Morphable Diffusion: 3D-Consistent Diffusion for Single-image Avatar Creation

Supplementary Material

In this supplementary document, we first provide additional evaluations of the novel view and expression synthesis of faces (Sec. 1 & 2). We further provide details for our implementation (Sec. 3) and comparison to baseline methods (Sec. 4). We also ablate our model for different design choices and training scheme (Sec. 5). Finally, we discuss the effect of the topology, expressiveness, and accuracy of the input meshes on the generated face images (Sec. 6).

1. Novel view synthesis of faces

Comparison to EG3D: We conduct an additional evaluation to compare our method with EG3D [2] for novel view synthesis on the FaceScape dataset [20]. Fig. 1 shows the qualitative results. The novel view synthesis with the optimized latent code obtained by GAN inversion of EG3D from the input view does not preserve the identity well and fails to generate realistic side views. This can be largely attributed to the fact that the model is trained on mostly frontal views. Note that since EG3D requires fixed camera distance, and the scale of the face on the pre-trained model is not available, it is not possible to use the real camera parameters for novel view synthesis. Instead, we render images with their camera parameters and use NeuS2 [17] for mesh reconstruction and novel view synthesis. However, it is still challenging to align the EG3D meshes accurately with the ground truth. Therefore, we do not perform a quantitative evaluation for this method and only show the qualitative results in the figure.

Geometry evaluation: In addition, we reconstruct the meshes for all the baselines and our method using NeuS2 [17] and compare the geometry quality. Since the hairstyle sometimes varies among each generated batch of 16 views in our method, we only reconstruct the mesh using one of the batches of generated images. For a fair comparison, we only use the first 16 generated views to reconstruct the meshes for all baselines. Instead of using the provided ground truth mesh scans in the FaceScape dataset, we also reconstruct the ground truth meshes with NeuS2, using all ground truth target views whose absolute camera azimuth is less than 90 degrees. We consider this evaluation strategy since some parts of the capture devices are visible in the ground truth mesh scans, but not in the reconstructed meshes of any of the compared methods.

Fig. 2 and Tab. 1 display the qualitative and quantitative results of the mesh reconstructions of faces. We report the Chamfer Distance and Volume IoU [13] for geometry accuracy. Even though our meshes demonstrate good geometric details, the quantitative scores are slightly outperformed by pixelNeRF, which however produces overly coarse meshes due to its blurry rendering results. We suspect the main reason to be that our model sometimes generates a hairstyle that is different from the one in the input image, *e.g*., long hair for male subjects. Overall, the meshes reconstructed from our method are visually comparable to the ones from SyncDreamer. However, our method preserves better resemblance, as shown in Fig. 3 of the paper, and has the additional advantage of the capability of facial expression rigging over SyncDreamer.

Method	Chamfer Distance ↓	Volume IoU ↑
zero-1-to-3 [10]	0.0950	0.0613
SyncDreamer [11]	0.0138	0.7947
SSD-NeRF [3]	0.0154	0.7801
pixelNeRF [21]	0.0118	0.8218
Ours	0.0130	0.8048

Table 1. **Quantitative evaluation of geometry for novel view synthesis of faces.** Colors denote the 1st and 2nd best-performing models. See Sec. 1 for details.

2. Novel facial expression synthesis

Expression rigging with ControlNet: We test the ability of ControlNet [22] of facial expression rigging by using the OpenPose [1] model for human pose conditioning and the projected ground truth facial keypoints as input image. We first personalize the Stable Diffusion 1.5 model [14] with DreamBooth [15] using one or multiple images of the test subject. Fig. 3 shows that the personalized model finetuned on the single input view tends to overfit to the specific facial expression in the input image. The model finetuned with 16 views of the test subject in random views and random facial expressions, however, fails to generate images with the correct facial expressions conditioning merely on the facial keypoint maps. We believe the main reason to be that the current human pose conditioning model isn't trained with various facial expressions. A specific facial expression conditioning model with facial keypoint maps as inputs could potentially enhance the ability of expression rigging.

Geometry evaluation: Fig. 4 and Tab. 2 show the qualitative and quantitative results of the mesh reconstructions on face novel expression synthesis. In this case, our method outperforms the baselines in terms of both visual qualities

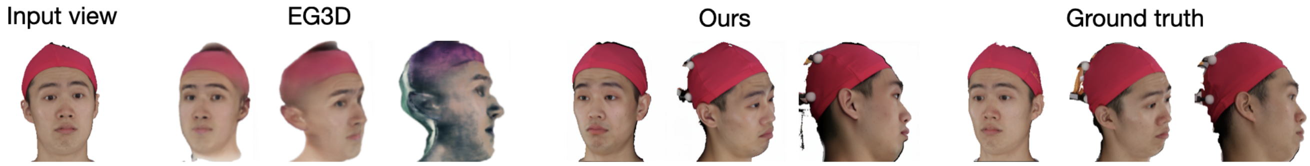

Figure 1. **Additional qualitative results on novel view synthesis of faces.** EG3D [2] fails to generate side views with good quality, while our method generates high-fidelity images in all views.

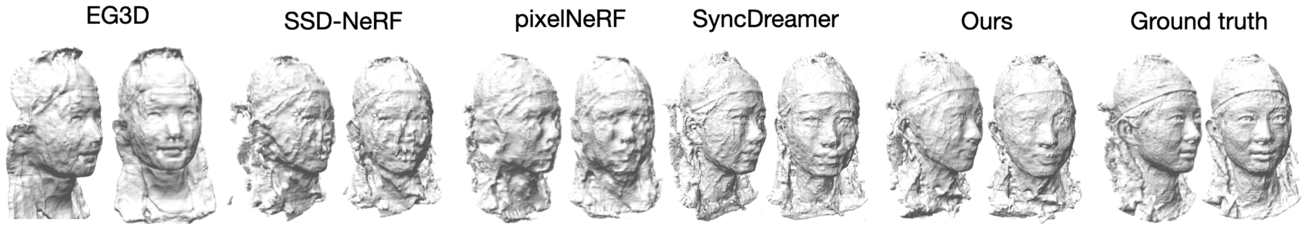

Figure 2. **Mesh reconstruction of faces.** EG3D [2] generates bad geometries for side views. SSD-NeRF [3] and pixelNeRF produce overly coarse geometries. SyncDreamer [11] and our method produces meshes with a comparable amount of details, but our method preserves better facial resemblance (Fig. 3 in the paper) and has the additional capability for expression rigging.

Single image input:	Chamfer Distance ↓	Volume IoU ↑
MoFaNeRF [23]	0.0234	0.7281
DECA [7]	0.0279	0.6818
DiffusionRig* [6]	0.0284	0.6383
Ours	**0.0113**	**0.7670**

Table 2. **Quantitative evaluation of geometry for novel facial expression synthesis.**

and geometry metrics.

Overall, we believe that although the precise geometry evaluation metrics are well suited for examining the reconstruction fidelity of multi-view NeRF-based methods, they are less reflective of the model performance in the case of generative models working in highly underconstrained single-image setups.

For the animation of the meshes and textures for all methods, please refer to our project website.

3. Implementation details

Layers	Layer Description	Output Dim
1-2	2 × (3 × 3 × 3 conv, stride=1)	D×H×W×16
3	3 × 3 × 3 conv, stride=2	½D×½H×½W×32
4-5	2 × (3 × 3 × 3 conv, stride=1)	½D×½H×½W×32
6	3 × 3 × 3 conv, stride=2	¼D×¼H×¼W×64
7-9	3 × (3 × 3 × 3 conv, stride=2)	¼D×¼H×¼W×64

Table 3. **Architecture of our SparseConvNet.**

We train our model with the AdamW [12] optimizer and a total batch size of 140 images for 6k steps (≈36 hours) using two 80GB NVIDIA A100 GPUs. The learning rate for training the UNet is increased from 1e-6 to 5e-5 with 100 warm-up steps [8]. The learning rate for the remaining trainable modules are set to 5e-4. During each training step, we randomly select 1 view as input and $N = 16$ target views. For inference, our method takes about 25 seconds to generate 16 target views from a single input image using 50 DDIM [16] steps with an NVIDIA RTX 3090 GPU.

Tab. 3 shows the architecture of our SparseConvNet f_θ (introduced in Section 3.2 in the paper). Given vertex features $\mathbf{V}_F \in \mathbb{R}^{n_v \times d}$, where n_v is the number of vertices and d is the dimensionality of the noise features, a sparse volume $\mathbf{V}_S \in \mathbb{R}^{D \times H \times W \times d}$ filled with the sparse vertex features is first constructed. Here, D, H, W are determined by the size of the bounding box of the face/full-body mesh, which differs for each mesh. Then, the SparseConvNet downsamples the size of the volume by 4 times, while increasing the number of channels to f_V by 4 times, and trilinearly interpolates the 3DMM-aware feature grid $\mathbf{F}_V$ from the downsampled volume. The SparseConvNet is implemented using the Spconv library [4].

In practice, we upsample the noisy image features to 16 channels with a 2D CNN block pretrained in SyncDreamer, and unproject and interpolate these noise features to construct $\mathbf{V}_F$. Therefore, we have $d = 16$ and $f_V = 64$. The size of the grid $\mathbf{F}_V$ is set as $x = y = z = 32$, and the size of the frustums $\mathbf{F}^{(i)}$ are set as $h_F = w_F = 32$, $d_F = 48$.

We set the voxel size for our SparseConvNet to 0.005 and the length of frustum volume to $\sqrt{3}/2$, same as their defaults in SyncDreamer.

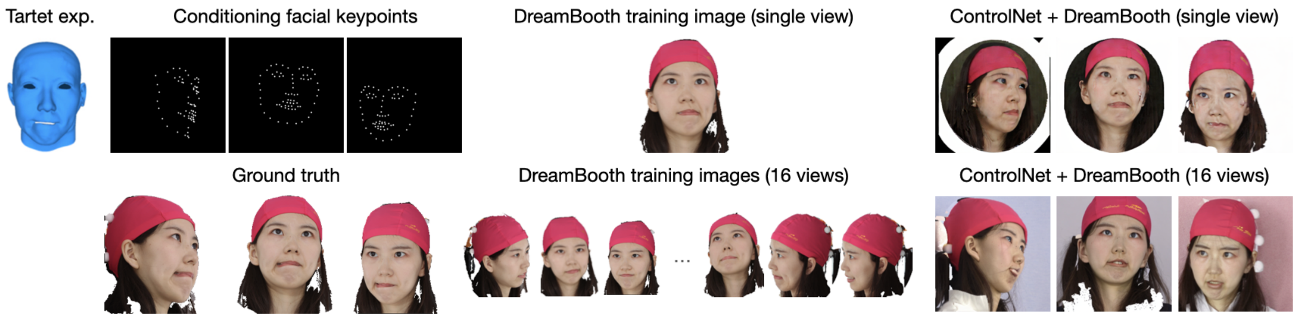
Tartet exp.
Conditioning facial keypoints
DreamBooth training image (single view)
ControlNet + DreamBooth (single view)
Ground truth
DreamBooth training images (16 views)
...
ControlNet + DreamBooth (16 views)

Figure 3. **Facial expression rigging with ControlNet [22] on personalized DreamBooth [15] models.** The model trained with a single view overfits to the expression in the training image, while the model trained with multiple views in different facial expressions fails to generate the correct facial expression based on the provided conditioning facial keypoints maps.

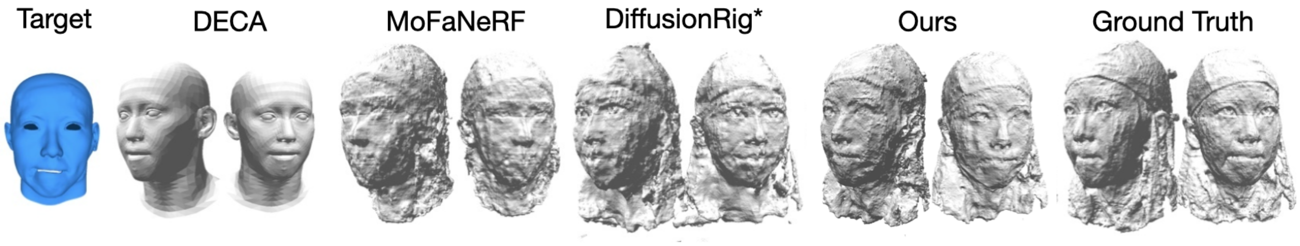

Figure 4. **Mesh Reconstruction for novel expression synthesis.** DiffusionRig requires per-subject finetuning with additional images and is thus denoted with the * symbol. For DECA [7], we simply show the pseudo ground truth FLAME [9] mesh since it only renders the predicted albedo map onto this coarse mesh. MoFaNeRF [23] and DiffusionRig [6] produce overly coarse geometries. Our method generates the highest amount of details and preserves the best geometry of the target subject.

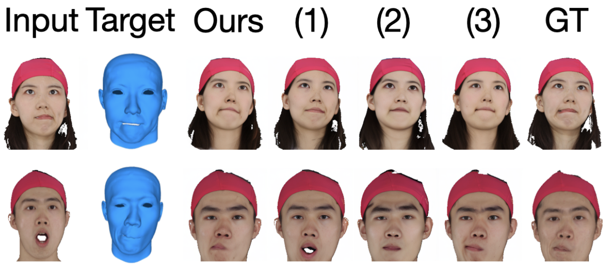

Figure 5. **Ablation studies of our proposed model v.s. different design choices and train- ing strategies**. Ablated models correspond to training with same facial expression for input and target views (1), not finetuning the UNet along with the conditioning module (2), and training with FLAME meshes fitted to 3D keypoints instead of the ground truth bilinear meshes (3). Model (1) overfits to the input views and tends to ignore the target facial expressions. Model (2) generates faces with worse resemblance. Model (3) generates faces in a comparable resemblance but slightly worse facial expressions due to the optimization loss of the FLAME model.

4. Additional details on baseline comparisons

For a fair comparison, we finetune zero-1-to-3 and SyncDreamer models pre-trained on Objaverse [5] on the FaceScape / THuman 2.0 datasets, for 6k steps. We finetune the UNet together with the conditioning module for SyncDreamer (see Sec. 5 for the detailed explanation). We train pixelNeRF and SSD-NeRF from scratch on the same datasets for 400k and 80k steps respectively, following the training schemes of the original pipelines. At inference, we use the same single input view for all methods. Finetuning and pre-training are performed with the default hyperparameters of the baseline methods. For EG3D, we use a custom GAN inversion repository [18] to optimize the latent codes for 1000 steps for all input images.

5. Ablation studies

We compare our method with several variants of pipeline designs and training strategies: 1) having the same facial expression for input and target views during training, 2) not finetuning the UNet, and 3) instead of using ground truth bilinear meshes provided by the FaceScape dataset [20], we use FLAME [9] meshes fitted to the ground truth 3D keypoints via optimization. Since the held-out facial ex-

	LPIPS ↓	SSIM ↑	FID ↓	PCK@0.2 ↑	PCK@0.2 (mouth) ↑	Re-ID ↑
Morphable Diffusion	**0.1693**	**0.8026**	14.34	**95.46**	**94.23**	99.89
w. same expression	0.1787	0.7881	**13.68**	92.39	84.12	**100.00**
w.o. finetuning UNet	0.1841	0.7910	20.70	93.44	90.30	98.35
w. FLAME [9] meshes	0.1764	0.7939	15.03	95.22	93.23	99.45

Table 4. **Ablation studies on different design choices and training strategies of our novel facial expression synthesis model.** The proposed pipeline demonstrates superior performance compared to the alternative designs on most metrics. Our proposed model produces the most accurate keypoints for the test facial expression, with the difference even larger for mouth keypoints only.

	LPIPS ↓	SSIM ↑	FID ↓	PCK@0.2 ↑	Re-ID ↑
SyncDreamer w. UNet	**0.1854**	**0.7732**	**6.05**	**94.07**	**99.60**
SyncDreamer w.o UNet	0.2026	0.7585	7.53	88.64	96.60

Table 5. **Ablation studies on the effect of finetuning UNet with SyncDreamer [11] for novel view synthesis on FaceScape [20].** The model that finetunes the UNet outperforms the model that does not on every metrics, suggesting the need to finetune the UNet when we apply the method to human faces.

pression “jaw_right” mainly involves rigging on mouth keypoints, we additionally report the PCK metric on the 20 mouth keypoints, which we denote as “PCK@0.2 (mouth)”. Tab. 4 and Fig. 5 show that our proposed pipeline quantitatively achieves the best performance on most metrics, while preserving the most accurate facial expressions and resemblance qualitatively.

The ablated model which employs the same facial expressions for both input and target views, exhibits overfitting to the task of novel view synthesis and tends to overlook the driving signal from the target expression mesh. This leads to diminished performance in terms of both image quality and expression preservation metrics. Such observations underscore the effectiveness of our proposed shuffled training scheme. This scheme, by using images with varying facial expressions for the input and target views, successfully disentangles the processes of reconstruction and animation, as elaborated in Section 4.2 of the paper.

Although SyncDreamer [11] proposes not to finetune the UNet when training their proposed conditioning module with a UNet pretrained on the same object dataset, we find that it's still beneficial to do so when we apply this baseline onto the human domain, as shown in Tab. 5. Therefore, for the baseline SyncDreamer models that we report in the main paper, we finetune the UNet together with the conditioning module. Similarly, we find improvement to finetune the UNet with our proposed method. As shown both qualitatively and quantitatively, our proposed model generates faces with better resemblance compared to the ablated model without finetuning the UNet.

We also fit FLAME model to the ground truth 3D facial keypoints for all meshes in the FaceScape dataset via optimization, and use the FLAME meshes instead of the ground truth bilinear meshes for both training and inference. This model produces faces that are comparable but slightly worse (due to the loss in the mesh-fitting optimization) to the results generated with our proposed model trained and tested with bilinear meshes. We experiment with this model mainly because we find that not changing the mesh topology during inference produces better results, and there are more off-the-shelf FLAME-based methods to reconstruct meshes from single images [7, 24]. Therefore, we provide this model for better generalization to in-the-wild face images, as shown in Fig. 6 of the paper. More details about mesh topologies will be discussed in Sec. 6.

6. Effects of mesh topology, expressiveness, and accuracy

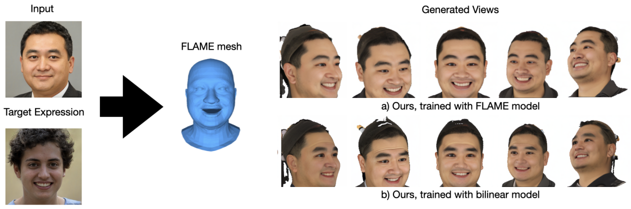
Input
Target Expression
FLAME mesh
Generated Views
a) Ours, trained with FLAME model
b) Ours, trained with bilinear model

Figure 6. **The effect of changing mesh topology during inference on the generated faces.** a) we test our model trained with FLAME meshes (with 5,023 vertices) on a target expression mesh also in FLAME topology. The shape and facial expression of the mesh are estimated from the input and the target expression images respectively using MICA [24]. b) we use the same input image and mesh but the model is trained with the bilinear mesh topology (with 26,317 vertices). We observe that using the same mesh topology for training / inference leads to results with more accurate facial structures and better resemblance.

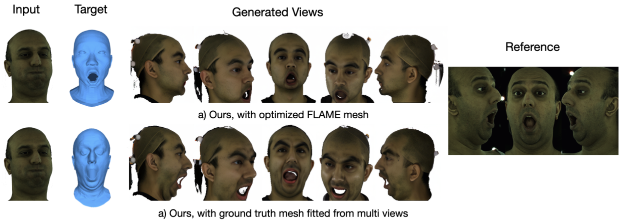
Input
Target
Generated Views
Reference
a) Ours, with optimized FLAME mesh
a) Ours, with ground truth mesh fitted from multi views

Figure 7. **The effect of the expressiveness of the mesh on the generated faces.** We compare the generated images using a) the FLAME mesh optimized from a single image of the reference expression using MICA [24] v.s. b) using the more expressive ground truth mesh provided in the dataset fitted from all views of the reference expression on a Multiface [19] subject. We find that using more expressive meshes of the same facial expression improves the resemblance of the generate faces, although neither results preserves good resemblance due to the limited generalization ability of our method on different ethnicities (as discussed in Sec. 5 of the paper).

Although our method is agnostic of the input mesh topology, thanks to the SparseConvNet's ability to process an arbitrary point cloud, we find that having the same mesh topology for training and inference leads to better results, as shown in Fig. 6.

However, as shown in Fig. 7, having more expressive meshes could still lead to better preservation of the subject identity. Future works could consider leveraging geometry information of the input image into the conditioning module, such as signed distance fields (SDFs), to improve the resemblance.

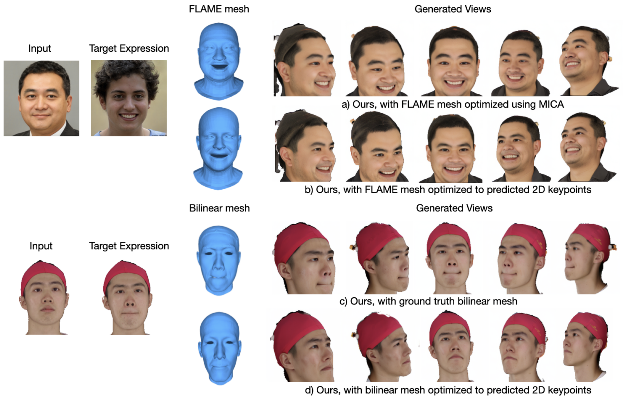
FLAME mesh
Generated Views
Input
Target Expression
a) Ours, with FLAME mesh optimized using MICA
b) Ours, with FLAME mesh optimized to predicted 2D keypoints
Bilinear mesh
Generated Views
Input
Target Expression
c) Ours, with ground truth bilinear mesh
d) Ours, with bilinear mesh optimized to predicted 2D keypoints

Figure 8. **The effect of mesh accuracy on the generated faces**. We compare the generated faces using a FLAME mesh obtained with the state-of-the-art FLAME reconstruction pipeline [24] (a) with the generated results using the FLAME mesh fitted to detected 2D keypoints only via optimization (b). We also compare results using the ground truth bilinear mesh (c) with the results using the bilinear mesh fitted to 2D keypoints (d). (a) and (b) are tested using our model trained with FLAME models while (c) and (d) are tested using our model triained on bilinear models. Meshes with more accurately reconstructed facial expression and shape parameters lead to more accurate preservation of the target expression and better resemblance in the generated images.

Additionally, we find that having an accurate mesh estimator can lead to better preservation of the facial expression and better resemblance, as shown in Fig. 8.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *PAMI*, 2019. 1

[2] 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. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 1, 2

[3] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *ICCV*, 2023. 1, 2

[4] Spconv Contributors. Spconv: Spatially sparse convolution library. `https://github.com/traveller59/spconv`, 2022. 2

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 3

[6] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *CVPR*, 2023. 2, 3

[7] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. Graph.*, 2021. 2, 3, 4

[8] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 2

[9] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *SIGGRAPH Asia*, 2017. 3, 4

[10] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 1

[11] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. In *ICLR*, 2024. 1, 2, 4

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2

[13] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 1

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[15] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1, 3

[16] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 201. 2

[17] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *ICCV*, 2023. 1

[18] Yiqian Wu. Eg3d inversion projector. `https://github.com/oneThousand1000/EG3D-projector`, 2022. 3

[19] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. *arXiv preprint arXiv:2207.11243*, 2022. 4

[20] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *CVPR*, 2020. 1, 3, 4

[21] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 1

[22] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1, 3

[23] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *ECCV*, 2022. 2, 3

[24] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *ECCV*, 2022. 4, 5