Single Image, Any Face: Generalisable 3D Face Generation

Wenqing Wang, Haosen Yang, Josef Kittler, and Xiatian Zhu

University of Surrey, United Kingdom

Abstract. The creation of 3D human face avatars from a single unconstrained image is a fundamental task that underlies numerous real-world vision and graphics applications. Despite the significant progress made in generative models, existing methods are either less suited in design for human faces or fail to generalise from the restrictive training domain to unconstrained facial images. To address these limitations, we propose a novel model, **Gen3D-Face**, which generates 3D human faces with unconstrained single image input within a multi-view consistent diffusion framework. Given a specific input image, our model first produces multi-view images, followed by neural surface construction. To incorporate face geometry information in a generalisable manner, we utilise input-conditioned mesh estimation instead of ground-truth mesh along with synthetic multi-view training data. Importantly, we introduce a multi-view joint generation scheme to enhance appearance consistency among different views. To the best of our knowledge, this is the first attempt and benchmark for creating photorealistic 3D human face avatars from single images for generic human subject across domains. Extensive experiments demonstrate the superiority of our method over previous alternatives for out-of-domain singe image 3D face generation and top competition for in-domain setting.

Keywords: 3D Head Generation · Multi-view Diffusion · Novel View Synthesis

1 Introduction

Generating photorealistic 3D face avatars from a single image input benefits a wide range of real applications in computer graphics and computer vision, e.g. video conferencing, virtual modeling, entertainment, augmented and enhanced reality [13,36,53]. The majority of existing 3D face modelling methods not only need costly per-identity optimisation, but also demand a large amount of input in the form of text [17,64], and multi-view images or videos [21,39,60,69]. Text-guided 3D avatar generation often struggles to ensure authenticity and identity control, as it faces the daunting task of accurately capturing human identity and face appearance in high detail, unlike image/video-based approaches. On the other hand, the latter typically rely on multiple view calibrated images,

Fig. 1: 3D human face avatar from (a) a single unconstrained image by (c) prior art model [7] (note newly added hat and clear identity shift), vs. (d) our model.

making them less useful and applicable in practice as in many situations such input data is just unavailable.

Inspired by the remarkable success of generative diffusion models [20, 46] and driven by the aforementioned challenges, *single-image 3D face generation* has become a trendy topic with the key challenges in figuring out both geometry and appearance information from only a single face image of a generic human identity. This seemingly impossible task now becomes hopeful for two reasons: *The first* lies in the availability of unprecedentedly rich and comprehensive knowledge captured by off-the-shelf generative models, providing a chance of extracting and transferring useful information for particular downstream tasks (human face in this work) [29, 38, 54]. For example, Stable Diffusion was trained with a massive (unknown) text-image pairs from the Internet, including a diversity of facial images from a broad range of subjects like the celebrities [67]. *The second* is the enormous technical advance in multi-view image generation [30, 31, 48, 51], and 3D representation, reconstruction, and generation [24, 34, 57]. Combining these building blocks all together properly could be the basis of plausible solutions to tackling this challenge.

Building on the pillars discussed above, an intuitive approach is to learn a generic 3D face generation model from a large, diverse collection of data with multi-view images per human identity, so that the model could generalize to generic unseen single face images. There are some early attempts pursuing this strategy by training on large synthetic digital avatars created by 3D artists [55]. This however raises a synthetic domain to real domain generalisation challenge, resulting in unrealistic face generation. Besides, the collection of human face data is much more restricted, due to both the intrinsic complexity and diversity, as well as the intricate privacy considerations. As a result, existing 3D face benchmarks are often limited in size and diversity in practice, e.g. containing only a few hundreds identities [25, 35, 61], making them insufficient for model training.

To mitigate this data scarcity challenge, the latest attempt for single-image 3D face generation leverages the human geometric priors by incorporating ground-truth mesh in multi-view synthesis [7]. A promising finding from this work is

that properly blending image appearance and mesh's geometric knowledge enables the model to work across different views at good quality. However, we find that their method suffers from several limitations that significantly hampers its generalisation to unconstrained face images shown in Figure 1: (i) *Over reliance on the ground-truth mesh*, which is often unavailable in practice; (ii) *Overfitting to the training domain* due to the stringent need with training data, thus limited data availability, so that the model just cannot generalise to different unseen styles.

In this work, to overcome these limitations we propose a novel diffusion generative approach, **Gen3D-Face**, for more generalisable 3D face generation using unconstrained single images. Our model first generates consistent multi-view images and then conducts the neural surface construction. Instead of requiring ground-truth mesh, we exploit input-conditioned mesh estimation for not only mitigating the model's over reliance on the geometric prior, but also enabling the model to generalise to typical cases without the ground-truth mesh, and with distinct appearance styles. To enhance data diversity, we further generate synthetic 3D face images with off-the-shelf model [2]. To improve multi-view consistency, we introduce a multi-view joint generation scheme.

Our **contributions** are summarised as follows: **(1)** Investigating the understudied single-image 3D face generation problem with a particular focus on the ability of generalising the model to unconstrained unseen face imagery so that the developed more would be more practically useful and deployable. To the best of our knowledge, this is the very first attempt at tackling this meaningful problem setting in single image 3D face generation. **(2)** A novel approach Gen3D-Face characterised by generalisable incorporation of face geometric priors, multi-view joint generation, and joint mining of both real and synthetic 3D face data. **(3)** Extensive evaluation on generalised single image 3D face generation demonstrating the superior performance of our model over the state-of-the-art alternatives.

2 Related Work

Novel view synthesis Neural fields [21, 34, 59] and 3D Gaussian Splatting [24, 66] have emerged as the most effective 3D object and scene representations, capable of producing photorealistic images from arbitrary novel views of a scene. However, the first generation is reconstruction-based, necessitating densely captured views. To relax this assumption, follow-up approaches [22, 44, 50, 63] propose learning-based methods requiring only a few views, by utilising scene priors from other existing datasets [63], or explicitly mapping the input image to one 3D Gaussian per pixel [50]. Commonly, these methods tend to be restricted to reconstructing relatively simple objects, or confined to low-resolution, due to their limited expressive capacity.

3D avatars from a single image In addition to reconstruction techniques, various methods have been developed to generate 3D avatars utilising Generative Adversarial Networks (GANs) [2, 4, 11] or more recently diffusion models [20, 46]. 3D-aware GANs learn 3D representation by integrating tri-planes [4, 11] or tri-

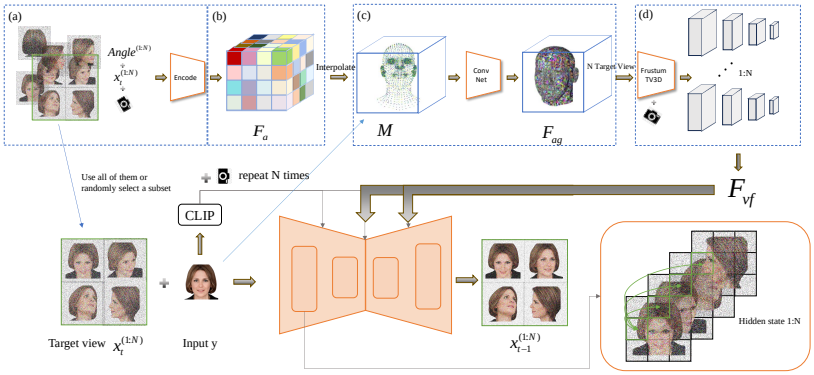

Fig. 2: Overview of our Gen3D-Face method. It adopts the latent diffusion paradigm involving the learning of multi-step denoising. Each step denoises N novel views with condition on a single face image y and the mesh $\mathbb{M}$ estimated from y, following the process as below: (a) Using an encoder to integrate the previous step's noise multi-view images $\mathbf{x}_t^{(1:N)}$ with with camera angles and time embedding; (b) Interpolating its output with a predefined 3D voxel to obtain the *appearance feature volume* F_a; (c) Further combining F_a with the geometry prior $\mathbb{M}$ to obtain the *hybrid feature volume* F_{ag}; (d) Finally obtaining the denoised views $\mathbf{x}_{t-1}^{(1:N)}$ by injecting F_{ag} to a light 3D CNN to get view frustum volume F_{vf}, which into the diffusion backbone as the condition.

grids [2] combined with camera position. To achieve a single image to 3D avatar generation, typically GAN inversion to fit the input image is required, which is computationally expensive and time-consuming. Live3D [53] train an image-to-triplane encoder to map an unposed image to a canonical triplane 3D representation instead of GAN inversion, while still limited in large output angles. On the other hand, diffusion methods specifically designed for human avatar suffer from limited training data [1,7,47], as a 3D diffusion model is hard to learn from 2D image collections. Therefore, these methods rely on pretrained models [29,46] and incorporate 3D physical constraints [28,32,61] as prior knowledge. However, their stringent input requirements significantly restrict their ability to generalise across out-of-domain face images and the situations without ground-truth mesh. In this work, we tackle these challenges with proper model design and data synthesis.

Multi-view diffusion models Recent works [5,16,29–31,48,52] extend 2D diffusion models to generate consistent multi-view images from single-view images. Their success benefits from the existence of large-scale 3D datasets [8, 9]. Expanding this direction, our work focuses on human face avatar generation with special requirement on model generalisation to unconstrained imagery.

Learning from synthetic data Photorealistic synthetic data is effective in handling data scarcity [18, 41, 56, 62]. Recent methods have been developed to utilise synthetic data, either explicitly [26] or implicitly [53], to enhance perfor-

mance in generative tasks. In this work, we extend and validate this generic idea for more challenge single image 3D face generation in unconstrained settings.

3 Method

Given a single face image y as input, we aim to generate a 3D face avatar for this person. To that end, we propose a new latent diffusion approach, **Gen3D-Face**, with the architecture depicted in Figure 2. It generates multi-view consistent images from the single image, which can then be fed into existing neural surface construction methods (e.g. [57]). For the former, we adopt the off-the-shelf Stable Diffusion [46] as the backbone where the diffusion and denoising take place in a latent feature embedding space (e.g. a pretrained VAE [43]). For self-containing, we first brief 2D diffusion and 3D diffusion next.

3.1 Preliminaries: 2D Diffusion and 3D Diffusion

Diffusion models [20,46] aim to gradually generate structured outputs of a target distribution from random noise through learning an iterative denoising model. Given a noise input x_t, where $t \in (0, T)$ denotes the steps with a total of T, the model is trained to predict the added noise, with which removed, a less noisy version x_{t-1} can be revealed. Whilst these models can generate novel-view images, it is shown that multi-view consistency is hard to be maintained [29].

To address this issue, multi-view diffusion has been recently developed [30, 48]. The key idea is to jointly denoise the images for multiple predefined viewpoints with condition on the same input y, so that a conditional joint distribution of all these views $p_\theta(x_0^{(1)}, \cdots, x_0^{(N)}|y)$ can be learned instead, where N specifies the view number. The forward process adds noise to every viewpoint independently at time t, and the reverse process is constructed as:

$$p_\theta(\mathbf{x}_{0:T}^{(1:N)}) = p(\mathbf{x}_T^{(1:N)}) \prod_{t=1}^{T} \prod_{n=1}^{N} p_\theta(\mathbf{x}_{t-1}^{(n)}|\mathbf{x}_t^{(1:N)}), \tag{1}$$

where the per-step per-view denoising is formulated by a Gaussian distribution:

$$p_\theta(\mathbf{x}_{t-1}^{(n)}|\mathbf{x}_t^{(1:N)}) = \mathcal{N}(\mathbf{x}_{t-1}^{(n)}; \mu_\theta^{(n)}(\mathbf{x}_t^{(1:N)}, t), \sigma_t^2 \mathbf{I}), \tag{2}$$

where the learnable mean for the n-th view at step t is defined as:

$$\mu_\theta^{(n)}(\mathbf{x}_t^{(1:N)}, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t^{(n)} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta^{(n)}(\mathbf{x}_t^{(1:N)}, t) \right), \tag{3}$$

where $\epsilon_\theta^{(n)}$ denotes the trainable noise predictor for the n-th view, β_t specifies the noise schedule, α_t and $\bar{\alpha}_t$ are two scaling constants derived from β_t .

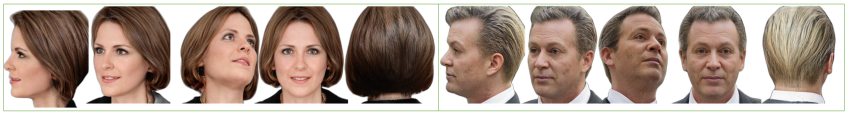

Fig. 3: Examples of synthetic face images.

3.2 Gen3D-Face

Extending prior multi-view diffusion, we take a step forward for generalisable single image 3D face avatar generation where single unconstrained face images are present without ground-truth mesh. To that end, we first need to address the data scarcity issue as discussed earlier by multi-view face images synthesis for training data augmentation.

Multi-view face synthesis We adopt the Panohead [2] to generate additional training images. We generated 25,000 virtual human identities, each represented by 48 images, with azimuth ranging from -180 to 180 degrees, and elevation angle from -40 to 40 degrees (see Figure 3).

As the synthesis process is less controllable, the output quality is often varying [26]. To filter out low-quality face images, we design a pruning process for dealing with two problems. (1) *The Janus problem*: We observe cases where the back-view images present blurry faces. To identify such cases, we construct a binary classifier with CLIP [42] using the class names as `back of human head` and `human front face`, and then classify all the back-view face images. We remove those back-view images with the score of the `human front face` class exceeding a threshold τ_{bv}. (2) *Identity inconsistency*: Multi-view face images generated by Panohead [2] are likely to be identity inconsistent. To detect this, we estimate the identity consistency using the average pairwise similarity of views with face embeddings [10] for every individual identity and keep only the top-τ_{ii} virtual identities for model training.

Face geometry prior To facilitate the 3D face modeling from single images, we integrate the human head mesh as a prior as [7]. A key difference however is that we suggest to use the estimated mesh from the input image, rather than requiring the ground-truth as [7]. The reasons are two-fold: (1) Often no ground-truth mesh in many real applications; (2) Using ground-truth mesh tends to make the model over rely on this prior, whilst largely ignore the appearance of the input image. Specifically, we estimate the FLAME mesh $\mathbb{M}$ with v vertices from a single image [12] during both training and inference. As we show in experiments, this design choice is a key to make our model more generalisable.

Joint conditioning of appearance and geometry The key in our context is how to effectively condition the multi-view diffusion process with both the appearance of the single image y and the geometry of the estimated mesh $\mathbb{M}$ (Sec. 3.1).

Specifically, let N noisy target views at time t denoted as $\mathbf{x}_t^{(1:N)}$ in our multi-view diffusion process. To impose viewpoint information, we deploy an encoder to project the camera angles and time embedding to the latent image space,

which is then added to each novel view's feature embedding $\mathbf{x}_t^{(1:N)}$ respectively. To represent these views in 3D space, we construct a 3D volume with its vertex $\mathbb{V} \in \mathbb{R}^{L\times L\times L}$ extracted by linear sampling along each dimension (where L is the number of voxels in each dimension). For each novel view n, we then warp $\mathbb{V}$ according to this view's extrinsic camera parameters into which the view's feature embedding $\mathbf{x}_t^{(n)}$ is interpolated. This results in an *appearance feature volume* F_a containing N noisy target view features.

To integrate the geometry prior from the estimated mesh $\mathbb{M}$, we adopt a sparse 3D ConvNet [15] to interpolate F_a with $\mathbb{M}$, leading to a *hybrid feature volume* F_{ag} with both appearance and geometry information. With F_{ag}, we produce the *view frustum volume* F_{vf} with a light FrustumTV3DNet [30]. This F_{vf} would then serve as a joint condition for multi-view diffusion by injecting into the backbone diffusion model (e.g., Stable Diffusion's UNet).

As seen from Eq (2), previous methods [7,30] often denoise a single view each time individually with condition on the previous step's output of all the views. This design requires N denoising times each for one view, which we consider is inferior in maintaining the view consistency. To overcome this issue, we propose a ***multi-view joint generation*** algorithm that instead denoises all the views concurrently at one time so that multi-view information interaction can be imposed and exploited. Specifically, instead of feeding one view $\mathbf{x}_t^{(n)}$ as the decoder's query at a time, we input all the views $\mathbf{x}_t^{(1:N)}$ together. This difference enables us to additionally perform the 3D self attention operation [3,48,68] among all the novel views $\mathbf{x}_t^{(1:N)}$ and the input y for information exchange and enhancing view consistency.

Model training Our objective function is a multi-view diffusion loss defined as

$$\ell(\theta) = \mathbb{E}_{t,y,c,\mathbf{x}_0^{(1:N)},(1:N),\epsilon^{(1:N)}}\left[\|\epsilon^{(1:N)} - \epsilon_\theta^{(1:N)}(\mathbf{x}_t^{(1:N)}, t)\|_2\right], \tag{4}$$

where y is the input image, c is the camera parameters, $\mathbf{x}_0^{(1:N)}$ denotes the N target-view images, $\epsilon^{(1:N)}$ is the added Gaussian noise, and $\epsilon_\theta^{(1:N)}$ is the noise predictor.

4 Experiments

Datasets: For model training, we use the 323 out of 359 identities from the Facescape dataset [61], following the setting of [7]. The same real training data is used for all compared models, whilst our model also uses synthetic data purposefully. For `out-of-domain` generalised evaluation, we randomly select 1,024 images from FFHQ [23] with background removed using [40]. For `in-domain` evaluation, as [7] we use the same 36 test identities.

Metrics: For generalised `out-of-domain` evaluation without access to the ground-truth images, we use four metrics: (1) Frechet Inception Distance (FID) [19], (2) Output-to-output ID consistency: calculated as the mean of Arcface cosine similarity [10] across all pairs of the 40 target views generated from the same input

Table 1: *Out-of-domain* single image 3D face generation results on FFHQ.

Method	FID↓	Output-to-Output ID Consistency↑	CLIP↑	Input-to-Output ID Consistency↑
Zero-1-to-3 [29]	78.8543	0.4483	0.5597	0.1300
SyncDreamer [30]	68.0294	0.4420	0.5883	0.1572
EG3D [4]	66.1578	0.4623	0.5142	0.1531
PanoHead [2]	58.1578	0.4821	0.5644	0.1611
Morphable Diffusion [7]	89.7443	**0.5171**	0.5359	0.1146
Gen3D-Face (Ours)	**55.4901**	0.4536	**0.6765**	**0.1716**

image, (3) Input-to-output ID consistency: averaging the Arcface cosine similarity [10] between the input image and all generated views, which we propose here to emphasise the importance of preserving the identity of input image, (4) CLIP Similarity [42]. For typical `in-domain` evaluation, following [7] we adopt four metrics: SSIM [58], LPIPS [65], FID [19], and face re-identification accuracy (Re-ID) [37], calculated between the ground truth and the generated images. For Re-ID metric, we consider two variants: (a) Re-ID(match): As [7], we calculate the percentage of matching the generated image with the ground truth at the Euclidean distance threshold of 0.6; (b) Re-ID(dist): The average Euclidean distance between the generated image and the ground truth, which offers additional insight into the quantity degree of matching.

Implementation: We use the AdamW [33] optimizer with a batch size of 8 for 48k iterations. The learning rate for training the backbone UNet has been raised from 1e-6 to 5e-5 with 100 warm-up steps [14], and is kept at 5e-4 for all other trainable modules. For inference, it takes about 25 seconds to generate 16 target views from a single input images using 50 DDIM [49] steps with an NVIDIA RTX 3090 GPU. We set $N = 16$ viewpoints, $\tau_{bv} = 0.93$ for back-view image filtering, $\tau_{ii} = 70\%$ for identity consistency filtering.

Competitors: We compare extensively with existing art nerf-based methods pixelNeRF [63], SSD-NeRF [6], and diffusion models including Era3D [27], Zero-1-to-3 [29], SyncDreamer [30], Morphable Diffusion [7], and GAN-based methods EG3D [4] and our data generator PanoHead [2]. Under the proposed out-of-domain setting, we exclude pixelNeRF [63] and SSD-NeRF [6] due to no precise camera parameters as required, and improve the generalisation of Morphable Diffusion [7] by using the FLAME [28] meshes obtained by fitting the ground truth 3D keypoints, otherwise (originally using ground truth bilinear meshes), it completely falls apart. All methods are fine tuned on in-domain training data except Era3D [27] as it claims that can have good results on human head.

4.1 Evaluations

Out-of-domain evaluation: From the quantitative results in Table 1, we observe that: **(1)** Interestingly, generic object diffusion models (Zero-1-to-3 [29], SyncDreamer [30]) and earlier GAN models (PanoHead [2]) even outperforms the latest face focused diffusion model (Morphable Diffusion [7]) on the three

out of four metrics. This is an *opposite phenomenon* under this more challenging setting as compared to what was discovered in [7], suggesting that the evaluation setting *matters*, fundamentally. This also implies that the way of imposing human geometry as [7] pays the generalisation cost implicitly in exchange of better performance for the limited in-domain setting. **(2)** Overall our Gen3D-Face is the best performer, except being secondary to Morphable Diffusion [7] on the output-to-output ID consistency metric. We note that looking at this metric *alone* however is not comprehensive and even misleading since it overlooked the divergence of generated images from the input (e.g. being consistent multi-view images of a totally difference identity). Instead, we should jointly consider both input-to-output and output-to-output ID consistency. Fusing the two metrics could make the comparison easier but hard to make sense.

Qualitative evaluation is presented in Figure 6 and Figure 8. We try to make sure that each method shows the same view for each row, but still have tiny difference even we give same evaluation camera views because different methods use different camera parameters for training, especially for Era3D [27]. We make these observations: **(1)** Zero-1-to-3 [29] tends to produce cartoon style images; **(2)** Era3D [27] is the newest single-image-to-3D method, we do not calculate the quantitative metrics because it can only generate 6 views, and it shows unrealistic geometry in visual. **(3)** SyncDreamer [30] and Morphable Diffusion [7] struggle in preserving the identity; **(4)** Morphable Diffusion [7] has more consistent generated images but suffers from overfitting to the training domain (e.g. added hat for all cases); **(5)** EG3D [4] and Panohead [2] tends to yield more blurry images, despite taking 20× more training time by PTI inversion [45]; **(6)** Our Gen3D-Face achieves the overall best result in terms of ID preservation and consistency, and wider pose variation.

In-domain evaluation: While this work stresses the importance of out-of-domain generalisation, we still evaluate the conventional in-domain setting. From Table 2 we observe that our method performs on par with the previous art model Morphable Diffusion [7]. This suggests that our model does not sacrifice the training domain performance while enhancing the model generalisation. The qualitative evaluations in Figure 7 to display our method keep good identity consistency.

Table 2: *In-domain* single image 3D face generation result on Facescape.

Method	SSIM↑	LPIPS↓	FID↓	Re-ID(match)↑	Re-ID(dist)↓
pixelNeRF [63]	0.7898	0.2200	92.61	0.9746	0.3912
Zero-1-to-3 [29]	0.5656	0.4248	10.97	0.9677	0.4193
SSD-NeRF [6]	0.7225	0.2225	34.88	0.9874	0.3855
SyncDreamer [30]	0.7732	0.1854	**6.05**	0.9960	0.3391
PanoHead [2]	0.7871	0.1914	7.10	0.9915	0.3412
Morphable [7]	**0.8064**	**0.1653**	6.73	**0.9986**	**0.3372**
Gen3D-Face (Ours)	0.7995	0.1701	6.1231	0.9981	0.3375

4.2 Ablation studies

Table 3: Ablation on the effect of synthetic and real training data.

Method	FID↓	Output-to-Output ID Consistency↑	CLIP↑	Input-to-Output ID Consistency↑
Real data only	71.2882	0.4317	0.5910	0.1428
Synthetic data only	105.5801	**0.5355**	0.5149	0.1011
w/o Data Pruning	57.3138	0.4451	0.6624	**0.1659**
w/ Data Pruning	**55.4901**	0.4536	**0.6765**	**0.1716**

Training data We evaluate the effect of synthetic and real training data. As shown in Table 3, we find that (1) both real and synthetic data contribute positively, and real data is more useful despite smaller size; (2) Using both could significantly boost the performance, validating our motivation of training data expansion by synthesis; (3) The significant decrease for only use synthesis data is because training and evaluating camera view not same.
Data pruning We show examples of Janus problem and Identity inconsistency in Figure 4, which are filtered out using our pruning process, and the effect of pruning shows in Table 3.

Fig. 4: Janus problem and Identity inconsistency in Synthesis Dataset.

Mesh prior effect We evaluate the effect of models trained with different mesh priors: (1) ground-truth fitted mesh [7], and (2) input-estimated FLAME mesh as proposed in our work. And provide different input images and the same mesh which is randomly chosen from the Facescape testing set to different mesh prior models separately. From Figure 5 we find that using ground-truth mesh will lead to the challenge of preserving the input identity and appearance information, and the tendency of overfitting to the training domain. In contrast, our idea can largely mitigate these issues, validating our design choice.

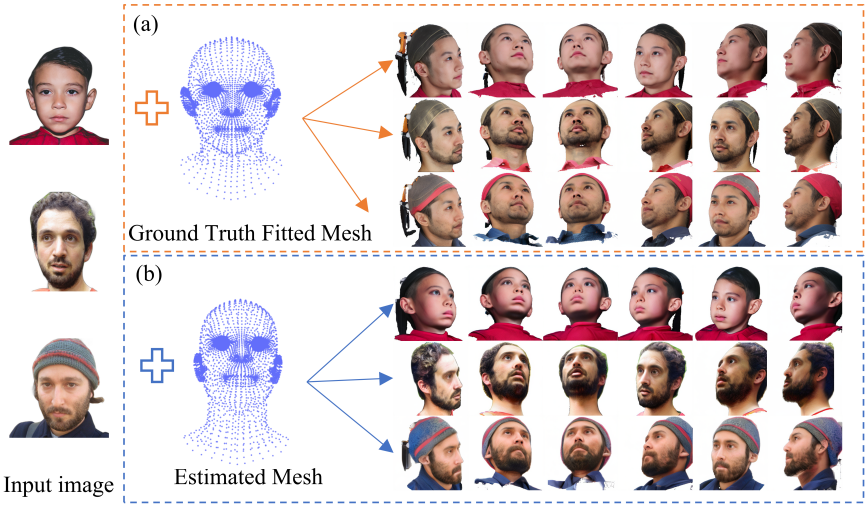

Fig. 5: Geometry priors: (a) Ground-truth fitted mesh [7]; (b) Our input-estimated mesh.

Table 4: Ablation on the effect of multi-view joint generation (MVJG) on the whole dataset.

Number of subset	FID↓	Output-to-Output ID Consistency↑	CLIP↑	Input-to-Output ID Consistency↑
No	58.4648	0.4441	0.6181	0.1571
Yes	**55.4901**	**0.4536**	**0.6765**	**0.1716**

Multi-view joint generation We evaluate the effect of our multi-view joint generation. From Table 4 we find that this design helps improve all the metrics, suggesting a good contribution.

5 Conclusion

In this work, we investigate for the first time the single image 3D face generation problem in unconstrained, out-of-domain scenarios. Established on recent multi-view diffusion, we present a novel generative method, Gen3D-Face, that can generate photorealistic 3D human face avatars from single, unconstrained images. We show that specific designs such as enhanced training data, input-conditioned mesh estimation, and multi-view joint generation matters to the final model generalisation. We benchmark this more challenging task with existing generative methods using more comprehensive metrics. Extensive experiments show that our method excels in creating unconstrained avatars for generic human subjects, whilst achieving competitive performance under the constrained in-domain setting.

Fig. 6: Examples of novel view generation on FFHQ (*out-of-domain* setting).

Fig. 7: Examples of novel view generation on Facescape (*in-domain* setting).

Fig. 8: More examples of novel view generation on FFHQ (*out-of-domain* setting).

References

1. AlBahar, B., Saito, S., Tseng, H.Y., Kim, C., Kopf, J., Huang, J.B.: Single-image 3d human digitization with shape-guided diffusion. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023) 4
2. An, S., Xu, H., Shi, Y., Song, G., Ogras, U.Y., Luo, L.: Panohead: Geometry-aware 3d full-head synthesis in 360deg. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20950–20959 (June 2023) 3, 4, 6, 8, 9
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023) 7
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. pp. 16123–16133 (2022) 3, 8, 9
5. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., De Mello, S., Karras, T., Wetzstein, G.: Generative novel view synthesis with 3d-aware diffusion models. arXiv preprint arXiv:2304.02602 (2023) 4
6. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. arXiv preprint arXiv:2304.06714 (2023) 8, 9
7. Chen, X., Mihajlovic, M., Wang, S., Prokudin, S., Tang, S.: Morphable diffusion: 3d-consistent diffusion for single-image avatar creation (2024) 2, 4, 6, 7, 8, 9, 10, 11
8. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36** (2024) 4
9. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023) 4
10. 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. pp. 4690–4699 (2019) 6, 7, 8
11. Dong, Z., Chen, X., Yang, J., Black, M.J., Hilliges, O., Geiger, A.: Ag3d: Learning to generate 3d avatars from 2d image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14916–14927 (2023) 3
12. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. vol. 40 (2021), https://doi.org/10.1145/3450626.3459936 6
13. Gafni, G., Thies, J., Zollhofer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8649–8658 (2021) 1
14. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017) 8
15. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9224–9232 (2018) 7

16. Gu, J., Trevithick, A., Lin, K.E., Susskind, J.M., Theobalt, C., Liu, L., Ramamoorthi, R.: Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In: International Conference on Machine Learning. pp. 11808–11826. PMLR (2023) 4
17. Han, X., Cao, Y., Han, K., Zhu, X., Deng, J., Song, Y.Z., Xiang, T., Wong, K.Y.K.: Headsculpt: Crafting 3d head avatars with text. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) 1
18. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? arXiv preprint arXiv:2210.07574 (2022) 4
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 7, 8
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 2, 3, 5
21. Hong, Y., Peng, B., Xiao, H., Liu, L., Zhang, J.: Headnerf: A real-time nerf-based parametric head model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20374–20384 (2022) 1, 3
22. Hu, S., Hong, F., Pan, L., Mei, H., Yang, L., Liu, Z.: Sherf: Generalizable human nerf from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9352–9364 (2023) 3
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 7
24. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4), 1–14 (2023) 2, 3
25. Kirschstein, T., Qian, S., Giebenhain, S., Walter, T., Nießner, M.: Nersemble: Multi-view radiance field reconstruction of human heads. ACM Transactions on Graphics (TOG) **42**(4), 1–14 (2023) 2
26. Lan, Y., Tan, F., Qiu, D., Xu, Q., Genova, K., Huang, Z., Fanello, S., Pandey, R., Funkhouser, T., Loy, C.C., Zhang, Y.: Gaussian3diff: 3d gaussian diffusion for 3d full head synthesis and editing. arXiv (2023) 4, 6
27. Li, P., Liu, Y., Long, X., Zhang, F., Lin, C., Li, M., Qi, X., Zhang, S., Luo, W., Tan, P., et al.: Era3d: High-resolution multiview diffusion using efficient row-wise attention. arXiv preprint arXiv:2405.11616 (2024) 8, 9
28. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6), 194–1 (2017) 4, 8
29. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object (2023) 2, 4, 5, 8, 9
30. Liu, Y., Lin, C., Zeng, Z., Long, X., Liu, L., Komura, T., Wang, W.: Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453 (2023) 2, 4, 5, 7, 8, 9
31. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023) 2, 4
32. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023) 4
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 8

34. 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) 2, 3
35. Pan, D., Zhuo, L., Piao, J., Luo, H., Cheng, W., Wang, Y., Fan, S., Liu, S., Yang, L., Dai, B., et al.: Renderme-360: A large digital asset library and benchmarks towards high-fidelity head avatars. Advances in Neural Information Processing Systems **36** (2024) 2
36. Peng, S., Zhang, Y., Xu, Y., Wang, Q., Shuai, Q., Bao, H., Zhou, X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9054–9063 (2021) 1
37. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on pattern analysis and machine intelligence **22**(10), 1090–1104 (2000) 8
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 2
39. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. arXiv preprint arXiv:2312.02069 (2023) 1
40. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. vol. 106, p. 107404 (2020) 7
41. Qiu, H., Yu, B., Gong, D., Li, Z., Liu, W., Tao, D.: Synface: Face recognition with synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10880–10890 (2021) 4
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 6, 8
43. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019) 5
44. Rebain, D., Matthews, M., Yi, K.M., Lagun, D., Tagliasacchi, A.: Lolnerf: Learn from one look. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1558–1567 (2022) 3
45. Roich, D., Mokady, R., Bermano, A.H., Cohen-Or, D.: Pivotal tuning for latent-based editing of real images. ACM Transactions on graphics (TOG) **42**(1), 1–13 (2022) 9
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 2, 3, 4, 5
47. Sengupta, A., Alldieck, T., Kolotouros, N., Corona, E., Zanfir, A., Sminchisescu, C.: Diffhuman: Probabilistic photorealistic 3d reconstruction of humans. arXiv preprint arXiv:2404.00485 (2024) 4
48. Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., Yang, X.: Mvdream: Multi-view diffusion for 3d generation. arXiv:2308.16512 (2023) 2, 4, 5, 7
49. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 8
50. Szymanowicz, S., Rupprecht, C., Vedaldi, A.: Splatter image: Ultra-fast single-view 3d reconstruction. arXiv preprint arXiv:2312.13150 (2023) 3

51. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: Lgm: Large multi-view gaussian model for high-resolution 3d content creation. arXiv preprint arXiv:2402.05054 (2024) 2
52. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. arXiv preprint arXiv:2303.14184 (2023) 4
53. 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. In: ACM Transactions on Graphics (SIGGRAPH) (2023) 1, 4
54. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023) 2
55. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3d digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4563–4573 (2023) 2
56. Wang, W., Zhang, L., Pun, C.M., Xie, J.C.: Boosting face recognition performance with synthetic data and limited real data. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 4
57. Wang, Y., Han, Q., Habermann, M., Daniilidis, K., Theobalt, C., Liu, L.: Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3295–3306 (2023) 2, 5
58. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004) 8
59. Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel view synthesis with diffusion models. arXiv preprint arXiv:2210.04628 (2022) 3
60. Xu, B., Zhang, J., Lin, K.Y., Qian, C., He, Y.: Deformable model driven neural rendering for high-fidelity 3d reconstruction of human heads under low-view settings. arXiv preprint arXiv:2303.13855 (2023) 1
61. Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., Cao, X.: Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 2, 4, 7
62. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. Advances in Neural Information Processing Systems **36** (2024) 4
63. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021) 3, 8, 9
64. Zhang, L., Qiu, Q., Lin, H., Zhang, Q., Shi, C., Yang, W., Shi, Y., Yang, S., Xu, L., Yu, J.: Dreamface: Progressive generation of animatable 3d faces under text guidance. arXiv preprint arXiv:2304.03117 (2023) 1
65. 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 conference on computer vision and pattern recognition. pp. 586–595 (2018) 8

66. Zheng, S., Zhou, B., Shao, R., Liu, B., Zhang, S., Nie, L., Liu, Y.: Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. arXiv preprint arXiv:2312.02155 (2023) 3
67. Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., Wen, F.: General facial representation learning in a visual-linguistic manner. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18697–18709 (2022) 2
68. Zhou, Y., Zhou, D., Cheng, M.M., Feng, J., Hou, Q.: Storydiffusion: Consistent self-attention for long-range image and video generation. arXiv preprint arXiv:2405.01434 (2024) 7
69. Zielonka, W., Bolkart, T., Thies, J.: Instant volumetric head avatars. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4574–4584 (2023) 1