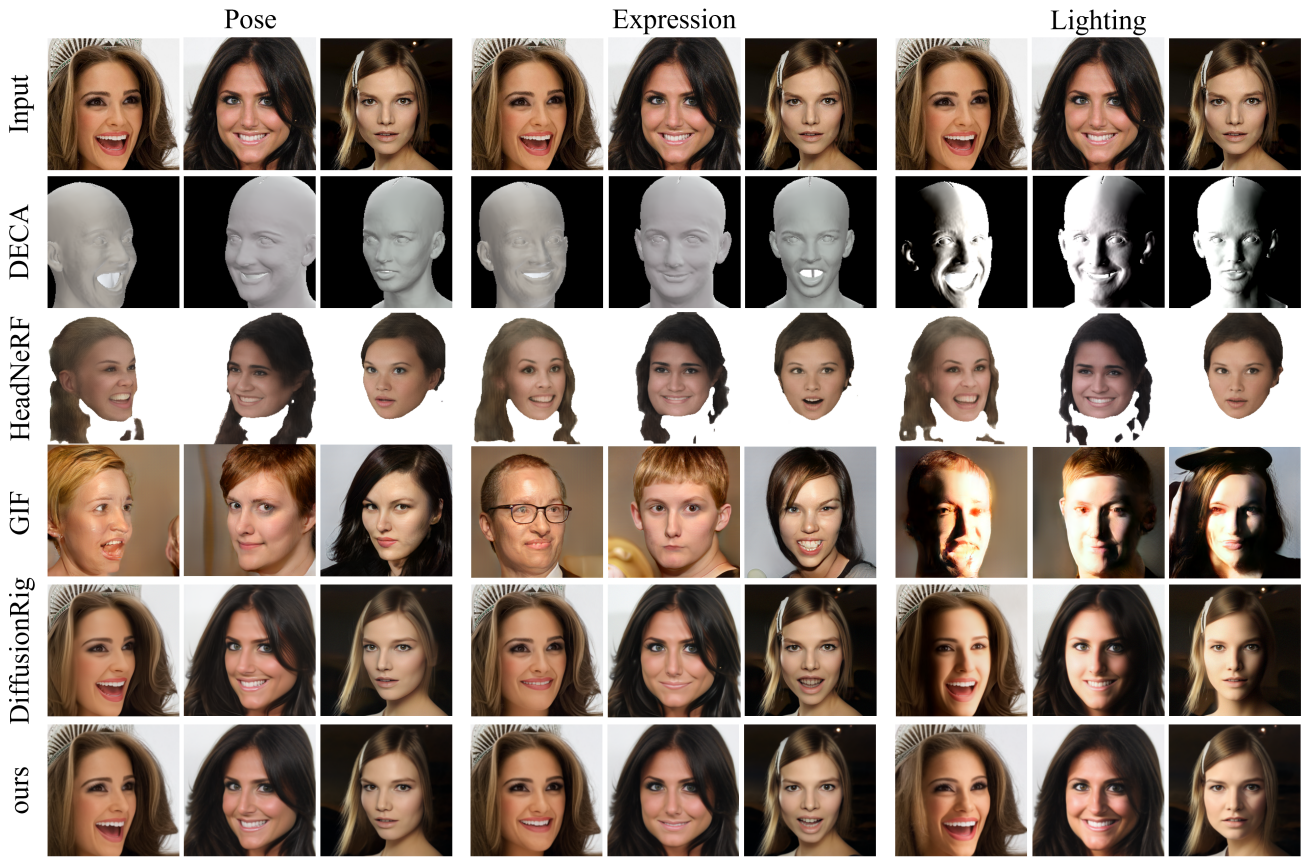

Figure 3: Qualitative comparison against baselines in one-shot editing. For each selected image, we use EMOCA [7] to estimate the corresponding explicit parameters, then synthesize the edited images using different methods based on the modified parameters of pose, expression, and lighting. We additionally provide the rendered shading shapes in the second row as the references of explicit control conditions. As can be seen, our DisControlFace can edit images that match well with the target control conditions while faithfully synthesizing facial appearances and editing-irrelevant details.

parameterized as:

$$\mathcal{L} = \mathbb{E}_{x_0, x_0^m, t, \epsilon}[\|\epsilon - \epsilon_\theta(x_t, t, z^m, f)\|_2^2] \tag{3}$$

where ϵ_θ is the U-Net of Diff-AE which predicts the noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ added in noisy image x_t. Throughout the generalized training, we only train $\mathcal{F}_\phi$ and freeze ϵ_θ and $\mathcal{E}_\eta$ with the pre-trained weights.

4.3 Exploiting One-shot Semantic Priors

At this point, the well-trained Exp-FaceNet is able to explicitly change pose, expression, and lighting of a facial image. However, there still exists identity shift or background changes during the face editing, which is especially evident when the target face for editing lies outside the domain of the pre-trained Diff-AE. Given this, it is meaningful to fully exploit the semantic priors of the to-be-edit image and inject them into the editing process. Concretely, in this stage, we freeze Exp-FaceNet and only fine-tune Diff-AE with the input portrait image using the aforementioned RSM training strategy. As a result, this one-shot fine-tuning enables the model to faithfully restore the personalized appearance details as well as editing-irrelevant factors such as background and accessories when performing explicit face editing.

4.4 Inference Editing

In practice, we first use EMOCA to predict all explicit parameters (mentioned in Sec. 4.1) of the input portrait x_0, then we modify the pose parameters θ, expression parameter ψ, and SH light parameter l by manually setting target values or directly transferring these parameters from a driving portrait. After this, we calculate the rendered shading shape snapshot s and surface normal snapshot n based on the modified parameters, and further generate explicit control conditions using Exp-FaceNet. On the other hand, since we should keep a consistent generative mechanisms in training and inference, here we also utilized masked input images to provide high-level semantic conditioning for DDIM decoder and accordingly design an intuitive patch masking strategy for inference editing.

	ID ↑	Shape↓	Pose ↓	Exp ↓	Light ↓
GIF [17]	0.22	3.0	5.6	5.0	0.40
DiffusionRig [12]	0.24	4.3	4.2	2.8	0.36
Ours	0.31	2.8	4.5	2.9	0.31

Table 1: Quantitative comparisons against compared baselines using identity consistency (ID) and DECA re-inference errors on shape, pose, expression, and lighting.

Concretely, for each timestep t, we generate the masked image x_t^m by randomly masking the patches of the input image x_0 with a linear ratio $\rho_t = 0.75 - 0.5(T-t)/T$, where the number of the inference denoising steps T is set to 20 in this work. Under this setting, we can generate x_t^m with high masking ratios ρ_t to emphasize the facial control of the intermediate denoising result z_{t-1} in the early stage of the inference, and then gradually decrease ρ_t to recover semantic information.

5 EXPERIMENTS

5.1 Implementation Details

We train the proposed Exp-FaceNet on the FFHQ dataset [26], which consists of 70k in-the-wild facial images. For evaluations, we select the images of the CelebA-HQ dataset [25] to perform one-shot fine-tuning and face editing. To balance generation quality and computational cost, we resize the images to a resolution of 256×256 for both training and inference. Accordingly, we utilize Diff-AE[1] pre-trained on FFHQ-256 for all experiments. We train Exp-FaceNet for 437,500 iterations, with a learning rate of $1e^{-4}$ and a batch size of 32, while during the one-shot fine-tuning stage, we only fine-tune the pre-trained Diff-AE for 1,500 iterations, with a learning rate of $1e^{-5}$ and a batch size of 4. In all training stages, we use AdamW [32] as the optimizer and set the denoising timesteps to 1,000.

5.2 Comparison

Baselines. We compare our methods against three generative methods for parametric face image synthesis and editing: HeadNeRF [23], GIF [17], and DiffusionRig [12]. Among these, HeadNeRF is a NeRF-based head model, while GIF and DiffusionRig are both generative face models built upon GAN and diffusion model, respectively. For a fair comparison, we utilize the models pre-trained on FFHQ at a resolution of 256^2 for each method.

Qualitative comparison. We evaluate our DisControlFace against baselines using images from CelebA-HQ and we perform one-shot fine-tuning for all methods. The qualitative comparison results are provided in Figure 3. For all methods, we visualize the editing results of pose, expression, and lighting on three identities. In order to intuitively measure the editing performance, we further give the rendered snapshot s of the shading shape corresponding to the modified explicit parameters as the references of explicit control conditions. (second row of Figure 3). We can find that compared to the other methods, our method synthesizes images with overall best identity consistency and parametric editing accuracy. Specifically, HeadNeRF cannot generate realistic facial appearance as well as the background of the original image. GIF has a good control ability of parametric face editing, however, is completely unable to preserve the face identity. DiffusionRig, another diffusion-based method achieves better editing results than HeadNeRF and GIF, especially in identity preservation, which can also be attributed to using one-shot fine-tuning to boost the diffusion model with more personal facial priors. Nevertheless, since DiffusionRig trains the whole model together from scratch which prevents the model from learning robust disentangled control, the edited results still have visible identity shift (*e.g.*, eyes, skin color, and hair) especially in lighting editing.

[1]https://github.com/phizaz/diffae.

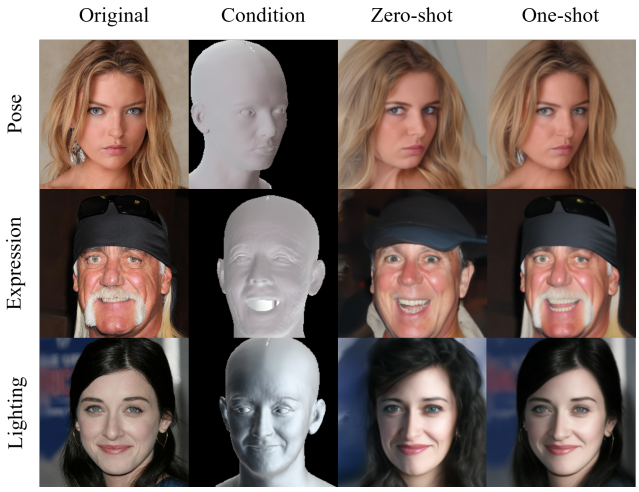

Figure 4: Ablation study on one-shot fine-tuning. DisControlFace can perform zero-shot explicit editing, where however, identity shift still exists. On this basis, adopting a simple one-shot fine-tuning can significantly improve the preservation of face identity as well as other editing-irrelevant semantic information.

Quantitative comparisons. Table 1 (top part) provides the quantitative editing results of all methods on 1000 in-domain images of FFHQ. Following previous work [9, 12, 17], we apply DECA re-inference on edited images and calculate the Root Mean Square Error (RMSE) between the input and re-inferred face vertices as well as spherical harmonics to evaluate the editing accuracy on shape, pose, expression, and lighting. We additionally measure the identity consistency (ID) score by computing the cosine similarity between the deep features generated by ArcFace [8] of the original and edited images. The results indicate that our method achieve the overall best performance on all metrics. Note that DisControlFace outperforms other methods by a large margin in ID score, which demonstrates the superiority of our method in identity preservation during one-shot editing.

5.3 Ablation Study

One-shot fine-tuning. Based on the deterministic semantic reconstruction capability of Diff-AE [44], DisControlFace can extract global semantics of the input face image. However, since the target

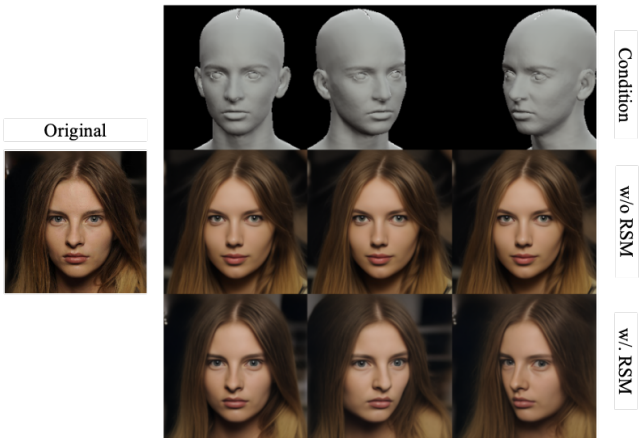

Figure 5: The necessity of the proposed random semantic masking (RSM) training. Without RSM training, it is infeasible to train Exp-FaceNet with explicit face control.

Figure 6: The ablation studies on our disentangled pipeline (a) and Exp-FaceNet structure (b). The disentangled control setting in DisControlFace trained with RSM strategy can significantly improve the identity preservation in explicit editing. Compared to adopting ControlNet with a light-weight decoder and zero convolutions, using our Exp-FaceNet can improve the explicit control accuracy by a large margin.

portrait image tends to have different face priors with pre-trained Diff-AE, there still exists semantic information shift under the zero-shot editing scenario, as shown in Figure 4. Given this, we adopt a simple one-shot fine-tuning on the pre-trained Diff-AE to fully exploit the semantic priors of the to-be-edit image and inject them into the editing process. We can observe that by means of this, the face identity and other high-level semantic information (*e.g.,* background, accessories, and hair) can be well preserved.

Effectiveness of RSM training. In this paper, we achieve a disentangled training of the proposed Exp-FaceNet by freezing the pre-trained weights of Diff-AE and according designing a RSM training strategy. To demonstrate the necessity of the proposed RSM training, we separately train Exp-FaceNet with and without random patch masking for the input image of semantic encoder of the Diff-AE backbone. Figure 5 shows that both training strategies enable the model to reconstruct the input image. However, only the model trained with PSM strategy can generate images with novel poses. This result is consistent with our claim that since the pre-trained Diff-AE backbone can already allow deterministic image reconstruction, limited gradients can be generated during error back-propagation for an effective training of Exp-FaceNet.

Figure 7: Visualization of disentangled face control. We separately utilize the encoded global semantic code of one image and the estimated 3DMM parameters of another image to provide semantic conditioning and explicit parametric conditioning in facial image generation.

Effectiveness of our disentangled pipeline To further demonstrate the advantages of our method over traditional pipeline in terms of disentangled control, we train the whole model (Diff-AE+Exp-FaceNet) from scratch without using RSM training. The result in Fig. 6 (a) shows that our disentangled pipeline can significantly reduce identity shift, both visually and in terms of quantitative metrics.

Exp-FaceNet structure. Our Exp-FaceNet is inspired from ControlNet [65], a popular deep network which has been widely employed to add various spatial visual guidance (*e.g.,* edge maps, pose maps, depth maps, *etc.*) to Stable Diffusion (SD) [45] for text-to-image generation. However, there exists many differences between two models. First, ControlNet is specially designed for SD which can generate specific content based on the input prompts, while there still exists uncertainty and randomness in the generation controlled by visual guidance and text prompts. In contrast to this, this paper focuses on semantics preservation and explicit editing of the input portrait image. Given this, our Exp-FaceNet is designed to be compatible with a Diff-AE backbone, aiming to provide semantically deterministic DDIM conditioning. Second, we construct Exp-FaceNet as a U-Net instead of adopting zero convolutions introduced in ControlNet. With this setting, we can endow Exp-FaceNet with strong capability of U-shape models in extracting spatial deep features from pixel-aligned visual guidance map (rendered explicit snapshots in this work). Here we compare the generation performance between using Exp-FaceNet and ControlNet to learn explicit face control. The results in Fig. 6 (b) show that compared to ControlNet, our Exp-FaceNet can help to edit the face image with less DECA re-inference error and visualized better identity consistency.

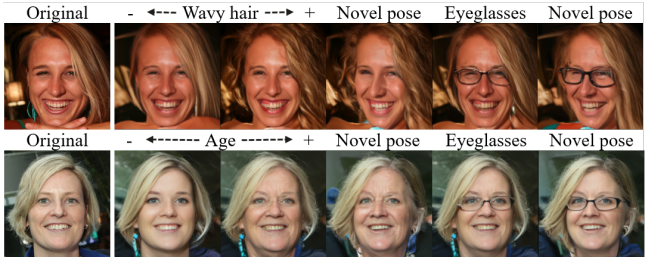

Figure 8: Synthesis results in semantic manipulation. We follow Diff-AE [44] to manipulate the global facial attributes (age, hairstyle, and eyeglasses) by linearly editing z. Owing to the disentangled control mechanism, our model can simultaneously perform semantic manipulation and explicit editing of the input facial image.

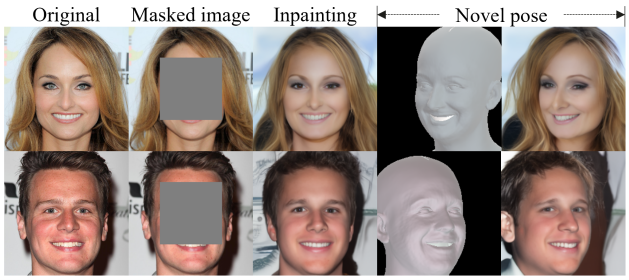

Figure 9: Zero-shot inpainting and subsequent explicit editing on images from CelebA-HQ. DisControlFace can restore the masked face regions smoothly and naturally. On this basis, Our method can further edit the restored facial images based on the modified explicit 3DMM parameters.

5.4 Visualization of Disentangled Control.

To further analyze the disentangled conditioning of our DisControlFace, we show the synthesis results of mixed conditioning generation in Figure 7. Specifically, we replace the encoded global semantic code of facial image A with global code extracted from another image B, while fixing all the estimated all explicit face parameters of A. We can observe that the synthesis face has the same face shape, pose, expression, and lighting condition with image A. Meanwhile, all facial appearance priors like skin color, eyes, and lips color as well as non-facial high-level semantics such as background and hair of image B have been successfully transferred to the synthesized image. Note that we can achieve robust facial semantics transfer without requiring a few-shot fine-tuning on the personalized images of an specific individual as adopted by DiffusionRig [12]. This result can intuitively show the effectiveness of the disentangled face control capability of DisControlFace. Also benefiting from this, the proposed DisControlFace can preserve original high-level semantics while performing fine-grained explicit face editing.

5.5 Applications

Semantic manipulation. Since we adopt Diff-AE [44] as the reconstruction backbone and freeze the pre-trained weights without tuning, our DisControlFace inherits the encoding capability of global facial semantics. Therefore, compared to previous diffusion-based model such as DiffusionRig [12], our model also supports manipulating face attributes (*e.g.,* hairstyle, age, and accessories) of the input portrait by linearly editing global semantic codes. The visualized results of both semantic manipulation and explicit editing are shown in Figure 8 , which once again demonstrate the effectiveness and flexibility of the proposed disentangled conditional generation mechanism.

Image inpainting. Benefiting from the proposed RSM training strategy, our model inherently supports image inpainting. Figure 9 shows the results of zero-shot inpainting on center-masked facial images as well as the subsequent editing. We can observed that the restored face in the inpainted image is smooth and natural, also has good similarity with the original face. Besides, the explicitly edited image still shows a consistent identity with the restored facial image. Meanwhile, since it is inappropriate to use masked images to perform one-shot fine-tuning on Diff-AE, some editing-irrelevant high-level semantics such as hairstyle and background might not be well preserved in the generated images.

6 LIMITATIONS AND CONCLUSION

Limitations and future work. In this work, both adopted Diff-AE and constructed Exp-FaceNet were trained using the FFHQ dataset, which consisting 70,000 in-the-wild images. However, this data size is still not sufficient to train the model to learn a more generalized face priors. It can be expected that collecting much more face data with abundant face conditions (*e.g.,* pose and expression) for training can substantially improve the editing performance, especially for some challenging tasks like zero-shot explicit editing. We adopt EMOCA to estimate explicit face parameters of the input image and generate spatial-aware conditions representing control information based on them. However, EMOCA struggles to model the detailed geometry of eyeballs as well as some extreme expressions, which could hinder the model from restoring the corresponding facial details when generating edited images. Furthermore, our editing framework is constructed as a denoising diffusion pipeline, which therefore is still unable to compete with GANs in terms of generation speed. In the future, with the development of fast sampling algorithms, we expect the generation time of our model to further decrease.

Conclusion. We have presented DisControlFace, a novel diffusion framework for one-shot facial image editing. Through exploiting disentangled conditioning on high-level semantics and explicit 3DMM parameters in the generation process, our model excels in explicit and fine-grained face control while preserving semantic information and facial priors in face editing. This may boost a series of related applications including various semantic and explicit facial image editing, zero-shot image inpainting, and cross-identity face driving.

REFERENCES

[1] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 843–852.

[2] 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.

[3] Volker Blanz and Thomas Vetter. 2023. A morphable model for the synthesis of 3D faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 157–164.

[4] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 20, 3 (2013), 413–425.

[5] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16123–16133.

[6] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5799–5809.

[7] 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.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4690–4699.

[9] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. 2020. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5154–5163.

[10] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. 2022. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10673–10683.

[11] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. 2021. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12819–12829.

[12] 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.

[13] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–13.

[14] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.

[15] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–13.

[16] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. 2018. Morphable face models-an open framework. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 75–82.

[17] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. 2020. Gif: Generative interpretable faces. In *Proceedings of the 2020 International Conference on 3D Vision (3DV)*. IEEE, 868–878.

[18] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985* (2021).

[19] Shalini Gupta, Kenneth R Castleman, Mia K Markey, and Alan C Bovik. 2010. Texas 3D face recognition database. In *Proceedings of the IEEE Southwest Symposium on Image Analysis & Interpretation (SSIAI)*. IEEE, 97–100.

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16000–16009.

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 6840–6851.

[22] Fangzhou Hong, Zhaoxi Chen, LAN Yushi, Liang Pan, and Ziwei Liu. 2022. EVA3D: Compositional 3D Human Generation from 2D Image Collections. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

[23] 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.

[24] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Hk99zCeAb

[26] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4401–4410.

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8110–8119.

[28] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. 2023. 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 (CVPR)*. 394–403.

[29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (ToG)* 36, 6 (2017), 194–1.

[30] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–16.

[31] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Richard Zhang, and SY Kung. 2022. 3d-fm gan: Towards 3d-controllable face manipulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 107–125.

[32] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[33] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4460–4470.

[34] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[35] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. 2019. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7588–7597.

[36] Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. 2022. Mystyle: A personalized generative prior. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–10.

[37] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. 2022. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 597–614.

[38] Xingang Pan, Xudong Xu, Chen Change Loy, Christian Theobalt, and Bo Dai. 2021. A shading-guided generative implicit model for shape-accurate 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 34. 20002–20013.

[39] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 165–174.

[40] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5865–5874.

[41] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14314–14323.

[42] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. 2020. Convolutional occupancy networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 523–540.

[43] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. 2023. DiFaReli: Diffusion Face Relighting. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[44] 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.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 234–241.

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22500–22510.

[48] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 45, 4 (2022), 4713–4726.

[49] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33. 20154–20166.

[50] Matan Sela, Elad Richardson, and Ron Kimmel. 2017. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1576–1585.

[51] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. 2020. Implicit neural representations with periodic activation functions. In *Advances in neural information processing systems (NeurIPS)*, Vol. 33. 7462–7473.

[52] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. 2019. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2437–2446.

[53] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[54] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. 2023. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20991–21002.

[55] Keqiang Sun, Shangzhe Wu, Zhaoyang Huang, Ning Zhang, Quan Wang, and Hongsheng Li. 2022. Controllable 3d face synthesis with conditional generative occupancy fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 16331–16343.

[56] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2019. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10812–10822.

[57] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. 2021. Learning complete 3d morphable face models from images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3361–3371.

[58] Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1126–1135.

[59] Luan Tran and Xiaoming Liu. 2019. On learning 3d face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 43, 1 (2019), 157–171.

[60] Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. 2023. Face0: Instantaneously Conditioning a Text-to-Image Model on a Face. *arXiv preprint arXiv:2306.06638* (2023).

[61] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. DisCo: Disentangled Control for Referring Human Dance Generation in Real World. *arXiv preprint arXiv:2307.00040* (2023).

[62] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. 2022. Anifacegan: Animatable 3d-aware face image generation for video avatars. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 36188–36201.

[63] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 601–610.

[64] Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. 2023. Inserting Anybody in Diffusion Models via Celeb Basis. *arXiv preprint arXiv:2306.00926* (2023).

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3836–3847.

[66] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13545–13555.