Joint Learning of Depth and Appearance for Portrait Images

Xinya Ji[1,2] Gaspard Zoss[3] Prashanth Chandran[3]
Lingchen Yang[1] Xun Cao[2] Barbara Solenthaler[1] Derek Bradley[3]
[1]ETH Zürich [2]Nanjing University [3]DisneyResearch|Studios
xinya@smail.nju.edu.cn caoxun@nju.edu.cn {lingchen.yang,solenthaler}@inf.ethz.ch
{prashanth.chandran,gaspard.zoss,derek.bradley}@disneyresearch.com

Abstract

*The field of 2D portrait manipulation has experienced significant advancements in recent years. A lot of research has leveraged the prior knowledge embedded in large generative diffusion models to enable high-quality image editing and animation tasks. However, most generative methods only focus on creating RGB images as output, and the co-generation of consistent visual plus 3D output remains largely under-explored. In our work, we propose to **jointly** learn the visual appearance and depth of faces simultaneously in a diffusion-based portrait image generator. Our method embraces the end-to-end diffusion paradigm and introduces a new architecture suitable for learning this joint distribution, consisting of a reference network for target identity and a channel-expanded diffusion backbone. We extend the training objective to predict both RGB and depth from a single representation, enabling various applications such as joint generation, facial depth estimation, and depth-driven portrait manipulation. Our experiments demonstrate that joint learning not only surpasses separate conditional generation but also achieves state-of-the-art results on both facial depth estimation and portrait image animation, validating the benefit of a joint-learning approach for depth and appearance of portrait images.*

1. Introduction

The development of deep generative models, like GANs [24, 25] and diffusion models [4, 19, 37] have witnessed significant advancements in recent years. What we learned from large pre-trained models like Stable Diffusion [37, 41] is that image synthesis methods can produce stunning photo-realistic images of human faces, indistinguishable from reality. Based on these models, researchers have devised a host of algorithms for tasks like portrait relighting [29, 36, 40], appearance editing [9, 28, 44, 65] (*e.g.* changing the hair color, adding accessories or re-aging the person) and animation retargeting [12, 16, 22, 35, 39, 43, 46, 53, 58, 64].

To accomplish these tasks, modern approaches start with a pre-trained generative model [4, 11, 37, 38] as the backbone and learn to condition the generation on other signals, like landmarks or images. While such methods [17, 18, 20, 61] have proven to be very powerful in portrait image manipulation, one issue is that the backbone generators only learn to generate the appearance (*i.e.* the RGB color) of the face, which can limit downstream applications that would benefit from additional information such as depth. Moreover, recent work [13, 26] has demonstrated that pre-trained diffusion models encode rich 3D structural priors, which facilitates 3D tasks like depth estimation and shape reconstruction. Even though some studies [32, 71] have attempted to leverage 3D representations as condition to guide portrait generation, these methods take the representations purely as an auxiliary signal without fully exploiting the intrinsic 3D information embedded within the diffusion model. In contrast, we propose a simple yet effective joint-learning framework based on a diffusion generative model where visual appearance and 3D depth are learned simultaneously. We directly integrate the appearance and depth information into the diffusion process by learning a joint distribution, so that 3D priors are better utilized, yielding various applications including joint appearance plus depth generation, facial depth estimation, and depth-driven portrait animation.

The correlation between the appearance and depth channels is of critical importance for our jointly learning architecture, *i.e.* the generated depth map must match the generated face image. Our new diffusion-based portrait image generator is built on top of Stable Diffusion [37, 41], but adapted to learn this joint distribution. Similar to related work [20, 48, 60], we employ a reference network designed to extract the identity of an RGB reference photo, which guides the image diffusion process. We expand the traditional Stable Diffusion backbone to de-noise a 6-

channel input image, which consists of separately-noised RGB and depth latent images, and we extend the training objective to predict both RGB and depth from a single representation. The shared UNet in the diffusion step also ensures good correlation between the appearance and depth outputs. Finally we train the model on a combination of studio-captured facial images with ground truth 3D geometry obtained from a facial scanner, and also in-the-wild facial videos with approximate 3D reconstructed geometry. As we will show, this combination allows our model to both learn accurate depth generation and also generalize to outdoor settings.

Once our model is trained it can be adapted for several applications. In addition to unconditional sampling to achieve coupled RGB and depth images, we show applications of channel-wise inpainting. Specifically, for a given image we can inpaint the depth channel, achieving facial depth estimation with our model. Alternatively, for a given depth image, we can inpaint the RGB channels to obtain an artistic way to control face image generation using either a 3D morphable face model or the estimated depth from a separate image. Furthermore, following previous works [18, 20, 61] that introduce temporal layers in diffusion models, our approach can be extended to generate temporally consistent visual results. Importantly, our experiments show that joint learning outperforms separate conditional learning (*e.g.* from depth to RGB or RGB to depth) for facial depth estimation and portrait animation.

Specifically, our contributions are:

1. A novel architecture for joint learning of depth and appearance of portrait images, implicitly learning the relationship of 2D and 3D information, with better performance than separate conditional generation,
2. A new training scheme for learning paired image and depth maps from a combination of in-studio and in-the-wild facial data,
3. The demonstration of several applications in portrait manipulation including both image-to-depth and depth-to-image channel-wise inpainting, achieving state-of-the-art results for facial depth estimation, depth-driven image editing and animation.

2. Related Works

Diffusion Model for Geometric Estimation. Diffusion models trained on large image datasets for high-quality generation tasks have been proven to contain a rich understanding of the underlying scene structure. This capability has extended the diffusion model to 3D geometric estimation tasks, including depth estimation [1, 2, 15, 21, 62], normal estimation [54, 59], and view synthesis [33, 34, 45, 51]. Recently, Marigold [26] leverages the diffusion priors by fine-tuning large pre-trained diffusion models specifically for depth estimation. Wonder3D [33] designs a cross-domain diffusion model with attention across different modalities for information exchange. Geowizard [13] proposes to jointly estimate depth and normals and involve a scene distribution decoupler strategy to discern different scene layouts. Recently, Khirodkar *et al*. [27] proposes Sapiens, a human-centric foundation model capable of pose estimation, body segmentation, depth and normals estimation. Several approaches have been developed to jointly denoise cross-domain representations utilizing the prior of large pretrained model. JointNet [66] achieves joint generation by replicating the original network, enabling it to handle multiple geometric tasks within a unified framework. Additionally, HyperHuman [31] proposes to learn the correlation between appearance and geometric structure by denosing the depth and surface-normal along with the RGB image. Different from these methods, we introduce a reference network designed for portrait images to extract identity information, with no need for an additional joint network to keep the generalization of the diffusio model. Instead, our unified framework efficiently achieves state-of-the-art results on depth estimation and portrait animation.

Diffusion Model for Portrait Animation. Diffusion-based generative models have shown remarkable capabilities in generative tasks, demonstrating diversity and adaptability across various multimedia domains. The development of large pre-trained models, such as Stable Diffusion [41], has spurred numerous applications leveraging its robust model priors. By extending the pre-trained model from 2D image generation to 3D video generation, researchers have explored tasks for animating human images. For example, AnimateDiff [18] introduces a plug-and-play temporal module designed to adapt flexibly to different motion patterns without model-specific tuning. In animating specific characters, DreamPose [23] introduces a dual clip-image encoder for image encoding. Similarly, methods like Animate Anyone [20], MagicAnimate [61], and Talk-Act [14] resort to a ReferenceNet with symmetrical U-Net architecture to maintain appearance consistency. Intermediate representations like landmarks, skeletons, or segmentation maps are used as control signals in this process for fine-grained control. Our work builds upon the diffusion priors of Stable Diffusion, achieving video generation by integrating a motion module for improved temporal consistency.

3. Joint Learning Method

We now describe our joint learning framework, starting with preliminary background details on latent diffusion models, followed by our new model for joint learning.

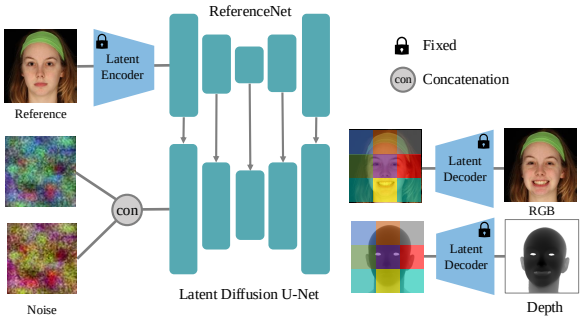

Figure 1. The overview of the proposed pipeline. Given a reference image, our model jointly generates the appearance (RGB) and depth of the identity under various expressions and poses, by simply sampling random noise in the latent space.

3.1. Preliminaries: Latent Diffusion Models

Diffusion models have set the new standard for generative models due to its ability to generate high-quality samples and perform a wide range of tasks with finetuning techniques. The diffusion model is trained to generate images by iteratively adding noise to the image and then removing the noise level-by-level, so that the model learns to generate the image from the gaussian noise. Different from the diffusion models that directly work on the image space, the latent diffusion models perform diffusion in a latent space, providing computational compactness and scalability to higher resolution images. The latent space is obtained from a pretrained variational auto-encoder (VAE) [30].

For a given sampled image $\mathbf{x}$, the encoder $\mathcal{E}$ of the VAE encodes the image into this latent space, as $\mathbf{z} = \mathcal{E}(\mathbf{x})$. The forward pass of the diffusion process adds noise to the latent code $\mathbf{z}_0$ according to the uniformly sampled noise level l:

$$\mathbf{z}_l = \sqrt{\bar{\alpha}_l}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_l}\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, $\bar{\alpha}_l$ is associated with the variance schedule of a diffusion process with L noise levels so that $\mathbf{z}_L$ becomes a gaussian distribution. In the reverse process, the denoising network $\boldsymbol{\epsilon}_\theta(\cdot)$, parameterized with learnable parameters θ, gradually removes noise from $\mathbf{z}_l$ to get $\mathbf{z}_{l-1}$, so as to obtain the fully denoised $\mathbf{z}_0$. The decoder $\mathcal{D}$ of the VAE then decodes $\mathbf{z}_0$ to generate the image $\mathbf{x}$. During training, the parameters θ are updated by minimizing the following loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),l\sim\mathcal{U}(L)} \left\|\epsilon - \boldsymbol{\epsilon}_\theta(\mathbf{z}_l, l)\right\|^2. \tag{2}$$

Equipped with conditional information injected using cross-attention modules [50], the latent diffusion models can be extended to perform various tasks, such as text-to-image generation [41] and image-to-image translation [67].

In this work, we propose to leverage a pretrained latent diffusion model and adapt it to perform the task of co-generation of depth and appearance for portrait image animation, conditioned on a reference image.

3.2. Joint Learning of Depth and Appearance

As demonstrated in Fig. 1, given a reference image $\mathbf{r}$ of the identity of interest, our task is to jointly generate the appearance (RGB) $\mathbf{x}$ and depth $\mathbf{d}$ of the subject under various expressions and poses. We model this as a conditional joint distribution in the latent diffusion U-Net [42] model, as $p(\mathbf{z}_0^\mathbf{x}, \mathbf{z}_0^\mathbf{d}|\mathbf{r})$, where $\mathbf{z}_0^\mathbf{x}$ and $\mathbf{z}_0^\mathbf{d}$ are the latent features for the appearance and depth, respectively. The final maps are generated by decoding the latent codes with the decoder of the VAE, as $\mathbf{x} = \mathcal{D}(\mathbf{z}_0^\mathbf{x})$ and $\mathbf{d} = \mathcal{D}(\mathbf{z}_0^\mathbf{d})$.

The reference image $\mathbf{r}$ is essential to generate consistent appearance and depth of the identity. In order to capture intricate details of the target, we use a reference network $\mathcal{R}$ to extract the identity features $\mathbf{z}^\mathbf{r}$ from the reference image $\mathbf{r}$, as $\mathbf{z}^\mathbf{r} = \mathcal{R}(\mathbf{r})$, which are then injected into the latent diffusion model using the spatial attention modules. Now the denoising process is conditional on the reference features, as $\mathbf{z}_{l-1} = \mathbf{z}_l - \boldsymbol{\epsilon}_\theta(\mathbf{z}_l, \mathbf{z}^\mathbf{r}, l)$.

To model the joint distribution, we generalize the latent diffusion process to handle multiple latent codes. Specifically, in the forward pass, we use the encoder $\mathcal{E}$ to separately encode the appearance and depth, as $\mathbf{z}_0^\mathbf{x} = \mathcal{E}(\mathbf{x})$ and $\mathbf{z}_0^\mathbf{d} = \mathcal{E}(\mathbf{d})$. We then independently add noise to each latent code, as follows:

$$\mathbf{z}_l^\mathbf{x} = \sqrt{\bar{\alpha}_l}\mathbf{z}_0^\mathbf{x} + \sqrt{1-\bar{\alpha}_l}\boldsymbol{\epsilon}^\mathbf{x}, \tag{3}$$

$$\mathbf{z}_l^\mathbf{d} = \sqrt{\bar{\alpha}_l}\mathbf{z}_0^\mathbf{d} + \sqrt{1-\bar{\alpha}_l}\boldsymbol{\epsilon}^\mathbf{d}, \tag{4}$$

where $\boldsymbol{\epsilon}^\mathbf{x}$ and $\boldsymbol{\epsilon}^\mathbf{d}$ are independently sampled from $\mathcal{N}(0, \mathbf{I})$. This follows the same reverse process as the original latent diffusion model, except now the denoising network $\boldsymbol{\epsilon}_\theta$ is modified to denoise both latent codes. For simplicity, we concatenate the noised appearance and depth latent codes, as $\mathbf{z}_l = [\mathbf{z}_l^\mathbf{x}, \mathbf{z}_l^\mathbf{d}]$. Then the denoising network $\boldsymbol{\epsilon}_\theta(\cdot)$ is modified to denoise the concatenated latent code, as $[\mathbf{z}_{l-1}^\mathbf{x}, \mathbf{z}_{l-1}^\mathbf{d}] = \mathbf{z}_l - \boldsymbol{\epsilon}_\theta(\mathbf{z}_l, \mathbf{z}^\mathbf{r}, l)$. During training, $\boldsymbol{\epsilon}_\theta$ learns to predict the concatenated noise:

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{\epsilon}^*\sim\mathcal{N}(0,\mathbf{I}),l\sim\mathcal{U}(L)} \left\|[\boldsymbol{\epsilon}^\mathbf{x}, \boldsymbol{\epsilon}^\mathbf{d}] - \boldsymbol{\epsilon}_\theta([\mathbf{z}_l^\mathbf{x}, \mathbf{z}_l^\mathbf{d}], \mathbf{z}^\mathbf{r}, l)\right\|^2. \tag{5}$$

3.3. Network Architecture

Diffusion Backbone. We aim to leverage the expressive knowledge stored in a pretrained latent diffusion model to learn our proposed conditional joint distribution with limited available data. However, since the latent diffusion model is originally trained to generate only RGB images,

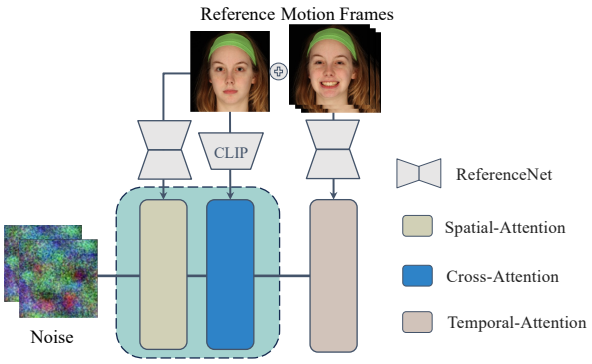

Figure 2. Detailed architecture of the building block of our extended model for portrait video generation, equipped with additional motion modules to achieve temporal consistency.

it must be adapted to co-generate depth and appearance. Here, we adopt a straightforward solution: expanding the input and output channels of the latent denoising network ϵ_θ. Specifically, the additional parameters in the input layer are initialized to zero, while the parameters in the output layer are duplicated from the original ones. We find it sufficient for our task, likely due to the rich priors learned in pretrained model, which enhances the model's capability to produce satisfactory results.

ReferenceNet. ReferenceNet is designed to enhance and stabilize the generation process by leveraging existing images as reference. It mirrors the layer structure of the denoising model, ensuring compatibility. Both networks produce feature maps with matching spatial resolutions and semantically aligned characteristics. This alignment allows ReferenceNet to effectively integrate extracted features into the diffusion model, resulting in improved visual quality. The weights of our ReferenceNet are initialized from the denoising network and trained together with it.

4. Adapting for Applications

Once trained, our model can be adapted to achieve several applications, which we will highlight in Sec. 5 and the supplemental video. For example, sampling from the model to achieve novel expressions with corresponding depth is straightforward. However, the model can also be adapted for bi-directional prediction of image or depth conditioned on the other signal, allowing for tasks such as monocular depth estimation and depth-based image editing/animation. In this section, we discuss the details of adapting our model for these tasks.

Our joint distribution of depth and appearance can be transformed into a conditional distribution in both directions by domain-wise inpainting. With a light fine-tuning process, our model is capable of both depth-to-image generation and image-to-depth generation. Specifically, we employ masked latent as an additional input condition [41] and design asymmetric masks for appearance and depth while fine-tuning. The task of image-to-depth, *i.e.* monocular depth estimation, can then be achieved by setting pure white for the depth mask and pure black for the image mask, and vice-versa for depth-to-image generation. Note that the involvement of ReferenceNet here enables the generation of the RGB image matching the appearance of the reference image, which allows possibilities for various applications like facial attribute editing and animation. Our model can be easily extended to generate portrait videos by incorporating temporal modules into diffusion backbone with attention modules. Similar to existing methods [18, 20, 61], we add an additional temporal-attention module in each building block of the denoising U-Net to maintain consistency between the generated frames. Fig. 2 illustrates the detailed architecture of the building block of our extended model.

Method	AbsRel ↓	δ_1 ↑	RMSE ↓
Marigold [26]	0.529	0.538	0.055
GeoWizard [13]	0.392	0.644	0.050
DepthAnything V2 [63]	0.457	0.612	0.050
Sapiens-0.3B [27]	0.313	0.526	0.056
Sapiens-0.6B [27]	0.297	0.549	0.048
Sapiens-1B [27]	0.197	0.696	**0.047**
w/o Joint Learning	0.260	0.738	0.050
Ours-Wild-Only	0.313	0.658	0.059
Ours	**0.162**	**0.765**	**0.047**

Table 1. Quantitative comparison for monocular depth estimation on portrait images.

5. Experiments

We begin describing our experimental setups (Sec. 5.1), and then we show that joint learning surpasses separate conditional learning for the same tasks (Sec. 5.2). We then demonstrate several applications including monocular depth estimation (Sec. 5.3), depth-conditioned portrait editing (Sec. 5.4) and depth-driven portrait animation (Sec. 5.5), showing that our method outperforms existing techniques. Finally, we end with limitations of our method (Sec. 5.6). Please refer to the supplemental document for additional results and ablation studies.

5.1. Experimental Setups

Implementation Details. The joint learning of portrait RGB images and depth takes about three days on four 4090 GPUs. We use a batch size of 32 and a constant learning rate of 1e-5 and train our model for 30000 steps. We then fine tune our model by incorporating different masks for the inpainting task. Note that our model is still

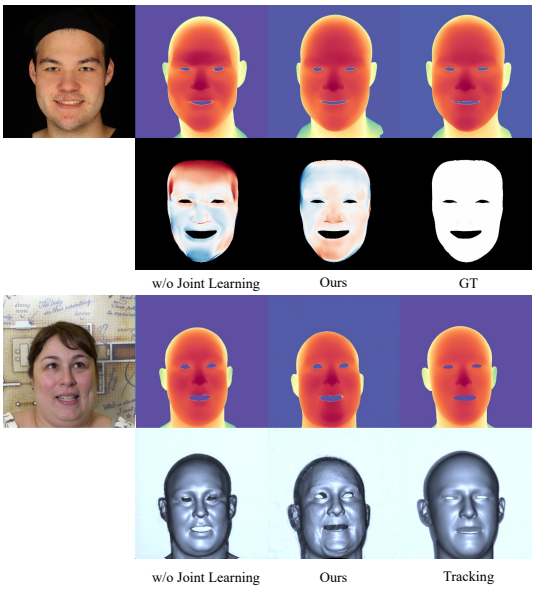

Figure 3. Benefit of joint learning for depth estimation.

capable of jointly generating RGB and depth by setting the mask to be pure white for both domains. To train our depth-driven animation network, we incorporate temporal attention modules. The training is performed on 4x RTX 4090 GPUs, and it takes about 3 days for the joint-learning model training, 15 hours for the inpainting fine-tuning and 2 days for the motion module training respectively.

The weights of the motion module are initialized from Animatediff [18], and we retain all other parameters from the first stage. We generate 14 frames at once and use the first 4 ground truth frames of each training sample as the motion frames during training.

Datasets. We train our model on a combination of datasets collected from both studio and in-the-wild scenarios. Particularly, we use a high-quality multi-view studio face dataset [5], comprising of images from 336 subjects performing various facial expressions, along with corresponding high-fidelity registered meshes. We render ground truth depth maps from these registered meshes to obtain the paired RGB-Depth data for training our method. To improve the generalization of our model to real world data, we incorporate in-the-wild visual sequences, from selected clips of the HDTF [70] and VFHQ [57] datasets. As these in-the-wild datasets do not contain corresponding geometry, we use a state of the art monocular face tracking approach [6, 7] to estimate 3D geometry from these videos, using which we can extract pseudo ground truth depth maps for training. In total we collect around 3 hours of video data from HDTF and VFHQ, containing 2,423 clips with diverse identities, facial expressions and head poses. All videos are sampled at 25 FPS and the images are cropped to a resolution of 512×512. The combination of studio data and in-the-wild data provides a solid foundation, enabling our network to jointly generate high-quality image and depth across various practical scenarios.

Method	L1 ↓	SSIM↑	LPIPS↓	FID↓
w/o Joint Learning	0.058	0.640	0.178	29.720
Ours	**0.055**	**0.691**	**0.174**	**27.536**

Table 2. Quantitative comparison for appearance generation.

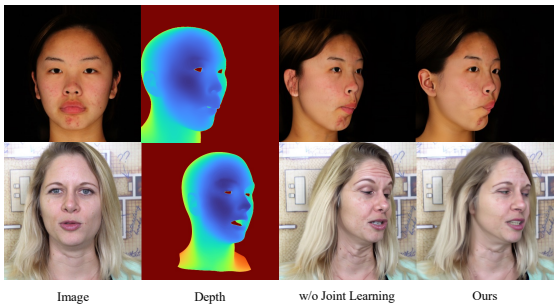

Figure 4. Benefit of joint learning for appearance generation.

5.2. Benefit of Joint Learning

In order to evaluate the effectiveness of joint learning, we compare our approach with separate individual networks in both directions. To this end, we train two separate networks that take image or depth as a condition for the U-Net to predict the corresponding depth or image, respectively. We use the same dataset and training settings to train the separate conditional networks for fair comparison.

Image to Depth. We first train the depth generation network and compare it with our model on a monocular depth estimation task on an unseen studio dataset, as this provides us with precise 3D depth maps that we can consider as ground truth. This evaluation dataset consists of 1264 images from 55 identities consisting of various facial expressions and poses. We follow the relative-depth evaluation protocols proposed in MiDaS [3] and LDM3D [47], and evaluate standard metrics including absolute relative error (AbsRel), δ_1 accuracy and root mean squared error (RMSE). As the ground truth depth maps derived from the 3D mesh in the studio dataset contain only the facial skin region, we apply a mask and remove regions outside this area to ensure a fair comparison of methods. We show quantitative results at the end of Tab. 1 and qualitative results in Fig. 3. Our approach outperforms separate generation on all metrics. We also present qualitative results for monocular depth estimation on in-the-wild face portraits. Since there is no groundtruth depth for them, we show estimated depth from a fitting method [6, 7] that was used to generate the depth component of our in-the-wild training data. Our approach generates depth that are

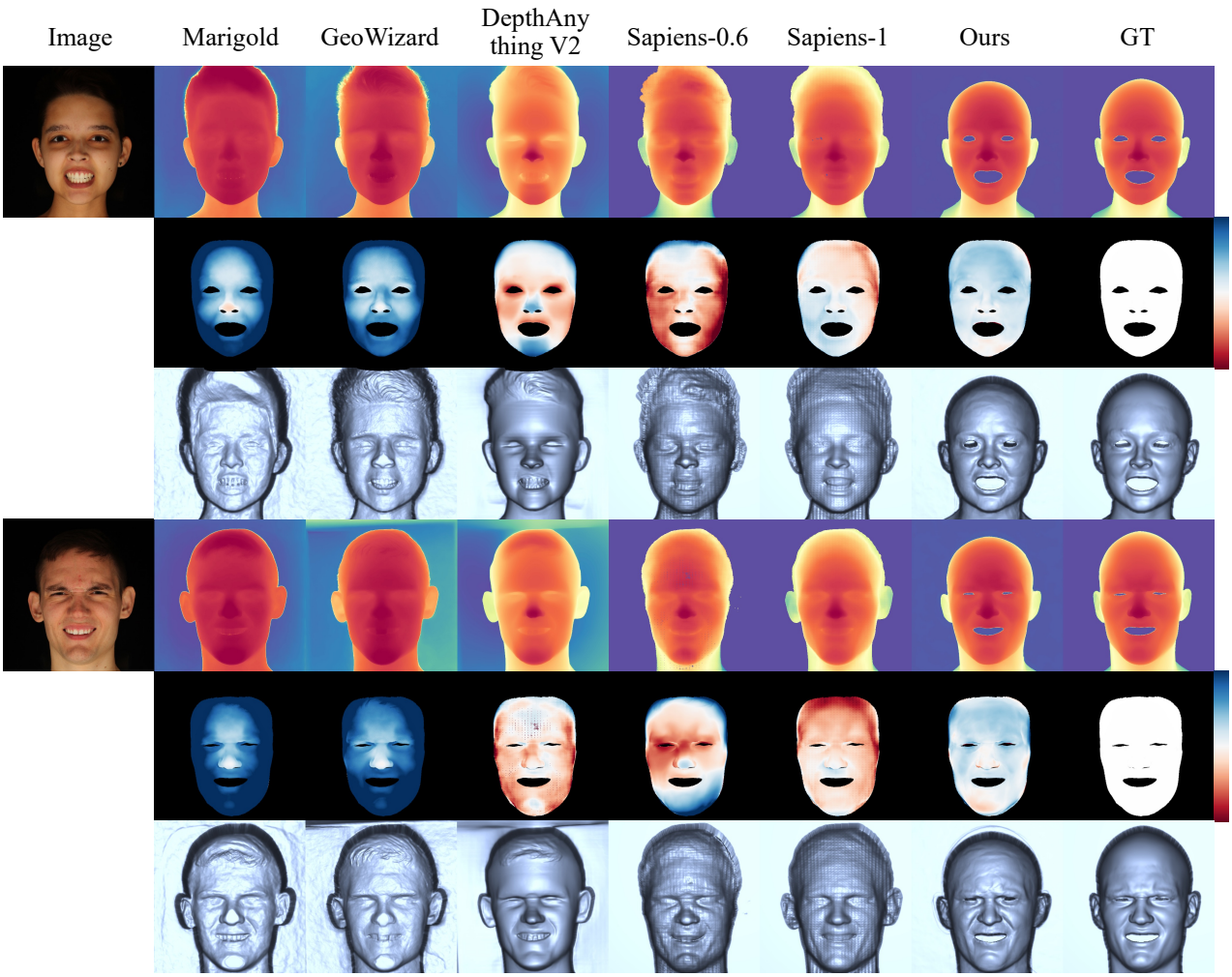

Figure 5. Qualitative comparisons with the state-of-the-art methods for monocular depth estimation on studio images. Several existing methods produce flat or exaggerate faces (please see supplemental video for geometry side-views).

most aligned with the image, even surpassing fitting-based methods, demonstrating the effectiveness of joint-learning.

Depth to Image. We then train the depth-conditioned image generation network and use the same studio dataset for evaluation. As shown in Tab. 2, our approach has better performance on all image metrics, L1 error, SSIM, LPIPS and FID. The samples in the Fig. 4 also highlight how joint learning of depth enhances appearance generation, particularly in facial expression and shape accuracy.

Now that we have highlighted the benefit of joint over separate learning, we continue with illustrating applications and comparisons using our joint learning method.

5.3. Depth Estimation

Recently there has been great interest in fine-tuning foundational models to predict depth from monocular RGB input [13, 15, 26, 62, 62, 63]. As illustrated in Sec. 4, our model can readily be used for the task of monocular depth estimation (or RGB conditioned depth prediction) after a light fine-tuning. We use the same test set as in Sec. 5.2 and compare our method against state-of-the-art monocular depth estimators including Marigold [26], Geowizard [13], DepthAnything V2 [63] and three different backbones from the human-centric foundation model Sapiens [27]. Quantitative results are listed in Tab. 1 and qualitative results are shown in Fig. 5. As we see in Tab. 1, our method outperforms all other models on this task, including the Sapiens-1B model. Qualitatively our method captures the facial shape and expression similar to Sapiens-1B, while containing significantly fewer grid-like artifacts. We also present qualitative results for monocular depth estimation on in-the-wild face portraits (Fig. 6), and compare our estimated depth to the result of fitting a 3D morphable model [6, 7] to the input RGB image as in Sec. 5.2. Our results on unseen in-the-wild images have better mouth and face structure when compared to the 3DMM fit, and

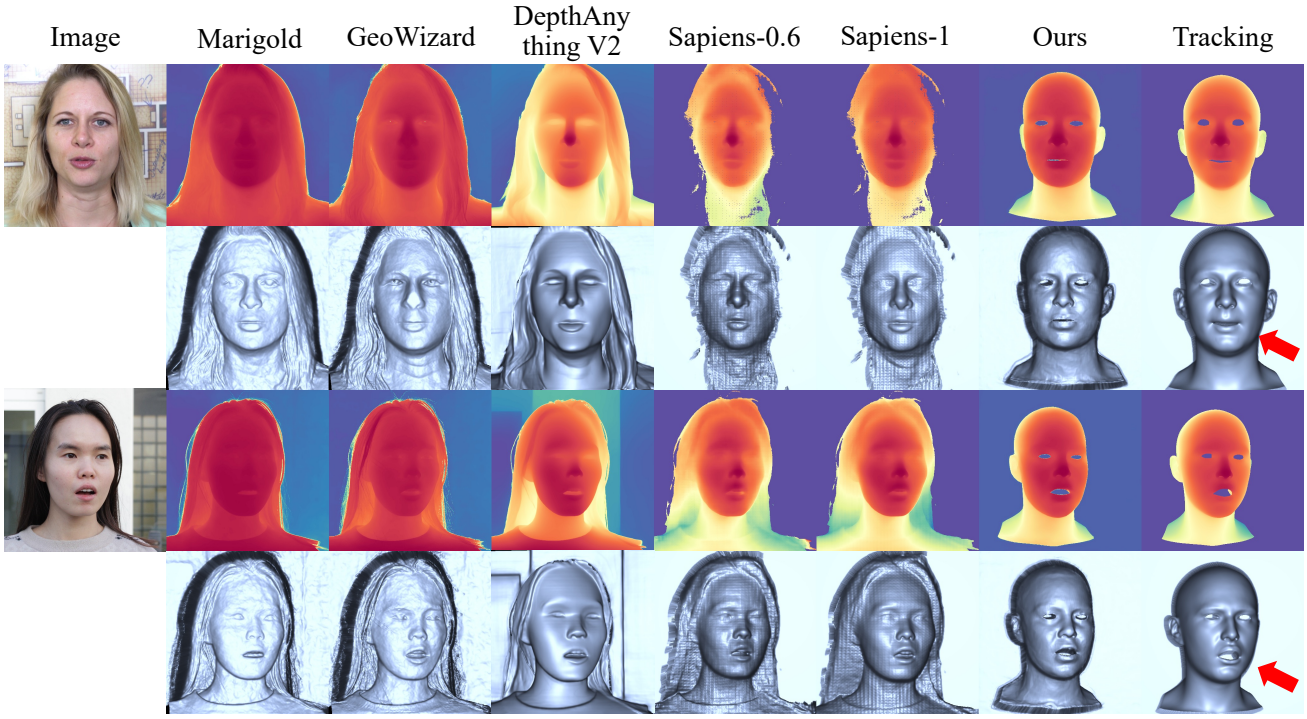

Figure 6. Qualitative comparisons for monocular depth estimation on wild faces. Note that even 3D facial tracking (right column) can sometimes fail. Our method can achieve a better depth due to the high-quality studio data as a subset of our training data.

correspond better to the RGB image. This is due to the fact that our method learned accurate depth correlation from the joint learning and combined studio training data.

5.4. Depth-Conditioned Portrait Editing

In addition to monocular depth estimation, the joint learning of RGB and depth modalities also enables us to generate an RGB image by providing a depth map as input. This application of our model can be particularly useful in having precise control over the generated RGB image for editing applications. In Fig. 7, we show examples of editing an RGB image, by modifying its corresponding depth map, and requiring our model to re-generate an RGB image corresponding to the edited depth. The inpainting mask spatially guides the model to the regions it is expected to modify in the given image. Our approach generates photo real images that respect the identity of the original RGB image and the edited depth maps.

5.5. Depth-Driven Portrait Animation

As illustrate in Sec. 4, our approach enables portrait animation by involving temporal layers. Here we show comparisons with concurrent diffusion-based portrait animation methods, including AniPortrait [56], EchoMimic [8], X-Portrait [58], Follow-Your-Emoji [35] and MegActor-Σ [64]. The qualitative results are presented

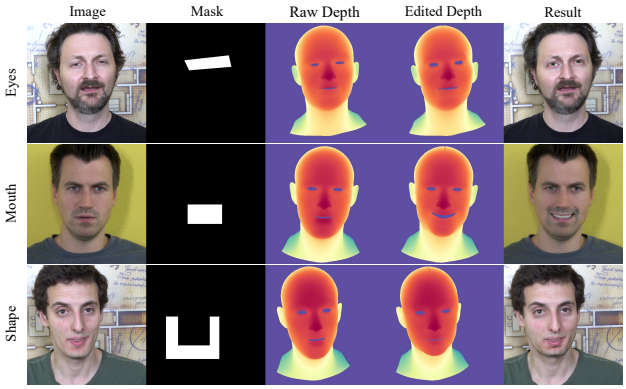

Figure 7. Depth-based face editing results.

in Fig. 8. We find that previous work suffers from artifacts, especially under large variations on the head pose or expressions. Additionally, previous condition-based methods struggle to have fine control over the head pose and facial expression. For quantitative rsults, we evaluate all the methods on the test video data from HDTF and VFHQ. For self reenactment task, we use four metrics to assess the video quality including L1 error, SSIM [55] and LPIPS [68] and FVD [49]. We also evaluate cross reenactment on three metrics, ArcFace score [10] to

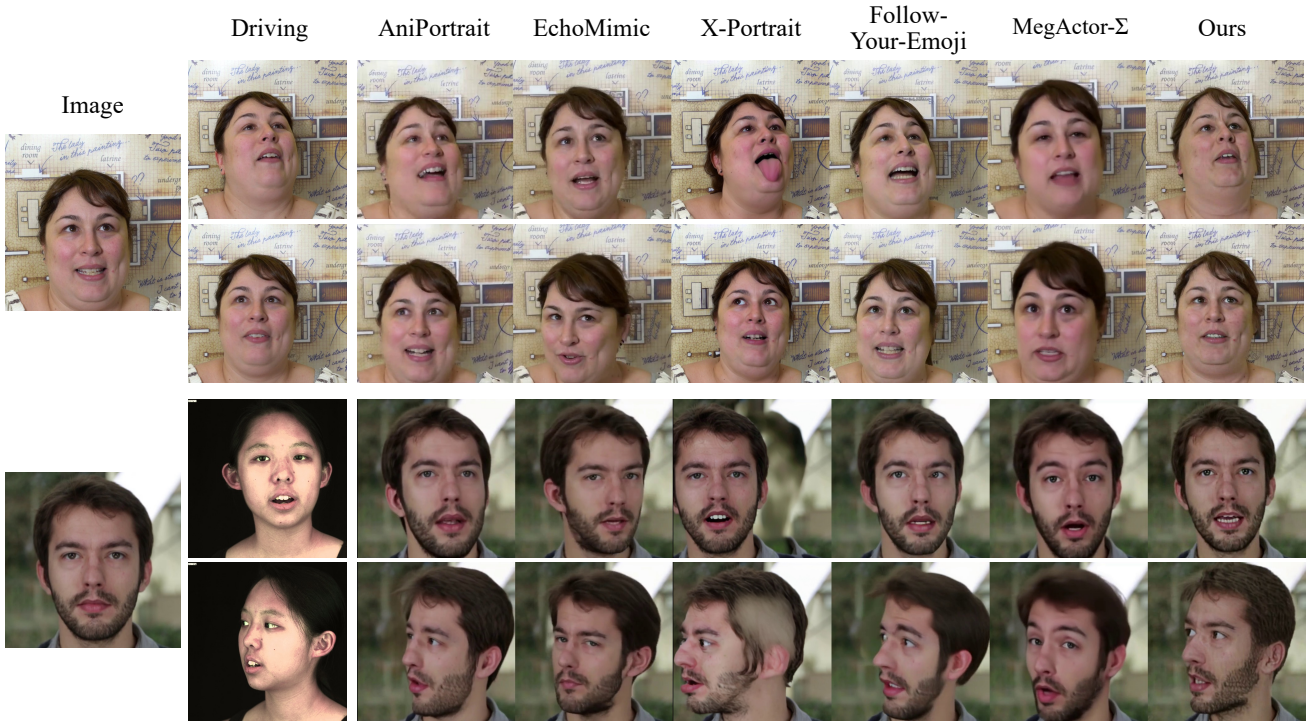

Figure 8. Qualitative comparisons with the state-of-the-art methods for portrait animation.

Method	Self Reenactment				Cross Reenactment		
	L1 ↓	SSIM ↑	LPIPS ↓	FVD ↓	ID Similarity ↑	Image Quality ↑	Expression/Pose ↓
AniPortrait [56]	0.049	0.732	0.197	204.194	0.763	51.381	0.71/2.99
EchoMimic [8]	0.085	0.616	0.311	516.652	0.727	44.902	0.65/4.94
X-Portrait [58]	0.078	0.624	0.284	233.384	0.766	51.339	0.68/5.50
Follow-Your-Emoji [35]	0.042	0.755	**0.145**	**180.305**	0.794	50.107	0.67/3.07
MegActor-Σ [64]	0.081	0.619	0.310	202.550	0.672	42.980	0.64/4.79
Ours	**0.041**	**0.760**	0.152	192.630	**0.798**	**54.324**	**0.61/2.59**

Table 3. Quantitative comparisons with state-of-the-art methods on self reenactment and cross reenactment tasks.

measure identity similarity, HyperIQA [69] to assess the image quality, and the extracted facial blendshapes and head poses from a 3DMM model [6, 7] to access the expression and pose accuracy. The results are presented in Tab. 3. Our approach demonstrates comparable performace on self-reenactment and surpasses other methods across all cross-reenactment metrics. Notably, our model achieves competitive performance while requiring significantly less training data than compared methods.

5.6. Limitations

Although our method is not limited to the facial skin region in principle, as it currently relies on depth maps derived from registered 3D geometry for training, it can only predict depth maps only for the skin region when given a new RGB image. Secondly, since our model uses depth maps for portrait manipulation, it can't handle fine-grained movements such as eye gaze or hair dynamics. Thridly, due to our modest computational resources, we were limited from scaling our training datasets and training times to those comparable with existing portrait animation methods. Therefore our current results could also be improved with longer training on larger datasets.

6. Conclusion

In this work we propose a new generative model for face portrait images, with a focus on jointly learning the visual appearance and the 3D depth in a unified framework. To accomplish this task we introduce a new diffusion-based architecture and corresponding training scheme, which ensures correlation between the two different output signals. Here we have demonstrated our joint-learning is

superior than spearate individual network through extensive experiments. Further applications are also possible with our joint learning framework, which believe advances the state-of-the-art in portrait depth estimation and image animation.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021. 2

[2] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[3] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 5

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1

[5] Prashanth Chandran, Derek Bradley, Markus Gross, and Thabo Beeler. Semantic deep face models. In *2020 international conference on 3D vision (3DV)*, pages 345–354. IEEE, 2020. 5

[6] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, and Derek Bradley. Continuous landmark detection with 3d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16858–16867, 2023. 5, 6, 8, 1

[7] Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, and Derek Bradley. Infinite 3d landmarks: Improving continuous 2d facial landmark detection. In *Computer Graphics Forum*, page e15126. Wiley Online Library, 2024. 5, 6, 8, 1

[8] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 7, 8

[9] Yuhao Cheng, Zhuo Chen, Xingyu Ren, Wenhan Zhu, Zhengqin Xu, Di Xu, Changpeng Yang, and Yichao Yan. 3d-aware face editing via warping-guided latent direction learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 916–926, 2024. 1

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 7

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1

[12] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2663–2671, 2022. 1

[13] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2025. 1, 2, 4, 6

[14] Jiazhi Guan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Shengyi He, Zhiliang Xu, Haocheng Feng, Errui Ding, Jingdong Wang, Hongtao Xie, et al. Talk-act: Enhance textural-awareness for 2d speaking avatar reenactment with diffusion model. *arXiv preprint arXiv:2410.10696*, 2024. 2

[15] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9233–9243, 2023. 2, 6

[16] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 1

[17] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023. 1

[18] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1, 2, 4, 5

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[20] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1, 2, 4

[21] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 2

[22] Jianwen Jiang, Gaojie Lin, Zhengkun Rong, Chao Liang, Yongming Zhu, Jiaqi Yang, and Tianyun Zhong. Mobileportrait: Real-time one-shot neural head avatars on mobile devices. *arXiv preprint arXiv:2407.05712*, 2024. 1

[23] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633. IEEE, 2023. 2

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving

the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[26] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 1, 2, 4, 6

[27] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2025. 2, 4, 6

[28] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 1

[29] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25096–25106, 2024. 1

[30] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Xian Liu, Jian Ren, Aliaksandr Siarohin, Ivan Skorokhodov, Yanyu Li, Dahua Lin, Xihui Liu, Ziwei Liu, and Sergey Tulyakov. Hyperhuman: Hyper-realistic human generation with latent structural diffusion. *arXiv preprint arXiv:2310.08579*, 2023. 2

[32] Cui Liyuan, Xu Xiaogang, Dong Wenqi, Yang Zesong, Bao Hujun, and Cui Zhaopeng. Cfsynthesis: Controllable and free-view 3d human video synthesis. *arXiv preprint arXiv:2412.11067*, 2024. 1

[33] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 2

[34] Yuanxun Lu, Jingyang Zhang, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, Xun Cao, and Yao Yao. Direct2. 5: Diverse text-to-3d generation via multi-view 2.5 d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8744–8753, 2024. 2

[35] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 1, 7, 8

[36] Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4):43–1, 2021. 1

[37] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[40] Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun BR, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, et al. Lite2relight: 3d-aware single image portrait relighting. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024. 1

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 4

[42] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[44] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 1

[45] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2

[46] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 1

[47] Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023. 5

[48] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 1

[49] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[50] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[51] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[52] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021. 1

[53] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 1

[54] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 539–547, 2015. 2

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[56] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 7, 8

[57] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 5

[58] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 1, 7, 8

[59] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J Black. Econ: Explicit clothed humans optimized via normal integration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 512–523, 2023. 2

[60] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 1

[61] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 1, 2, 4

[62] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 2, 6

[63] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2025. 4, 6

[64] Shurong Yang, Huadong Li, Juhao Wu, Minhao Jing, Linze Li, Renhe Ji, Jiajun Liang, and Haoqiang Fan. Megactor: Harness the power of raw video for vivid portrait animation. *arXiv preprint arXiv:2405.20851*, 2024. 1, 7, 8

[65] Dongxu Yue, Qin Guo, Munan Ning, Jiaxi Cui, Yuesheng Zhu, and Li Yuan. Chatface: Chat-guided real face editing via diffusion latent space manipulation. *arXiv preprint arXiv:2305.14742*, 2023. 1

[66] Jingyang Zhang, Shiwei Li, Yuanxun Lu, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, and Yao Yao. Jointnet: Extending text-to-image diffusion for dense distribution modeling. *arXiv preprint arXiv:2310.06347*, 2023. 2

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[68] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[69] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081, 2023. 8

[70] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 5

[71] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 1

Joint Learning of Depth and Appearance for Portrait Images

Supplementary Material

In this supplementary material, we provide more details about the network architecture and data processing. More information on our training details, more results of our method, and an ablation study are also provided. We recommend watching the supplementary video for even more results.

7. Implementation Details

We describe more implementation details on the network architecture and the training details used in Sec. 3 of the main paper.

7.1. Network Architecture

Here, we present the details of our network architecture for portrait animation. For the depth-driven animation task, we add a temporal-attention module in each block.

- **Temporal-Attention Module.** We use the same Temporal-Attention layers as in recent advances [48]. This module is designed to ensure smooth transitions across synthesized frames. To capture the dependencies between consecutive frames, we apply self-attention mechanisms on the temporal dimension of the features. Specifically, we first reshape the input feature $F \in \mathbb{R}^{b\times c\times f\times h\times w}$, where b, c, f, h, w represent the batch size, feature channel, the number of the generated frames in a sequence and the height and width of the feature map, to $F \in \mathbb{R}^{(b\times h\times w)\times c\times f}$. Then we apply self-attention across the temporal dimension f. However, motion consistency can only be guaranteed inside each sequence in this way, constraining the application for long video generation. Therefore, we draw inspiration from existing works [52] and take the last n generated frames from the preceding sequence as the motion frames. Here, we first feed these motion frames into the ReferenceNet to extract multi-resolution motion features. Then in each block, we concatenate the temporal module input and the motion feature along the temporal dimension f to get the self-attention layer input. In this way, the motion information from the previous sequence can be involved, so to ensure the coherence among different clips.

7.2. Training Details

As described in Sec. 3, our method contains three different training stages. In the first stage, we use the multi-view studio face dataset along with the in-the-wild dataset to train our joint-learning network. In the second stage, we slightly fine tune our model by incorporating different masks for the inpainting task. Here we design asymmetric masks for the RGB and depth branches, so that after fine-tuning, our model is capable of down-stream tasks including depth estimation, relighting and depth-based image editing. Note that our model is still capable of jointly generating RGB and depth by setting the mask to be pure white for both domains. In the third stage, we extend our first stage model to the depth-driven animation task by incorporating temporal attention modules. Only in-the-wild datasets are employed in this stage due to the lack of video data in studio face datasets. As mentioned in Sec. 5.1, we fix the parameters from the first stage while training. The training is performed on 4x RTX 4090 GPUs, and it takes about 3 days for the joint-learning model training, 15 hours for the inpainting fine-tuning and 2 days for the motion module training respectively.

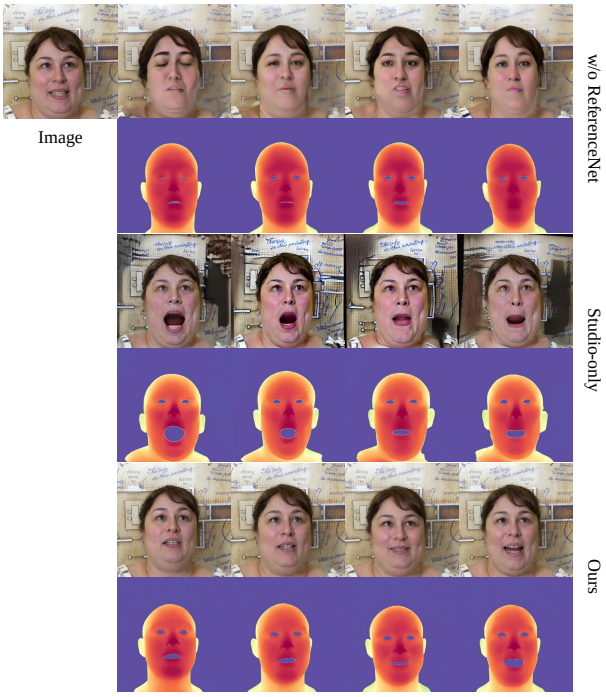

Figure 9. Ablation studies to show the effect of our architecture without the ReferenceNet (top), and trained on studio-only data (middle), compared to our proposed method (bottom).

7.3. Data Processing

In order to obtain the corresponding depth map for monoculor in-the-wild datasets, we use an off-the-shelf face tracking tool [6, 7] to fit a face mesh for each frame and then render out the depth map. The fitted mesh is represented by the blendshape weights of a PCA-based face

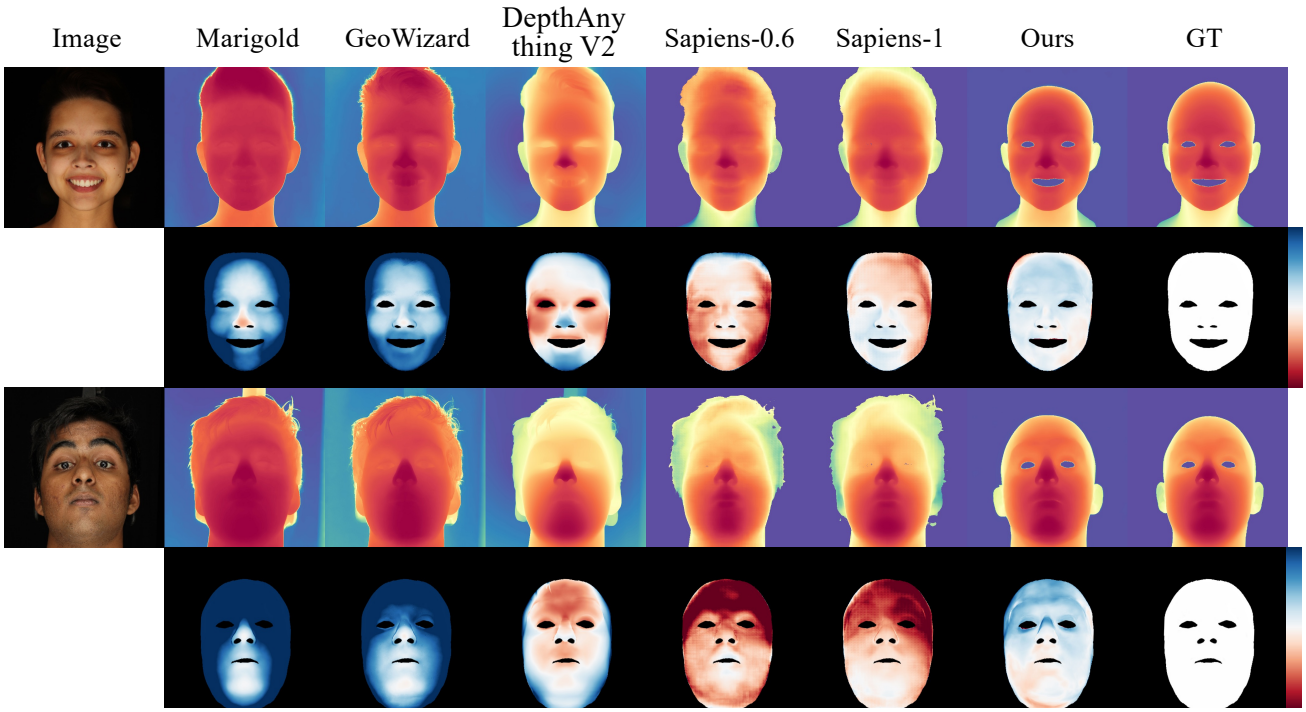

Figure 10. Qualitative comparisons with the state-of-the-art methods for monocular depth estimation on studio faces. We also show error map under facial mask for each sample. Note that here white means no error.

model, which includes 50 eigen faces for identity and 25 for expression. A landmark loss and a photometric loss are utilized to optimize the weights. To ensure the stability and smoothness of the tracking on a video sequence, we solve a global identity code for each clip.

8. More Results

8.1. Depth Estimation

As demonstrated in Sec. 5.3, we apply a mask on the face skin region while calculating the depth estimation metrics with ground truth. In Fig. 10, we show more qualitative results along with the error map under the mask. Here we set the error range to be -0.1 to 0.1. As demonstrated in the figure, our method outperforms other methods for generating depth maps with accurate geometry under various expressions and poses.

8.2. Image Relighting

One benefit of our joint learning of appearance and depth is that the facial depth can be used for downstream tasks like portrait relighting. Fig. 11 illustrates an example where the generated depth maps are used to compute surface normals for basic lighting changes in the generated image. Here the normals are used for rendering a diffuse shading layer that is multiplied with the image as a post-process.

Figure 11. Portrait image relighting is possible using our generated depth map.

9. Ablation Studies

We first evaluate the influence of the ReferenceNet on the quality of our generated results. We train a version of our network where we remove the ReferenceNet, and instead provide the latent reference RGB image as an additional input to the denoising U-Net. After training, we

jointly generate RGB images and depth maps from multiple different noise inputs, which are shown in the first row of Fig. 9. The generations without the ReferenceNet fail to capture the identity of the reference image, highlighting its importance in our architecture.

Secondly we also evaluate the importance of training our method on both studio data with ground truth depth, and in-the-wild data with pseudo ground truth depth. We first verify whether our method trained only on studio data can generalize to unseen in-the-wild identities. As we seen in the second row of Fig. 9, training only on studio data results in poor generalization to in-the-wild data and degrades the visual quality of the generated RGB images. However training with data from both studio and in-the-wild sources, results in the best performance as we see in the last row of the Fig. 9. This is also confirmed by our quantitative results in Tab. 1.