

FlexAvatar: Learning Complete 3D Head Avatars with Partial Supervision

Tobias Kirschstein¹ Simon Giebenhain¹ Matthias Nießner¹
 Technical University of Munich¹

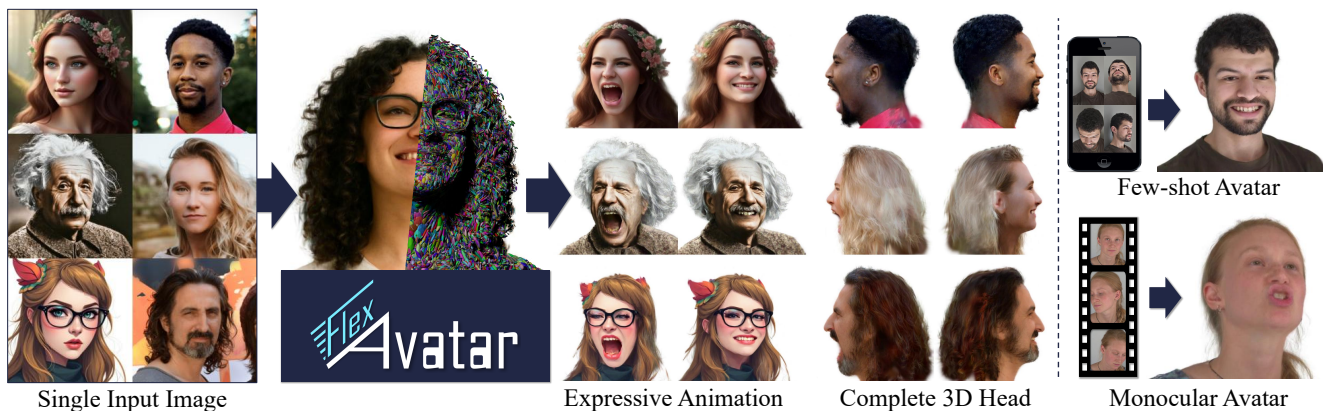


Figure 1. **FlexAvatar**. From just a single portrait image of a person, FlexAvatar creates a high quality 3D head avatar representation that can be freely animated and rendered from diverse viewpoints. Our model can be flexibly applied to other scenarios including creating avatars from a phone scan or from monocular videos. The entire avatar creation process can be executed within minutes.

Abstract

We introduce *FlexAvatar*, a method for creating high-quality and complete 3D head avatars from a single image. A core challenge lies in the limited availability of multi-view data and the tendency of monocular training to yield incomplete 3D head reconstructions. We identify the root cause of this issue as the entanglement between driving signal and target viewpoint when learning from monocular videos. To address this, we propose a transformer-based 3D portrait animation model with learnable data source tokens, so-called bias sinks, which enables unified training across monocular and multi-view datasets. This design leverages the strengths of both data sources during inference: strong generalization from monocular data and full 3D completeness from multi-view supervision. Furthermore, our training procedure yields a smooth latent avatar space that facilitates identity interpolation and flexible fitting to an arbitrary number of input observations. In extensive evaluations on single-view, few-shot, and monocular avatar creation tasks, we verify the efficacy of *FlexAvatar*. Many existing methods struggle with view extrapolation while *FlexAvatar* generates complete 3D head avatars with realistic facial animations.

Website: <https://tobias-kirschstein.github.io/flexavatar/>

1. Introduction

3D head avatars have many exciting applications in immersive teleconferencing, virtual try-on, personalized video games, or education. Ideally, users can create high-quality animatable 3D head avatars from one or a few input images without expensive capture equipment or long optimization times. The avatars could even be generated from text descriptions using existing text-to-image methods.

However, creating a high-quality 3D head avatar from just a single image is extremely challenging because it is underconstrained in two regards: (i) There are many unobserved regions complicating accurate 3D reconstruction. (ii) The model must infer realistic facial animation for a person without having seen any facial expressions of them. These issues are typically addressed by using multi-view video recordings for training, but those are hard to obtain for sufficiently many persons. Many existing approaches therefore rely on monocular portrait video datasets scraped from the internet because they offer broad identity coverage and in-the-wild variability. A natural disadvantage of these datasets is that they provide only a single viewpoint per identity and typically have a strong front-view bias. As a result, models trained solely on monocular data tend to reconstruct incomplete 3D heads.

Despite these challenges, existing works have successfully

trained single-image 3D head avatar pipelines, typically by relying heavily on geometric priors. The most common priors are 3D morphable models (3DMMs) such as FLAME [28] which provide a coarse but animatable head geometry. In these approaches, the predicted 3D primitives, such as meshes, radiance fields, or Gaussians, are typically rigged to the 3DMM, using its deformation field to drive facial motion. This approach reduces overfitting on monocular training data but limits expressiveness to the 3DMM’s predefined expression space. Still, many methods struggle with novel-view rendering.

We identify the underlying issue to be the entanglement of driving signal and target viewpoint in monocular training data. More specifically, models exploit the fact that in a monocular self-reenactment setting, the control for the facial expression is derived on the ground-truth target image itself, encouraging the model to guess the viewpoint from the expression input. Simply mixing monocular and multi-view training data does not prevent this behavior. We therefore introduce a transformer-based 3D portrait animation module with bias sinks that explicitly separate the model’s behavior on the two dataset types. In practice, we feed learnable tokens into the transformer depending on whether a training sample stems from a monocular or a multi-view dataset. During inference, we simply use the multi-view token, prompting the model to produce a complete 3D head regardless of the input image. We further avoid relying on a restrictive 3DMM and instead learn facial expressions directly from the data, yielding more flexible animation. Finally, to improve the quality of the renderings, we propose an upsampling architecture for the transformer based on a combination of PixelShuffle and StyleGAN [20] blocks. As a side product of our training, FlexAvatar learns a smooth latent space of 3D head avatars, allowing interpolation between identities and enabling flexible fitting to arbitrary numbers of input views. Therefore, our pipeline can be used not only in a single-input scenario but also in few-shot and monocular video avatar creation settings.

In summary, our contributions are as follows:

- A novel and efficient pipeline for creating high-quality 3D head avatars from a single image
- Learnable bias sinks that combine the strengths of monocular and multi-view training to provide both strong generalization and complete 3D head avatar reconstruction
- An efficient upsampler architecture based on StyleGAN2 and PixelShuffle for improved visual quality

2. Related Work

2.1. 3D Head Avatars from Sparse Observations

In 3D portrait animation, the goal is to predict 3D head avatars from a single image by utilizing 3D representations such as meshes [22], Neural Radiance Fields (NeRFs) [5,

10, 29, 30, 35, 48, 58] or 3D Gaussians (3DGS) [4, 14, 15, 21], which allows rendering of novel viewpoints. Many methods heavily rely on priors from 3D morphable models (3DMMs) such as FLAME [1, 28] for coarse geometry and animation of their 3D representation. For example, both LAM [15] and GAGAvatar [4] rig 3D Gaussians to the morphable FLAME mesh, inheriting its limited animation space. In contrast, our method avoids these limitations in expressiveness by learning facial motion directly from data.

Another line of work reconstructs avatars from one or a few observations of a person. Regression-based methods [25] can provide an avatar near instantly but often struggle to generalize to out-of-domain inputs or varying numbers of observations. Distillation-based methods [46, 47, 54] instead use a pre-trained multi-view image or video generation network to synthesize additional views, which are then used to reconstruct a high-quality avatar [41]. While this generally improves quality, distillation is inherently slow due to the cost of invoking image or video generation. Our method generalizes well to any image domain while reconstructing high-quality avatars within minutes.

A different approach is to learn a photorealistic 3D head prior which can later be fitted to any set of input images [16, 55, 56, 60, 65]. These models are typically Autodecoder-based [38] and trained on multi-view data. Because multi-view recordings are limited, recent methods also leverage synthetic data [2, 44]. Our approach also learns a latent space of avatars which can be utilized for fitting to arbitrary observations. However, we use an encoder-decoder structure, avoiding the issue of growing dictionaries in Autodecoders and enabling fast inference.

2.2. Learnable Dataset Embeddings

Learnable embeddings or tokens are used in several settings. In task adaptation, they are used for parameter-efficient finetuning [18, 26, 43]. In multi-dataset training, dataset indicators help unify heterogeneous datasets into a shared feature space [34, 66]. Multi-modal transformers similarly use modality-specific embeddings to distinguish input types [17, 67]. In 3D reconstruction, NeRF-in-the-wild [32] learns a per-image embedding that captures aspects of the input that the subsequent generalized NeRF cannot explain. Similarly, methods like Nerfies [39] or Cafca [2] bake unwanted temporal variations of the input images into learnable embeddings.

The difference in our setting is that we introduce dataset-level embeddings to explicitly capture dataset-induced biases. This allows us to suppress these biases at inference time, enabling a model trained on mixed monocular and multi-view data to behave as if it were supervised by multi-view observations alone while keeping the generalization capabilities induced by the monocular training data.

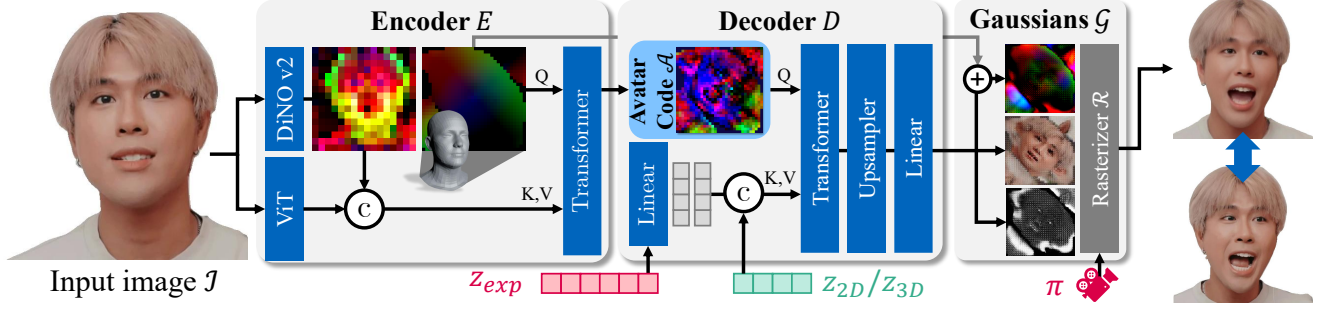


Figure 2. **Method Overview of FlexAvatar.** Given the single input image I , our method allows to change both viewpoint π and facial expression z_{exp} . The transformer-based encoder E first produces a compressed avatar code $\mathcal{A}$ via cross-attention. The decoder D then incorporates the effect of the facial expression z_{exp} into the avatar representation. Crucially, the corresponding bias sinks are concatenated to the expression tokens: z_{2D} if the input image I comes from a monocular dataset, and z_{3D} if it comes from a multi-view dataset. Finally, the upsampled avatar code is decoded into the 3D Gaussian attributes for rendering. During training, the bias sinks absorb data modality-specific biases such as the entanglement of driver expression and target viewpoint of monocular datasets. At inference time, only z_{3D} is used to inherit the disentangled behavior of multi-view datasets yielding both generalized and complete 3D head avatars.

3. Method

Given a single portrait image I , our goal is to create an animatable avatar representation $\mathcal{A}$ which we can control via animation codes z_{exp} and render from arbitrary viewpoints. A visual overview of our approach is depicted in Fig. 2. We adopt an encoder-decoder perspective and split the image synthesis process into multiple stages: (1) An Encoder E that finds a suitable avatar code $\mathcal{A}$ based on the input image I , (2) a decoder D that creates a set of articulated 3D Gaussians given an expression code z_{exp} , and (3) a renderer $\mathcal{R}$ which renders the 3D Gaussian representation from the desired viewpoint π :

$$\mathcal{A} = E(I) \quad (1)$$

$$\mathcal{G} = D(\mathcal{A}, z_{exp}) \quad (2)$$

$$I^{pred} = \mathcal{R}(\mathcal{G}, \pi) \quad (3)$$

In practice, we use the tile-based differentiable rasterizer from 3DGS [21] as $\mathcal{R}$ and expression codes z_{exp} from FLAME [28]. E and D are implemented via transformers [49], and $\mathcal{A} \in \mathbb{R}^{H_I \times W_I \times D}$ is a 2-dimensional latent code that lives in the UV-space of a template head mesh.

Crucially, the encoder-decoder design choice leads to the emergence of a smooth latent space of avatars during training. This enables applications that go beyond direct feed-forward prediction of a 3D head avatar from a single image (see Sec. 3.5).

3.1. Encoder E : Projecting onto an Avatar manifold

The general design of our encoder is inspired by LAM [15] with the focus on producing a compressed avatar representation. For this purpose, we employ a head template mesh with corresponding UV space which will host the avatar code’s features. We begin by first extracting image features

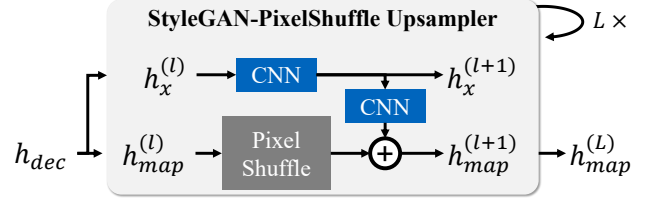


Figure 3. **Architecture of the StyleGAN-PixelShuffle block.**

f_{img} with a pre-trained DINOv2 [36] model and a shallow learnable ViT.

$$f_{img} = \text{MLP}([\text{DINO}(I), \text{ViT}([I, I^{pluck}])]) \quad (4)$$

where I^{pluck} are the plucker embeddings of the camera viewpoint of the input image I . To map the image features into the template’s UV space, we define queries Q anchored in UV space. This is done by uniformly sampling 3D surface positions in the UV space of the template mesh $\mathcal{T}$ and encoding them with sinusoidal frequencies:

$$x_{mesh}, x_{uv} \leftarrow \mathcal{T} \quad (5)$$

$$Q = \text{PE}(x_{mesh}) \quad (6)$$

Finally, we perform cross-attention from the UV-anchored queries Q to the image features f_{img} :

$$\mathcal{A} = \text{ATTENTION}(Q, f_{img}, f_{img}) \quad (7)$$

In practice, we use the attention implementation from MMDIT [11]. The result is a compact 2-dimensional latent code $\mathcal{A} \in \mathbb{R}^{H_I \times W_I \times D}$ that contains all relevant information from the input image but is agnostic to both viewpoint and facial expression.

3.2. Decoder D : Decode Articulated 3D Gaussians

The decoder’s goal is to incorporate the effect of facial expressions on the avatar representation and to produce the final 3D Gaussians for rendering. For animation modeling, we adopt the approach of Avat3r [25] and use cross-attention from the internal representation to a sequenced expression code $s_{exp} \in \mathbb{R}^{N_{exp} \times D}$. This model-free approach allows the network to learn facial animations from the data without being limited to the animation space of a pre-defined 3D face model:

$$s_{exp} = \text{MLP}(z_{exp}) \quad (8)$$

$$h_{dec} = \text{ATTENTION}(\mathcal{A}, s_{exp}, s_{exp}) \quad (9)$$

The expression code z_{exp} can be any description of the facial state, such as audio, 3DMM coefficients, or an image embedding derived from a driving image. In practice, we use the expression codes of FLAME [28]. However, note that our network design makes no assumptions about the structure of z_{exp} and can easily be applied to different driving signals.

The resulting decoder feature map $h_{dec} \in \mathbb{R}^{H_l \times W_l \times D}$ is then upsampled L times yielding $h_{map}^{(L)} \in \mathbb{R}^{L \cdot H_l \times L \cdot W_l \times \frac{D}{L^2}}$. This is crucial for decoding sufficiently many 3D Gaussians. Fig. 3 shows an overview of our upsampler design which uses a combination of PixelShuffle [45] and CNN blocks inspired by StyleGAN2 [20]:

$$h_x^{(l+1)} = \text{CNN}(h_x^{(l)}) \quad (10)$$

$$h_{map}^{(l+1)} = \text{PIXELSHUFFLE}(h_{map}^{(l)}) + \text{CNN}(h_x^{(l+1)}) \quad (11)$$

with $h_x^{(0)} = h_{map}^{(0)} = h_{dec}$. This is followed by bilinear grid sampling to extract one feature per Gaussian:

$$x = \text{GRIDSAMPLE}(h_{map}^{(L)}, x_{uv}) \quad (12)$$

where x_{uv} are the texel locations of the sampled points x_{mesh} on the template mesh. In practice, we use $L = 2$ upsampling steps and perform grid sampling with another 2x upsampling, yielding a total upsampling rate of 8x. The resulting features $x \in \mathbb{R}^{G \times \frac{D}{L^2}}$ hold information for each 3D Gaussian that are decoded with an MLP:

$$\mathcal{G} = \text{MLP}(x) \quad (13)$$

We also initialize the Gaussians’ positions on the template mesh surface x_{mesh} :

$$\mathcal{G}_{pos} \leftarrow \mathcal{G}_{pos} + x_{mesh} \quad (14)$$

The final 3D Gaussians $\mathcal{G}$ can then be rendered via the tile-based rasterizer of [21]:

$$I_{pred} = \mathcal{R}(\mathcal{G}, \pi) \quad (15)$$

In practice, we use the batched rendering implementation of gsplat [57] for better training performance.

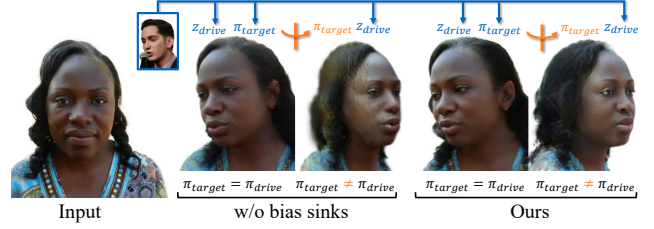


Figure 4. **Entanglement of driving signal and target viewpoint.** Naive training on monocular data works well as long as both expression code z_{drive} and rendering camera π_{target} are transferred to the avatar ($\pi_{target} = \pi_{drive}$). Artifacts occur when the rendering camera is moved, i.e., rendering and driving viewpoint differ ($\pi_{target} \neq \pi_{drive}$). This issue is fixed by our proposed bias sinks.

3.3. Fighting Entanglement with Bias Sinks

During training, 3D portrait animation models minimize an image loss $\mathcal{L}(f(I_{source}, z_{target}), I_{target})$ where the expression code $z_{target} = \text{TRACK}(I_{drive})$ is derived from a driving image that matches the target expression. In monocular video datasets, $I_{drive} = I_{target}$ since there is only a single camera available. In this case, the derived expression code z_{target} can leak information about the viewpoint π_{target} of the target image. The model may exploit this by predicting only a partial 3D head which is sufficient to satisfy the loss from that specific viewpoint. We refer to this failure mode as *entanglement of driving signal and target viewpoint*. Although acceptable when $\pi_{target} = \pi_{drive}$ (e.g., standard portrait animation), it breaks in applications requiring free-view rendering ($\pi_{target} \neq \pi_{drive}$), leading to incomplete heads as seen in Fig. 4.

Multi-view datasets break this entanglement by providing multiple viewpoints for the same facial expression, but they are too limited in scale for good generalization. To address this, we introduce bias sinks, which are two learnable tokens z_{2D} and z_{3D} that are concatenated to the expression code sequence s_{exp} before decoding:

$$s_{exp} \leftarrow [s_{exp}, z_{bias}] \quad (16)$$

During training, samples from monocular datasets use z_{2D} and multi-view samples use z_{3D} . This makes the decoder explicitly aware of a sample’s provenance absorbing the bias of a particular dataset type. In practice, the model learns to predict incomplete 3D heads whenever it sees the z_{2D} token and produces a complete avatar when z_{3D} is given. Crucially, this design still allows the model to share knowledge across dataset types. In particular, when feeding in z_{3D} , the model still benefits from the generalization obtained from the monocular video training. During inference, we always feed in the z_{3D} token to obtain both well generalized and complete 3D head avatars from a single image.

Task	#Inputs	Output assumption	Training data	Evaluation data	Fitting	Figures
3D Portrait Animation (§4.3)	1	$\pi_{target} = \pi_{drive}$	CelebV-Text	Ava256	VFHQ-Test	200 steps Tab. 2
Single-image Avatar Creation (§4.4)	1	$\pi_{target} \neq \pi_{drive}$	Hallo3	-	Ava256 ⁵ persons	200 steps Tab. 3, Fig. 5
Few-shot Avatar Creation (§4.5)	4	$\pi_{target} \neq \pi_{drive}$	NeRSemble	Ava256 ^{Avat3r} train	Ava256 ^{Avat3r} test	1000 steps Tab. 3
Monocular Avatar Creation (§4.6)	900	$\pi_{target} \neq \pi_{drive}$	Cafca	Ava256	NeRSemble Benchmark	2000 steps Tab. 4, Fig. 6

Table 1. **Overview of Experimental Results.** We evaluate FlexAvatar on 4 different tasks and 3 different datasets.

3.4. Training with Perceptual Losses

We use the L1 and SSIM losses from 3DGS:

$$\mathcal{L}_1 = \|I_{pred} - I_{target}\|_1 \quad (17)$$

$$\mathcal{L}_{SSIM} = 1 - \text{SSIM}(I_{pred}, I_{target}) \quad (18)$$

Inspired by PercHead [37], we additionally employ perceptual losses based on DINOv2 [36] and the Segment Anything Model (SAM) [42]:

$$\mathcal{L}_{DINO} = \|\text{DINO}_f(I_{pred}) - \text{DINO}_f(I_{target})\|_1 \quad (19)$$

$$\mathcal{L}_{SAM} = \|\text{SAM}_f(I_{pred}) - \text{SAM}_f(I_{target})\|_1 \quad (20)$$

where $\text{DINO}_f(\cdot)$ and $\text{SAM}_f(\cdot)$ extract intermediate feature maps of the given image. The final reconstruction loss is a combination of all terms:

$$\mathcal{L}_{rec} = \mathcal{L}_1 + \mathcal{L}_{SSIM} + \mathcal{L}_{DINO} + \mathcal{L}_{SAM} \quad (21)$$

3.5. Fitting A to Additional Observations

Often, more than one image of a person is available, e.g., a set of images $(\mathcal{I}^{many}, z_{exp}^{many}, \pi^{many})$ with corresponding expression codes and cameras. We can use our encoder E to get an initialization for $\mathcal{A}$ by using any one of the available observations:

$$\mathcal{A}^{init} = E(\mathcal{I}_0^{many}) \quad (22)$$

This initial estimate of the avatar can then be optimized by fitting it against all observations:

$$\mathcal{I}_{pred}^{many} = \mathcal{R}(D(\mathcal{A}^{init}, z_{exp}^{many}), \pi^{many}) \quad (23)$$

$$\mathcal{L}_{fit} = \mathcal{L}_{rec}(\mathcal{I}_{pred}^{many}, \mathcal{I}^{many}) \quad (24)$$

By minimizing $\mathcal{L}_{fit}$, one can obtain an animatable 3D head avatar representation $\mathcal{A}^{fit}$ that incorporates all available observations of the person. Crucially, we only make $\mathcal{A}^{init}$ learnable and keep the entire decoder $\mathcal{D}$ fixed to avoid overfitting on the sparse inputs.

This procedure is similar to how autoencoder-style photorealistic 3D head models such as GPHM [55], HeadGAP [65], or HeadNeRF [16] create an avatar of a person. However, our approach has two advantages: First, it can be trained on mostly monocular video datasets whereas autoencoder-style models typically require multi-view training. Second, our approach also has an encoder which speeds up the optimization process by providing already an initial guess of the latent avatar code.

4. Experimental Results

4.1. Training

We train FlexAvatar on 5 datasets: 2 monocular portrait video datasets (CelebV-Text [59] and Hello3 [6]), 2 multi-view datasets (NeRSemble [24] and Ava256 [33]), and the synthetic multi-view Cafca dataset [2]. We sample 40k clips from the monocular sets, use all Ava256 recordings, $\sim 25\%$ of NeRSemble, and neutral-expression frames from all Cafca identities. While monocular data provides generalization, NeRSemble and Ava256 offer high-quality expressions, and Cafca supplies full 360° supervision.

We extract cameras π and expression codes z_{exp} using Pixel3DMM [13]. For NeRSemble and Ava256, we only track the frontal camera. Training uses Adam [23] with a learning rate of $1e-4$. Perceptual losses are introduced after 400k steps to avoid early overfitting to high-frequency details. In total, the model is trained for 1M steps with a batch size of 20 on one A100 GPU, taking roughly 3 weeks.

4.2. Experiment Setup

Tasks. Tab. 1 shows an overview of our experiment setup. We evaluate FlexAvatar’s ability to create 3D head avatars in a variety of situations:

3D Portrait Animation. In this well-established task, the goal is to animate a portrait image by transferring both expression and head pose from a second image (which can be of a different person). In this setting, methods can exploit the entanglement of driving signal and target viewpoint since $\pi_{target} = \pi_{drive}$.

Single-image 3D Head Avatar Creation. Similar to 3D Portrait animation, a single image is given with the additional requirement to be able to freely change the camera viewpoint. In this setting, no connection between the driving signal and the rendering viewpoint can be assumed since $\pi_{target} \neq \pi_{drive}$.

Few-shot 3D Head Avatar Creation. In this task, 4 images of a person are provided with the goal to create a complete 3D head avatar that can be freely animated and rendered from any viewpoint.

Monocular 3D Head Avatar Creation. For the last task, one or several monocular videos of a person are available to create a 3D head avatar. We compare against recent state-of-the-art methods on the public leaderboard of the NeRSemble benchmark.

	Self Reenactment							Cross Reenactment		
	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	AED↓	APD↓	AKD↓	CSIM↑	AED↓	APD↓
GPAvatar [5]	21.04	0.807	0.150	0.772	0.132	0.189	4.226	0.564	0.255	0.328
Real3DPortrait [58]	20.88	0.780	0.154	0.801	0.150	0.268	5.971	0.663	0.296	0.411
Portrait4D [9]	20.35	0.741	0.191	0.765	0.144	0.205	4.854	0.596	0.286	0.258
Portrait4D-v2 [10]	21.34	0.791	0.144	0.803	0.117	0.187	3.749	0.656	0.268	0.273
GAGAvatar [4]	21.83	0.818	0.122	0.816	0.111	0.135	3.349	0.633	0.253	0.247
LAM [15]	22.65	0.829	0.109	0.822	0.102	0.134	2.059	0.651	0.250	0.356
Ours	23.47	0.837	0.099	0.830	0.075	0.010	2.965	0.663	0.223	0.026

Table 2. **3D Portrait Animation comparison on the VFHQ dataset.** We evaluate the ability to animate a single image by transferring facial motion and head pose from a driving video showing the same person (self-reenactment) or a different person (cross-reenactment).

		PSNR↑	SSIM↑	LPIPS↓	AKD↓	CSIM↑
		PSNR↑	SSIM↑	LPIPS↓	AKD↓	CSIM↑
Single-image	Portrait4Dv2 [10]	11.9	0.671	0.404	7.77	0.578
	LAM [15]	13.1	0.702	0.399	11.2	0.411
	GAGAvatar [4]	12.7	0.709	0.371	7.45	0.555
	Ours	16.9	0.762	0.265	5.52	0.695
Few-shot	InvertAvatar [64]	13.0	0.288	0.590	52.3	0.296
	GPAvatar [5]	20.0	0.700	0.291	5.72	0.341
	Avat3r [25]	20.8	0.715	0.310	5.66	0.616
	Ours	21.1	0.733	0.218	5.39	0.755

Table 3. **Single-image and Few-shot Avatar Creation comparison on the Ava256 dataset.**

Metrics. Across all our experiments, we employ three paired-image metrics to measure the quality of individual rendered images: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [50], and Learned Perceptual Image Patch Similarity (LPIPS) [63]. Furthermore, we make use of two face-specific metrics: Average Key-point Distance (AKD) measured in pixels with keypoints estimated from PIPNet [19], and cosine similarity (CSIM) of identity embeddings based on ArcFace [7]. Temporal consistency is measured with FovVideoVDP [31] (JOD) which is sensitive to flickering, noise and other temporal artifacts. Finally, we estimate 3DMM coefficients using the forward regressor of [8] to compute Average Expression Distance (AED) and Average Pose Distance (APD) by computing the L1 distance of the corresponding 3DMM coefficients.

4.3. 3D Portrait Animation

We follow the evaluation protocol of GAGAvatar [4] and evaluate both self-reenactment and cross-reenactment performance on the VFHQ test split [52]. The results can be seen in Tab. 2. Our method improves in all metrics except AKD over the previous state-of-the-art. This shows that FlexAvatar can generalize well to unseen persons and can animate portraits with different driving persons.

4.4. Single-image 3D Head Avatar Creation

We evaluate single-image 3D head avatar reconstruction on the Ava256 dataset [33]. We select one challenging sequence for 5 diverse subjects. The frontal frame of the first timestep serves as input, and we uniformly sample 10 target expressions from 4 distinct cameras per sequence, yielding 200 test images. Expression codes z_{exp} are extracted from the frontal view, unlike standard 3D portrait animation settings where the driving and target viewpoints coincide. This setup is more demanding as methods that exploit viewpoint information in z_{exp} are penalized. This evaluation better reflects real applications that require freely animating an avatar without assumptions about the rendering viewpoint.

Results in Tab. 3 and Fig. 5 show that our method substantially outperforms recent approaches, producing realistic, complete, and expressive 3D heads. For fairness, the entire Ava256 dataset is held out during training. Note that the publicly released version of LAM used in the comparison was trained on both monocular (VFHQ [52]) and multi-view (NeRsemble [24]) data. Hence, our gains cannot be attributed solely to multi-view supervision. Further analysis is provided in the ablation section.

4.5. Few-shot 3D Head Avatar Creation

Thanks to FlexAvatar’s smooth avatar latent space, we can seamlessly integrate multiple observations of a subject via fitting following Sec. 3.5. We evaluate this on the Ava256 dataset using the same protocol as Avat3r [25]: 4 input images of a person are provided, and the model must render a novel expression from a novel viewpoint. To build an avatar, we encode one of the 4 images to obtain an initial code $\mathcal{A}^{init}$ and then optimize it for 1000 steps (~ 7 minutes per avatar) to match all four inputs. For fair comparison, we train with Ava256 but exclude all test identities, following Avat3r. Metrics are computed on a subset of sequences where Pixel3DMM tracking succeeds. As shown in Tab. 3, our method outperforms Avat3r, particularly in sharpness (LPIPS) and identity preservation (CSIM).

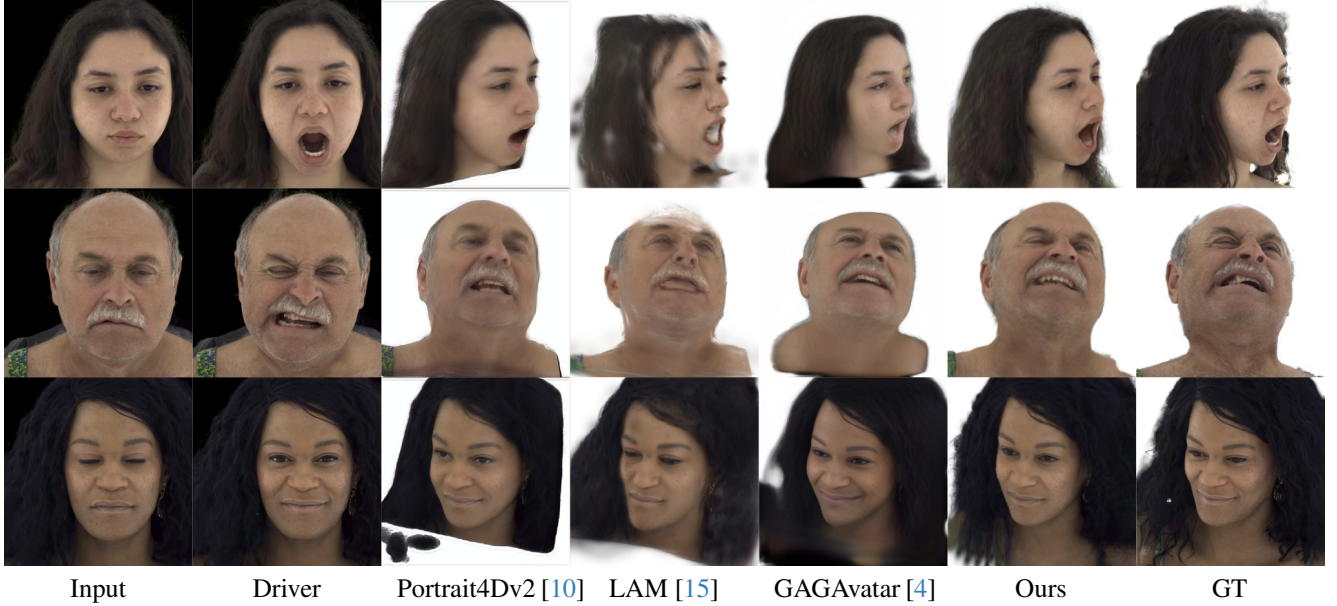


Figure 5. **Qualitative Single-image Avatar Creation comparison on the Ava256 dataset.** We compare our method to the recent state-of-the-art on 3D head avatar creation from a single portrait image. Our method produces more complete 3D head avatars and re-enacts the target expression more faithfully.

	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	JOD $\uparrow$	AKD $\downarrow$	CSIM $\uparrow$
INSTA [68]	15.8	0.771	0.344	4.83	5.22	0.631
FlashAvatar [51]	16.3	0.731	0.386	4.15	19.18	0.304
TaoAvatar [3]	18.2	0.789	0.267	5.28	5.50	0.715
FATE [62]	19.1	0.820	0.220	5.56	3.52	0.770
HRAvatar [61]	19.5	0.817	0.214	5.76	4.62	0.765
CAP4D [47]	19.8	0.821	0.185	5.79	4.19	0.793
RGBAvatar [27]	20.6	0.829	0.181	6.03	3.41	0.824
Ours	20.9	0.830	0.156	6.08	3.80	0.827

Table 4. **Monocular Avatar Creation comparison on the NeRSemble Benchmark.** We evaluate the ability to render novel views and novel expressions given monocular videos of 5 persons.

4.6. Monocular 3D Head Avatar Creation

Finally, we evaluate the scalability of FlexAvatar on the NeRSemble monocular 3D head avatar benchmark [24], which requires creating 3D avatars from video clips of 5 subjects. As in our few-shot experiments, we predict an initial avatar code $\mathcal{A}^{init}$ and fit it to 900 evenly sampled frames from the training videos for 2000 iterations (~ 10 minutes per avatar). No benchmark subject data is used during training. Results in Tab. 4 show that we outperform all baselines on nearly all metrics, with significant gains in sharpness (LPIPS). A visual comparison is shown in Fig. 6. Notably, our method surpasses CAP4D [47], that relies on a strong multi-view 3D head prior, while using fewer frames and achieving much faster fitting (10 minutes vs. 4 hours).



Figure 6. **Comparison on the NeRSemble Benchmark.**

4.7. Ablations

In Tab. 5 and Fig. 7, we present ablations of our dataset and architecture choices. The ablations are compared on the Ava256 dataset on the single-image 3D head avatar creation task. Crucially, we hold out the entire Ava256 dataset from training to measure performance on an unseen data domain.

Effect of Training Data. Training only on monocular data (*only 2D*) produces partial 3D heads due to entanglement between driving signal and target viewpoint. Multi-view training (*only 3D*) resolves this, yielding complete high-quality avatars, but generalization to unseen identities is poor, reflected in low CSIM scores.

Effect of bias sinks. Simply combining monocular and multi-view data (*w/o bias sinks*) does not produce complete 3D heads for unseen images. The model mainly learns to identify the dataset rather than resolve view-



Figure 7. **Qualitative Ablation of method components on the Ava256 dataset.**

	2D	3D	$\mathcal{B}$	$\mathcal{U}$	$\mathcal{F}$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	AKD $\downarrow$	CSIM $\uparrow$
only 2D	☒	☐	☐	☒	☐	13.7	0.736	0.358	6.59	0.593
only 3D	☐	☒	☐	☒	☐	13.2	0.699	0.378	10.4	0.119
w/o bias sinks	☒	☒	☐	☒	☐	14.5	0.747	0.351	5.98	0.583
w/o StyleGAN	☒	☒	☒	☐	☐	17.1	0.765	0.287	7.03	0.614
Ours ^{ref}	☒	☒	☒	☒	☐	17.2	0.768	0.285	6.34	0.621
Ours + fitting	☒	☒	☒	☒	☒	16.9	0.771	0.280	5.59	0.682

$\mathcal{B}$ = bias sinks $\mathcal{U}$ = StyleGAN-PixelShuffle Upsampler $\mathcal{F}$ = Fitting

Table 5. **Quantitative Ablation on the Ava256 dataset.** Ablation models are only trained for 500k iterations to save compute resources. Hence, the numbers for *Ours + fitting* differ slightly from *Ours* in Tab. 3 even though both use the same evaluation setup.

point-expression entanglement. Our full architecture (*Ours^{ref}*) successfully generates complete 3D heads.

Effect of StyleGAN-PixelShuffle upsampler. Replacing the StyleGAN-PixelShuffle block with standard PixelShuffle slightly reduces metrics and visual quality, particularly in sensitive facial regions like the eyes and mouth interior.

Effect of Fitting. A single forward pass already produces accurate avatars, but fitting further improves identity (CSIM), sharpness (LPIPS), and expression fidelity (AKD). Fitting is fast (~ 1 minute) as it optimizes only the avatar code $\mathcal{A}$ while keeping the network frozen.

4.8. Limitations.

While FlexAvatar generates high-quality and complete 3D head avatars from a single image, several limitations remain. First, lighting is baked from the input image, preventing explicit control. This can appear unnatural if the avatar is placed in a different virtual environment. Second, although the architecture is 3DMM-free, all experiments use FLAME expression codes, which limits fine details such as the tongue. However, thanks to its model-agnostic design, FlexAvatar can be trained with more expressive de-



Figure 8. **In-the-wild results.** We test FlexAvatar on highly diverse inputs and perform cross-reenactment.

scriptors, e.g., expression codes from implicit morphable models [12, 40] or features from generalized expression encoders [48, 53, 55].

5. Conclusion

We introduced FlexAvatar, a method for generating high-quality, complete 3D head avatars from a single image. Existing methods struggle with view extrapolation. We identify the entanglement between driving signal and target viewpoint in monocular training to be a key issue. To address this, we propose bias sinks, which combine the generalization of monocular datasets with the 3D completeness of multi-view supervision. Extensive experiments show that FlexAvatar generalizes well and produces realistic avatars. Its smooth latent space enables flexible applications, including few-shot avatar creation from phone scans or monocular videos. Our proposed design is quite general and makes little domain-specific assumptions. Extending it to different domains such as human bodies or generalized dynamic novel view synthesis is a promising research direction. Furthermore, we believe our findings on bias sinks may benefit other domains where scarce multi-view or 3D data has to be combined with partial supervision from monocular data.

Acknowledgements

This work was supported by the ERC Consolidator Grant Gen3D (101171131). We would also like to thank Angela Dai for the video voice-over and Karla Weighart for proof-reading.

References

- [1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 2
- [2] Marcel C Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, et al. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 2, 5
- [3] Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10723–10734, 2025. 7
- [4] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 37:57642–57670, 2024. 2, 6, 7
- [5] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. Gpavatar: Generalizable and precise head avatar from image (s). *arXiv preprint arXiv:2401.10215*, 2024. 2, 6
- [6] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21086–21095, 2025. 5
- [7] 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. 6
- [8] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 6
- [9] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 6
- [10] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *European Conference on Computer Vision*, pages 316–333. Springer, 2024. 2, 6, 7
- [11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3
- [12] Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Learning neural parametric head models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21003–21012, 2023. 8
- [13] Simon Giebenhain, Tobias Kirschstein, Martin Rünz, Lourdes Agapito, and Matthias Nießner. Pixel3dmm: Versatile screen-space priors for single-image 3d face reconstruction. *arXiv preprint arXiv:2505.00615*, 2025. 5
- [14] Chen Guo, Zhuo Su, Jian Wang, Shuang Li, Xu Chang, Zhaohu Li, Yang Zhao, Guidong Wang, and Ruqi Huang. Sega: Drivable 3d gaussian head avatar from a single image. *arXiv preprint arXiv:2504.14373*, 2025. 2
- [15] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025. 2, 3, 6, 7
- [16] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2, 5
- [17] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2
- [18] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pages 709–727. Springer, 2022. 2
- [19] Haibo Jin, Shengcai Liao, and Ling Shao. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *International Journal of Computer Vision*, 2021. 6
- [20] 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. 2, 4
- [21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3, 4
- [22] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer, 2022. 2
- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

- [24] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 5, 6, 7
- [25] Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12089–12100, 2025. 2, 4, 6
- [26] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 2
- [27] Linzhou Li, Yumeng Li, Yanlin Weng, Youyi Zheng, and Kun Zhou. Rgbavatar: Reduced gaussian blendshapes for online modeling of head avatars. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10747–10757, 2025. 7
- [28] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2, 3, 4
- [29] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17978, 2023. 2
- [30] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems*, 36:47239–47250, 2023. 2
- [31] Rafał K Mantiuk, György Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021. 6
- [32] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 2
- [33] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, et al. Codec avatar studio: Paired human captures for complete, driveable, and generalizable avatars. *Advances in Neural Information Processing Systems*, 37:83008–83023, 2024. 5, 6
- [34] Lingchen Meng, Xiyang Dai, Yinpeng Chen, Pengchuan Zhang, Dongdong Chen, Mengchen Liu, Jianfeng Wang, Zuxuan Wu, Lu Yuan, and Yu-Gang Jiang. Detection hub: Unifying object detection datasets via query adaptation on language embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11402–11411, 2023. 2
- [35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2
- [36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 5
- [37] Antonio Oroz, Matthias Nießner, and Tobias Kirschstein. Perchead: Perceptual head model for single-image 3d head reconstruction & editing. *arXiv preprint arXiv:2511.02777*, 2025. 5
- [38] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2
- [39] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5865–5874, 2021. 2
- [40] Rolandos Alexandros Potamias, Stathis Galanakis, Jiankang Deng, Athanasios Papaioannou, and Stefanos Zafeiriou. Im-head: A large-scale implicit morphable model for localized head modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10196–10206, 2025. 8
- [41] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 2
- [42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5
- [43] Mark Sandler, Andrey Zhmoginov, Max Vladymyrov, and Andrew Jackson. Fine-tuning image transformers using learnable memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12155–12164, 2022. 2
- [44] Jack Saunders, Charlie Hewitt, Yanan Jian, Marek Kowalski, Tadas Baltrusaitis, Yiye Chen, Darren Cosker, Virginia Estellers, Nicholas Gydé, Vinay P Nambodiri, et al. Gasp: Gaussian avatars with synthetic priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 271–280, 2025. 2
- [45] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 4
- [46] Felix Taubner, Ruihang Zhang, Mathieu Tuli, Sherwin Bahmani, and David B Lindell. Mvp4d: Multi-view portrait

- p>video diffusion for animatable 4d avatars.
- arXiv preprint arXiv:2510.12785*
- , 2025. 2
- [47] Felix Taubner, Ruihang Zhang, Mathieu Tuli, and David B Lindell. Cap4d: Creating animatable 4d portrait avatars with morphable multi-view diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5318–5330. IEEE Computer Society, 2025. 2, 7
 - [48] Phong Tran, Egor Zakharov, Long-Nhat Ho, Liwen Hu, Adilbek Karmanov, Aviral Agarwal, McLean Goldwhite, Ariana Bermudez Venegas, Anh Tuan Tran, and Hao Li. Voodoo xp: Expressive one-shot head reenactment for vr telepresence. *arXiv preprint arXiv:2405.16204*, 2024. 2, 8
 - [49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3
 - [50] 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. 6
 - [51] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 7
 - [52] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 657–666, 2022. 6
 - [53] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024. 8
 - [54] Sicheng Xu, Guojun Chen, Jiaolong Yang, Yizhong Zhang, Yu Deng, Stephen Lin, and Baining Guo. Vasa-3d: Lifelike audio-driven gaussian head avatars from a single image. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2
 - [55] Yuelang Xu, Lizhen Wang, Zerong Zheng, Zhaoqi Su, and Yebin Liu. 3d gaussian parametric head model. In *European Conference on Computer Vision*, pages 129–147. Springer, 2024. 2, 5, 8
 - [56] Haotian Yang, Mingwu Zheng, Chongyang Ma, Yu-Kun Lai, Pengfei Wan, and Haibin Huang. Vrmr: A volumetric relightable morphable head model. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
 - [57] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, et al. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025. 4
 - [58] Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024. 2, 6
 - [59] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023. 5
 - [60] Zhixuan Yu, Ziqian Bai, Abhimitra Meka, Feitong Tan, Qiangeng Xu, Rohit Pandey, Sean Fanello, Hyun Soo Park, and Yinda Zhang. One2avatar: Generative implicit head avatar for few-shot user adaptation. *arXiv preprint arXiv:2402.11909*, 2024. 2
 - [61] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Kangjie Chen, Minghan Qin, Yu Li, and Haoqian Wang. Hravator: High-quality and relightable gaussian head avatar. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26285–26296, 2025. 7
 - [62] Jiawei Zhang, Zijian Wu, Zhiyang Liang, Yicheng Gong, Dongfang Hu, Yao Yao, Xun Cao, and Hao Zhu. Fate: Full-head gaussian avatar with textural editing from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5535–5545, 2025. 7
 - [63] 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. 6
 - [64] Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, Jinli Suo, and Yebin Liu. Invertavatar: Incremental gan inversion for generalized head avatars. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–10, 2024. 6
 - [65] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiye Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, et al. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. In *2025 International Conference on 3D Vision (3DV)*, pages 946–957. IEEE, 2025. 2, 5
 - [66] Zangwei Zheng, Xiangyu Yue, Kai Wang, and Yang You. Prompt vision transformer for domain generalization. *arXiv preprint arXiv:2208.08914*, 2022. 2
 - [67] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16804–16815, 2022. 2
 - [68] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4574–4584, 2023. 7