# EMOPortraits: Emotion-enhanced Multimodal One-shot Head Avatars

Nikita Drobyshev Antoni Bigata Casademunt Konstantinos Vougioukas Zoe Landgraf
Stavros Petridis Maja Pantic
{n.drobyshev, a.bigata-casademunt22, k.vougioukas,
zoe.landgraf15, stavros.petridis04, m.pantic}@imperial.ac.uk
Imperial College London

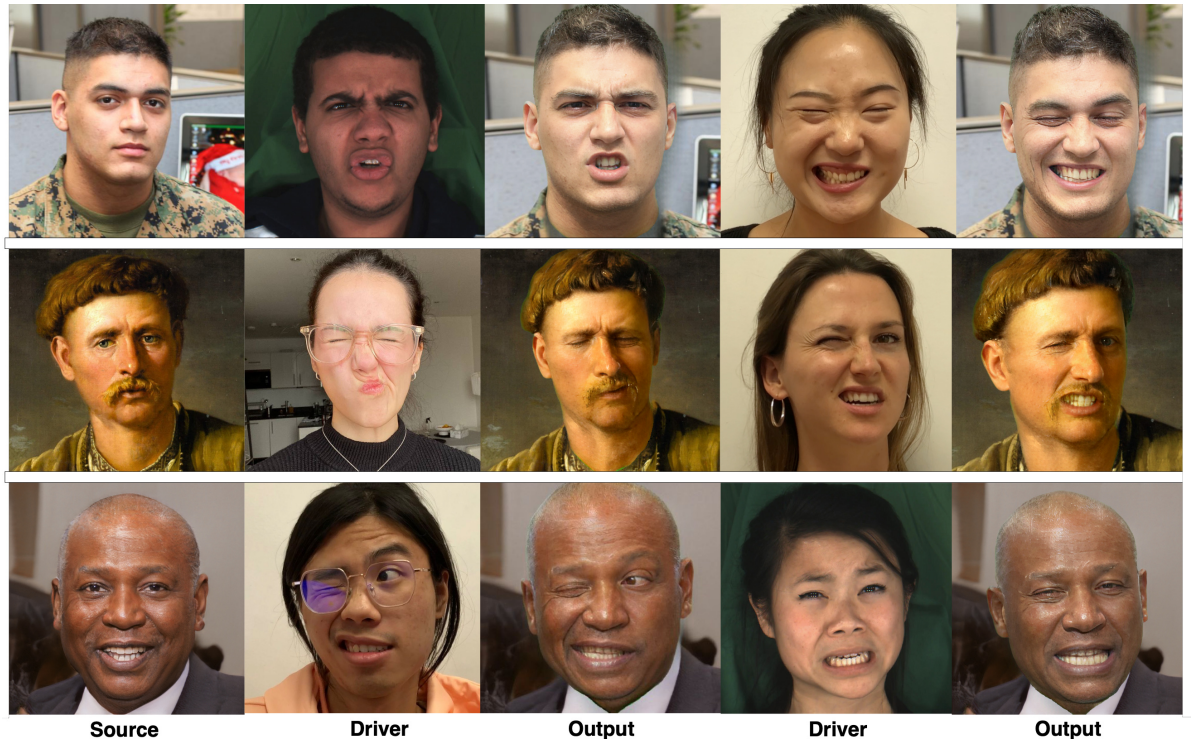


Figure 1. Selected animation results for image-driven mode of our method.

## Abstract

*Head avatars animated by visual signals have gained popularity, particularly in cross-driving synthesis where the driver differs from the animated character, a challenging but highly practical approach. The recently presented MegaPortraits model has demonstrated state-of-the-art results in this domain. We conduct a deep examination and evaluation of this model, with a particular focus on its latent space for facial expression descriptors, and uncover several limitations with its ability to express intense face motions. To address these limitations, we propose substantial changes in both training pipeline and model architecture, to introduce our* ***EMOPortraits*** *model, where we:*

*Enhance the model's capability to faithfully support intense, asymmetric face expressions, setting a new state-of-the-art result in the emotion transfer task, surpassing previous methods in both metrics and quality.*

*Incorporate speech-driven mode to our model, achieving top-tier performance in audio-driven facial animation, making it possible to drive source identity through diverse modalities, including visual signal, audio, or a blend of both.*

*Furthermore, we propose a novel multi-view video dataset featuring a wide range of intense and asymmetric facial expressions, filling the gap with absence of such data in existing datasets.*

# 1. Introduction

## 1.1. Representation of facial expression

Advancements in neural head technologies now enable the creation of realistic avatars from a few or even one image, with the latter being crucial when only one source image is available. Cross-driving synthesis, where avatars are animated with different identities is a key technique for virtual reality, filmmaking, photo animation, etc. However, accurate transfer of facial expressions, especially those that are intense and uneven, remains a substantial challenge in avatar animation, particularly for cross-driving synthesis. While previous research focused mainly on preserving identity and transferring moderate facial motions, our work also seeks to accurately drive high-intensity and asymmetric expressions.

Our research is built on and extends the MegaPortraits [9] model, notable for state-of-the-art results in cross-driving synthesis. An additional advantage using this method is that, unlike many avatar systems that depend on limited [3, 25] predefined motion descriptors, MegaPortraits, learns expression representations from scratch, allowing for greater adaptability to a wider range of expressions. We investigate the latent expression spaces and training methods of MegaPortraits to enhance its ability to depict a broad spectrum of facial expressions. Our comprehensive analysis reveals that, while the original model shows limited effectiveness in intense motion representing, it has significant potential for improvement through targeted architectural modifications, adjustments in the training approach, and the integration of our novel dataset.

## 1.2. Integrating speech driving

We integrate speech in our model to be used either as a complementary or an alternative driver, a crucial aspect for applications like virtual assistance or mixed reality when a primary visual signal is absent.

By improving disentanglement of facial expression latent space, we remove disturbing information from face motion descriptors and emphasize components that solely control lip movements. Based on this, we formulated a novel loss that helps to achieve desirable results. Furthermore, our method can generate plausible head rotations and blinks, which appear natural and enhance its applicability across various tasks. Thus, our final method can drive source identity through image, audio, or a mix of them.

## 1.3. FEED: Facial Extreme Emotions Dataset

Public data scarcity is a common obstacle in deep learning, the topic of human facial expressions and movements is not an exception. In particular, we note a lack of high quality video data capturing a wide range of facial expressions, with existing datasets not going beyond basic facial actions shown at Fig. 2. To address this, we collect a new dataset that includes basic expressions and also captures complex movements like blinks, winks, and head and tongue movements, along with varied extreme expressions which are difficult or impossible to categorize through basic facial actions. Given this variety in expressions in our dataset, we believe that it will be a valuable resource for research in human emotions and facial reconstruction fields. In summary, our main contributions are as follows:

- We introduce our new **EMOPortraits** model for one-shot head avatars synthesis, that is capable of transferring intense facial expression, showing *state-of-the-art* results in cross-driving synthesis. We achieve this through careful latent facial expression space development as well as novel losses and a minimal amount of domain-specific data *($\sim$ 0.1 % of the train set)*.
- We integrate a speech-driving mode in our model, that demonstrates cutting-edge results in speech-driven animations. It functions effectively alongside visual signals or independently, also generating realistic head rotations and eye blinks.
- We present a unique multi-view dataset that spans a broad spectrum of extreme facial expressions, filling the gap of absence of such data in existing datasets.

# 2. Related Work

## 2.1. Visual-driven head avatars

In recent years, the domain of neural avatars has branched into two prominent sub-fields: person-specific and person-agnostic avatars. While person-specific methodologies [10, 11, 30, 38, 39] excel in delivering stunning realism and motion fidelity for a particular individual, they encounter challenges representing an arbitrary person. Moreover, these approach demands multiple frames of training data, distinct training for each avatar, and can struggle to replicate motions not encountered during training.

Person-agnostic methods, on the other hand, don't need train or fine-tuning for each new person and present an alternative approach to talking-head synthesis. Earlier works in this domain generated avatars in a few-shot technique [3, 33], while subsequent studies introduced one-shot capabilities [2, 8, 9, 22, 29, 31, 32, 34]. Many of these works employ predefined motion representations, such as 3DMM's blendshapes [8, 15, 31, 32]. In contrast, some learned latent emotion representation from scratch [3, 9, 22, 34, 35]. This latter approach holds potential for better motion representation as in such settings, expression descriptors become entirely trainable and free from inheriting the limitations typical for blendshapes as was shown in [3, 25]. In our work, we adopt this strategy, remarkably improving upon the methodology used in MegaPortraits[9].

**FEED dataset internal structure**

| Method | BFEs | Head rotations | Winks | Eyes move | Asym. em. | Tongue em. | Extreme em. |
|---|---|---|---|---|---|---|---|
| Mode | Mild/Strong | Axes/random | Mild/Strong | Random | Strong | Strong | Extreme |
| # of part. | 21/23 | 21 | 23 | 21 | 23 | 23 | 23 |
| Avg. lenght | 1:38/1:52 | 0:58 | 0:49 | 0:33 | 1:41 | 1:53 | 3:02 |

Table 1. The tasks we asked our participants to perform for our FEED dataset. Here BFEs stands for basic facial expressions, shown in Fig. 2

**Datasets comparison**

| Dataset | #Actors | #Views | Basic exp. | Strong exp. | Extrime & Assym. exp. | Tongue exp. | Resolution |
|---|---|---|---|---|---|---|---|
| SAVESS | 4 | 1 | ✓ | ✗ | ✗ | ✗ | 1280 × 1024 |
| RAVGESS | 24 | 1 | ✓ | ✗ | ✗ | ✗ | 1920 × 1080 |
| CREMA-D | **91** | 1 | ✓ | ✗ | ✗ | ✗ | 1280 × 720 |
| MMI | 25 | 1 | ✓ | ✓ | ✗ | ✗ | 1920 × 1080 |
| MEAD | 48 | **7** | ✓ | ✓ | ✗ | ✗ | 1920 × 1080 |
| Ours (FEED) | 23 | 3 | ✓ | ✓ | ✓ | ✓ | **3840 x 2160** |

Table 2. Comparisons with modern, high-quality audio-visual datasets created in controlled settings.

## 2.2. Face expression datasets

Early image datasets[1, 19] predominantly offered annotated face expression data corresponded to up to 8 basic emotions. However, these datasets are not well-suited for training head avatars, which necessitate video or multi-view data.

The SAVEE [14] dataset captures facial expressions in speech, but involves just 4 actors. The MMI [20] dataset, more expansive in actor participation, offers spontaneous moderate-intensity expression but is restricted to single-view sequences. RAVDESS [18] distinguishes itself by capturing two motion intensities, but its limited recordings challenge broad applicability. CREMA-D [4] encompasses three expression intensities, and though MEAD [28] provides detailed multi-view data across three intensities, it's limited to the standard eight expression groups. Our novel FEED dataset aims to address the scientific community's demand for high-quality multi-view facial expression videos outside the standard categories: Joy, Fear, Sadness, Disgust, Anger Contempt and Surprise.

## 2.3. Speech-driven head avatars

While numerous studies introduce speech-driven avatars [12, 21, 24, 31, 36, 39, 40], few excel in producing high-quality talking heads with authentic rotations, blinks, and the capability to use both visual and audio inputs. For instance, Wav2Lip [21] aims at re-dubbing videos with accurate lip motions, but often falls short in realism with using single images. Diffused heads [24] struggles to generate long sequences and doesnt provide access to pose control. MakeItTalk [40] animates facial landmarks in a speaker-specific manner, yet struggles with head pose controls due to its non-reliance on 3D. PC-AVS [39] offers a solution, but demands a driving video for pose modulation. Newly developed models such as SadTalker [36] and StyleHEAT [31] have shown promising outcomes in generating high-quality talking heads. We conducted a quantitative evaluation of our audio-driven mode compared to the aforementioned models.

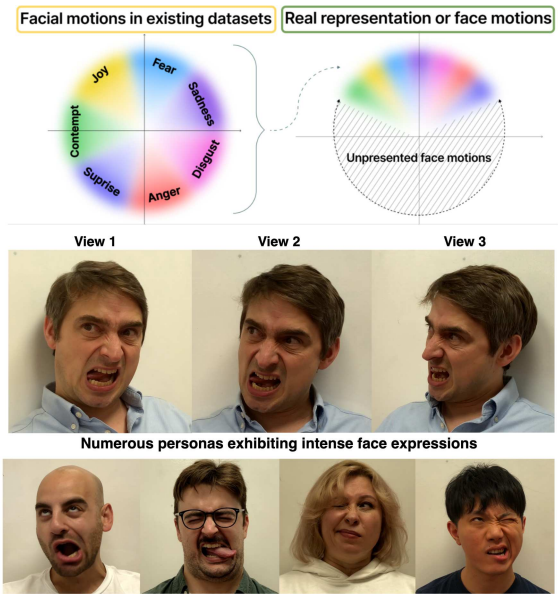


Figure 2. Illustration of the problem in publicly available face expression data and selected examples from our FEED dataset.

## 3. FEED dataset

As previously mentioned, the current publicly available datasets with human face expression videos fall short in capturing a broad range of facial manipulations (see illustration at Fig. 2). In our view, this gap could stem from the complexity of designing clear tasks for participants, coupled with the reality that not everyone is ready for performing extreme expressions on camera. To bridge this gap, our pioneering FEED dataset is designed to meet the scientific community's need for high-quality, multi-view emotion videos that extend beyond standard emotions. It contains various expressions, including strong asymmetric ones, tongue and cheeks movements, winks, head rotations, eye movements and more nuanced gestures. ;

Our dataset consists of 520 multi-view videos of 23 subjects, captured with 3 cameras. As extreme face motions are complex and their perception can be heavily influenced by subtle differences, we use a high resolution of 4k for all video, capturing the whole face up to the level of individual hair strands and wrinkles, as shown in Fig. 2. For more examples, please refer to the supplementary materials. Our participants were asked to perform 7 tasks (see Tab. 1) to cover as many facial and head movements as possible.

In our comparison with other expression datasets detailed in Tab. 2, the MEAD dataset [28] emerges as the closest to ours regarding viewpoints and maximal expression intensity. However, despite MEAD's larger number of actors and viewpoints, it offers significantly less variety of facial expressions and have lower resolution than our dataset.

## 4. Expression enhancement

We found that the MegaPortraits [9] model fails to transfer intense expressions correctly, as shown in Fig. 9. Using our FEED dataset of strong and uneven facial expression to fine-tune a pre-trained MegaPortraits model didn't enhance final results as shown at Tab. 4. Training from scratch on FEED leads to fast overfitting due to the small number of identities presented in the dataset compared to VoxCeleb2 [7] used in [9] (23 vs ∼5000). Our method effectively injects this small dataset into training, yielding desirable results. We list our key findings here, for all details, see supplementary materials. Our model's scheme, displayed in Fig. 4, shares similarities with [9], leading us to occasionally refer to certain elements of our scheme in the context of [9]'s pipeline.

### 4.1. Latent expression space

We begin by exploring MegaPortraits' latent expression descriptors' space ($\mathbf{z}_{s/d}$ in Fig. 4), which is crucial for expression transfer, through PCA analysis. Inspired by [16], which demonstrated that the area under the cumulative explained variance curve of singular values (denoted here as $\mathbf{AUC_z}$ ) can serve as an effective metric for dimensionality collapse in some latent spaces, we applied this measure to our study.

This metric allows us to forecast model performance: greater collapse suggests reduced entropy in the latent space, which correlates with lower representational quality. High quality of the expression space, produced by $\mathbf{E}_{motion}$ (as shown in Fig. 4), is crucial to capture and differentiate between the nuances of strong facial expressions. Our ablation study confirms (see Tab. 4) that there is a notable correlation between how broad and isotropic model's latent space is (expressed by $\mathbf{AUC_z}$) and its final performance. Visual comparison is available in supplementary materials. As shown in Fig. 3, our model's latent space outperforms MegaPortraits in expression representation ability. This is supported by both visual (Fig. 9) and quantitative analysis (Tab. 3).

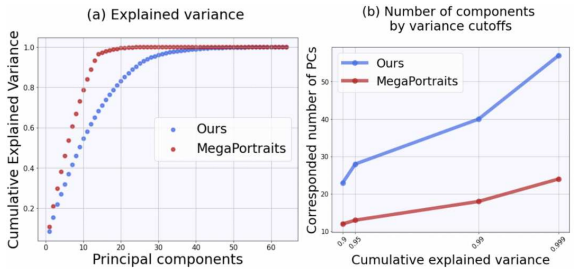


Figure 3. Comparison of latent spaces. Left plot shows that our model's latent space is wider and exhibits more even variance distribution. Also, as shown on the right plot, a greater number of principal components are involved in capturing variance across various thresholds. This implies a more robust representational capacity of expression space compared to [9]. The VoxCeleb2 test set was used for both plots.

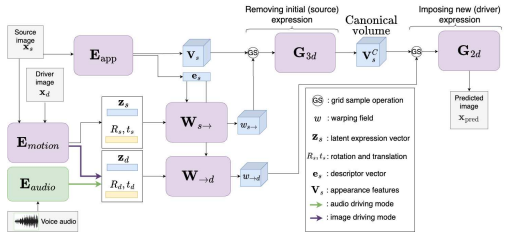


Figure 4. Method Overview. We use $\mathbf{E}_{\text{app}}$ to extract volume features $\mathbf{V}_s$ and a global descriptor $\mathbf{e}_s$ from the source image. Then $\mathbf{E}_{\text{motion}}$ or $\mathbf{E}_{\text{audio}}$ generates motion representations from source and driver, including head rotations $\mathbf{R}_{s/d}$, translations $\mathbf{t}_{s/d}$, and expression descriptors $\mathbf{z}_{s/d}$. Using them, we predict warpings $\mathbf{w}_{s\rightarrow}$ and $\mathbf{w}_{\rightarrow d}$. First warping and $\mathbf{G}_{\text{3D}}$ transform $\mathbf{V}_s$ into a canonical volume $\mathbf{V}_s^C$ by removing the source motions. Second warping and $\mathbf{G}_{\text{2D}}$ imposes the driver's motions and renders the final image.

Both metrics shown in Fig. 3 were vital in our model's development since they highly correlate with the final model's ability to transfer intense expressions. This was particularly useful for finding adjustments described in Sec. 4.2 and Sec. 4.3, as early visual results during training are not very telling, but changes in the latent space are noticeable after a few epochs. We also found that in MegaPortraits, just a few principal components significantly affect variance, with only 18 components making up 99% (Fig. 3), suggesting their 512 $\mathbf{z}$'s dimension is oversized. For our model, we found a 128-dimensional $\mathbf{z}$ is optimal, offering solid performance and reducing overfitting in imbalanced data, as our ablation study reveals. We calculate the Explained Variance ($\text{EV}_i$) for the $i^{th}$ component and ($\mathbf{AUC_z}$) using equations shown in Eq. (1), which involves standardizing the vector set $\mathcal{Z}$ for PCA and sorting the eigenvalues $\boldsymbol{\lambda}$ in descending order.

$$\begin{aligned} \mathbf{Z}_{\text{std}} &= \frac{\mathcal{Z} - \overline{\mathcal{Z}}}{\sigma(\mathcal{Z})}, \quad \boldsymbol{\lambda} = \text{sorted}_{\text{desc}}\left(\text{eig}\left(\frac{1}{n-1}\mathbf{Z}_{\text{std}}^{\top}\mathbf{Z}_{\text{std}}\right)\right), \\ \text{EV}_i &= \frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}, \quad \mathbf{AUC_z} = \frac{1}{d}\frac{\sum_{i=1}^{d}\sum_{k=1}^{i}\lambda_k}{\sum_{j=1}^{d}\lambda_j}, \end{aligned} \tag{1}$$

### 4.2. Canonical volume

Despite being designed to exclude face expression details, we found that canonical volume ($\mathbf{V}^C$ in Fig. 4) in MegaPortraits actually retains significant expression information from the source image, contributing to poor translation of intense expressions, as shown in our experiment detailed in Fig. 5. In this experiment, using portraits with varied expressions, we assessed the expression translation accuracy and how expression intensity affects the $\mathbf{V}^C$s. Our findings suggest that the canonical volumes in MegaPortraits are not truly neutral. As confirmed by our ablation study (Tab. 4), creating an expression-free $\mathbf{V}^C$ is essential for precise translation of intense motions. We believe that a more neutral canonical volume improves tractability and effectiveness in expression translation tasks.

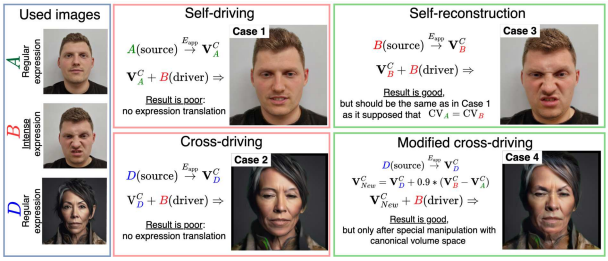


Figure 5. Canonical volumes ($\mathbf{V}^C$) in MegaPortraits [9] are not expression-neutral. To show it we use three portraits: A (regular expression), B (intense expression), D (regular expression, new identity). *Case 1* visualizes poor results in transferring B's intense expression to A using a self-driving mode, contrasting with the effective reconstruction of B when used as both driver and source in *case 3*. This discrepancy, indicative of expression leakage into the canonical volume, is further quantified by a 43% relative difference in $\mathbf{V}^C$s between B and C, contrary to expectations of their similarity for same identity. In cross-driving generation *case 2*, using B as the driver and D as the source yielded poor results. However, in *case 4*, after using additive operations on source canonical volumes, the output expression is much closer to driver than in *case 2*. This manipulation again confirms the significant retention of expression information in canonical volumes.

To overcome this issue, we propose to match canonical volumes from the different images of the same person $(\mathbf{V}^C_{s^n}, \mathbf{V}^C_{d^n})$ during training as following:

$$\mathcal{L}^{\text{n}}_{\text{CV}} = \mathcal{L}_{\text{MAE}}\left(\mathbf{V}^C_{s^n}, \mathbf{V}^C_{d^n}\right), \tag{2}$$

This loss ensures $\mathbf{V}^C$ remains stable and expression-independent, crucial for translating intense expressions, as shown in our ablation study (Tab. 4).

### 4.3. Source-driver mismatch loss

In addition to maintaining expression-free $\mathbf{V}^C$, it's vital to remove all identity information from the expression vector $\mathbf{z}$. While contrastive losses address this in [9], our experiments indicate they are insufficient to prevent overfitting in our imbalanced data scenario, where emotionally intense images represent only 0.1% of our training set but are sampled 25% of the time. We introduce a novel self-supervised loss which mitigates identity information in latent expression vectors:

$$\mathcal{L}_{sdm}(\mathbf{z}_s, \mathbf{z}_d) = w * \max(0, \cos(\mathbf{z}_s, \mathbf{z}_d) - \textit{margin}),$$
$$w; \textit{margin} = \begin{cases} 1; 0.5 & \text{if } \mathbf{z}_s, \mathbf{z}_d \text{ are from VC2} \\ 10; 0.25 & \text{if } \mathbf{z}_s, \mathbf{z}_d \text{ are from FEED} \end{cases} \tag{3}$$

Here, $\mathbf{z}_s, \mathbf{z}_d$ are emotion vectors from source and driver images, $w, \textit{margin}$ adjust the loss's intensity and strictness respectively. We increase $w$ for FEED due to its higher sampling rate and decrease *margin* for more assured expression variance in $s, d$. Refer to Fig. 6 for a visual summary of all self-supervised losses enhancing our latent emotion space.

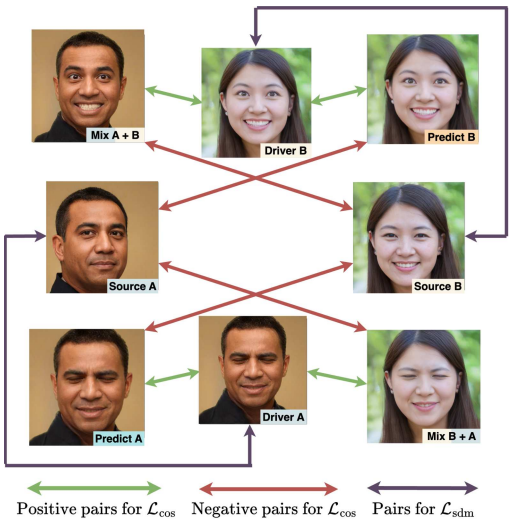


Figure 6. The visualisation of our self-supervised losses. $\mathcal{L}_{\text{cos}}$ is a contrastive loss, similar to one used in [9] but uses different pairs as described in Sec. 5.1. Our novel $\mathcal{L}_{sdm}$ is designed to prevent overfitting in an imbalanced dataset.

## 5. Incorporating speech-driving mode

### 5.1. Latent space disentanglement

During our latent space analysis described in Sec. 4, we found that combining expression from one driver with the head pose from another resulted in poor performance for MegaPortraits model, as illustrated in Fig. 7 (bottom row). Indeed, the base model lacks a mechanism to intentionally prevent head pose leakage. Disentangled expression latent space, besides expanding the model's use cases, plays the crucial role for the speech-driving mode. When the latent vector $\mathbf{z}$ is entangled with head pose data, predicting it from speech becomes challenging because speech lacks head rotation information, unlike images. Moreover, we would like to have a full control over head rotations during speech mode. We were able to make our latent space disentangled by changing the way we sample images for $\mathcal{L}_{\text{cos}}$ Eq. (5).

The MegaPortraits model uses $\mathcal{L}_{\text{cos}}$ - a modified *Large Margin Cosine Loss* [26] to mitigate the transfer of appearance features to expression descriptor (latent vector $\mathbf{z}$ on Fig. 4), which is crucial for cross-driving mode. Without this, differences in appearance between the source and driver, like hairstyle or skin tone, leak from driver to output image.

For computing this loss, the authors utilize a supplementary source-driving pair ($\mathbf{x}_{s^*}$ and $\mathbf{x}_{d^*}$) from a different video with another identity, ensuring a distinct appearance from the current $\mathbf{x}_s, \mathbf{x}_d$ pair. The base model is then employed to produce *cross-reenacted* image ($\hat{\mathbf{x}}_{s^* \rightarrow d} = \mathbf{G}_{\text{base}}(\mathbf{x}_{s^*}, \mathbf{x}_d)$). Concurrently, they determine a separate motion descriptor, $\mathbf{z}_{d^*} = \mathbf{E}_{\text{motion}}(\mathbf{x}_{d^*})$. Descriptors $\mathbf{z}_{* \rightarrow d}$ from the respective forward passes are also used for positive and negative pairs.


Figure 7. The illustration of disentanglement problem. Images in the right row indicate the expression descriptor $\mathbf{z}$ entangled with head pose for MegaPortraits, whereas we avoid it using changes proposed in Sec. 5.1.

Our innovation here is a new sampling strategy: for each $\mathbf{x}_s$, $\mathbf{x}_d$ pair, we sample one more random additional pair $\mathbf{x}_{s^m}$ and $\mathbf{x}_{d^m}$ apart from $\mathbf{x}_{s^*}$ and $\mathbf{x}_{d^*}$. We then apply the model to produce the following *cross-reenacted moved* image using source identity from $\mathbf{x}_{s^*}$, desired head pose from $\mathbf{x}_{s^m}$ and emotions from $\mathbf{x}_d$, represented as: $\hat{\mathbf{x}}_{s^{*m}\to d} = \mathbf{G}_{\text{base}}(\mathbf{x}_{s^{*m}}, \mathbf{x}_d)$. This creates a positive pair with the same desired emotion but varying head poses.

Motion descriptors are organized into *positive pairs* $\mathcal{P}$ for alignment, and *negative pairs* $\mathcal{N}$ for non-alignment: $\mathcal{P} = \big\{(\mathbf{z}_{s\to d}, \mathbf{z}_d), (\mathbf{z}_{s^*\to d}, \mathbf{z}_d), (\mathbf{z}_{s^{*m}\to d}, \mathbf{z}_d)\big\}$, and $\mathcal{N} = \big\{(\mathbf{z}_{s\to d}, \mathbf{z}_{d^*}), (\mathbf{z}_{s^*\to d}, \mathbf{z}_{d^*}), (\mathbf{z}_{s^{*m}\to d}, \mathbf{z}_{d^*})\big\}$ (see Fig. 6). These pairs are used to calculate the following cosine distance:

$$d(\mathbf{z}_i, \mathbf{z}_j) = s \cdot \big(\langle \mathbf{z}_i, \mathbf{z}_j \rangle - m\big), \tag{4}$$

where both $s$ and $m$ are hyperparameters. This distance is then used to calculate a large margin cosine loss (CosFace) [27]:

$$\mathcal{L}_{\text{cos}} = -\sum_{(\mathbf{z}_k, \mathbf{z}_l)\in\mathcal{P}} \log \frac{\exp\big\{d(\mathbf{z}_k, \mathbf{z}_l)\big\}}{\exp\big\{d(\mathbf{z}_k, \mathbf{z}_l)\big\} + \sum\limits_{(\mathbf{z}_i, \mathbf{z}_j)\in\mathcal{N}} \exp\big\{d(\mathbf{z}_i, \mathbf{z}_j)\big\}} \tag{5}$$

This loss prevents the head pose from leaking into the embeddings as shown in Fig. 7 (upper row)

By doing this, we achieved the capability to interpret $\mathbf{z}_i$ as a latent vector encapsulating the emotional content from an image $\mathbf{x}_i$. This advancement opened the door for an intriguing possibility during inference: the prediction of this vector could potentially be derived from an audio signal instead of a facial image using the motion encoder $\mathbf{E}_{\text{motion}}$, effectively transforming the model into a speech-driven system. The critical challenge here is aligning the audio input with the pretrained latent space. We addressed this by introducing an additional audio encoder, $\mathbf{E}_{\text{aud}}$, and employing $\mathbf{E}_{\text{motion}}$ as a teacher model for this new encoder. A crucial factor in achieving remarkable accuracy in lip synchronization was our ability to isolate specific components within the expression latent space, that are solely responsible for mouth movements. We describe it in the next subsection Sec. 5.2.

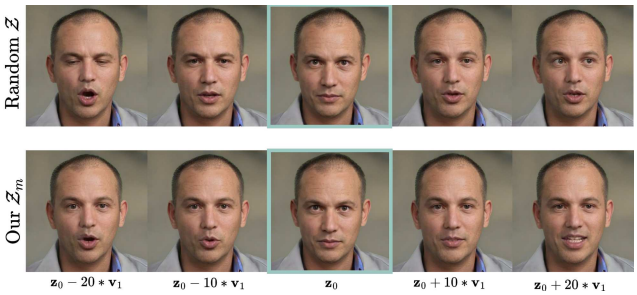


Figure 8. In the upper row, the first principal component from a random subset of expression vectors $\mathcal{Z}$ affects not just the mouth but also blinks and gaze, making it difficult to interpret and isolate mouth movement components, needed for training $\mathbf{E}_{\text{aud}}$. However, as shown in the bottom row, using $\mathcal{Z}_m$ from Sec. 5.2, the first principal component, $\mathbf{v}_1$, focused on mouth movements, is effectively isolated, $\mathbf{z}_0$ - latent vector, corresponded to the central image.

## 5.2. Mouth movements

To train the audio encoder $\mathbf{E}_{\text{aud}}$ for predicting $\mathbf{z}$ from speech, we chose to use the motion encoder $\mathbf{E}_{\text{motion}}$ as a teacher model. However, direct matching vectors $\mathbf{z}_i^{\text{aud}}$ from $\mathbf{E}_{\text{aud}}$ and pseudo-ground truth $\mathbf{z}_i$ from $\mathbf{E}_{\text{motion}}$ gives poor results (see Tab. 6). We believe that the reason behind this is that speech-derived $\mathbf{z}_i^{\text{aud}}$, while able to capture lip movements, struggles with other facial motions, especially in the upper face. Thus, we focused on isolating mouth movement components in the expression latent space using PCA analysis.

Employing PCA on a broad array of expression vectors $\mathcal{Z}$, generated from images with varied facial emotions, yields principal components with their explained variances ($\text{EV}_i$). Altering these components, particularly those with high $\text{EV}_i$, significantly changes facial expressions in an image. However, they typically affect a combination of facial features-Fig. 8 (upper row), not isolated areas like the mouth or eyes.

To focus on mouth movements, we created a unique set of $\mathcal{Z}_m$ from just one video of a person performing mouth manipulations and manually edited it to minimize upper face movements. This involved using a still upper face from the first frame in all subsequent frames, ensuring that the principal components mainly represented mouth movements. Applying PCA of $\mathcal{Z}_m$ reveals that most principal components are responsible for solely mouth movements. For illustration of the first component, see Fig. 8 (bottom row). Based on distilled components, we introduce mouth PCA mouth loss:

$$\mathcal{L}_{\text{PCA}}(v_i, v_j, n) = \frac{1}{n}\sum_{k=1}^{n} \big|\text{PC}_{\mathbf{v}_i}(k) - \text{PC}_{\mathbf{v}_j}(k)\big| \tag{6}$$

Where $n$ is the number of principal components to be considered, $\text{PC}_{\mathbf{z}_i}(k)$ and $\text{PC}_{\mathbf{z}_j}(k)$ are the $k$-th principal components of vectors $\mathbf{z}_i$ and $\mathbf{z}_j$ respectively. Implementing this loss has demonstrably enhanced the quality and accuracy of our results, as our ablation study confirms (see Tab. 6).

## 6. Experiments

In this section, we describe experiments we set for comparison for both our image-driven and speech-driven mode. Please see implementation details such as all used losses, component's sizes and architecture details, training procedure, data preprocess in supplementary materials.

### 6.1. Image-driven comparison

**Methods**. We evaluated our image-driven model alongside various one-shot models. For this comparison, we included models like NOFA [32] and StyleHEAT [31], which use 3DMM expression blendshapes without disentanglement training, and models with trainable latent face motion representations like FOMM [22], UVA [23], MetaPortrait [35] and MegaPortraits [9]. It's noteworthy that NOFA and UVA require per-source optimization, unlike the others, including our model.

**Evaluation metrics and data**. For our evaluation, we use 100 random, non-child, single-person images from the FFHQ dataset as sources. From the MEAD dataset, we chose 100 images covering all high-intensity emotions except neutral, and another 100 from the FEED dataset's extreme emotion task, featuring 5 identities not seen during training of our model. Each source image was paired with 20 driving images from these sets, totaling 2000 pairs.

For assessing reenactment quality, we employ various metrics including the Frechet Inception Distance (FID) [13] to measure the distributional discrepancy between synthetic and real images. Cosine similarity (CSIM) from a face recognition network [5] quantifies the identity preservation in generated images. Additionally, we conducted user studies to determine preferences regarding motion (UMTN) and appearance (UAPP) preservation, presenting participants with triplets—either a driving or source image and two method outputs—and asking them to choose the one with better motion or appearance preservation. Our user study includes 34 participants and around 300 unique questions. Both metrics numerically reflect preference percentages. As our visual results Fig. 9, and quantitative metrics Tab. 3 show, our method demonstrates exceptional ability in translating strong and extreme expressions, notably surpassing other image-driven avatar techniques in user preference for facial expression translation (UMTN) and FID scores.

### 6.2. Speech-driven comparison

**Methods**. To assess our model's speech-driven capabilities, we compared it with StyleHEAT [31], which also operates using both image and audio inputs, and audio-exclusive models including SadTalker [36], MakeItTalk [40], and PC-AVS [39].

**Evaluation metrics and data**. We employ the evaluation procedure described in [12]. We measure video realism using Frechet Inception Distance (FID) [13]. We also extract

| Driver: | MEAD | | | | FEED | | | |
|---|---|---|---|---|---|---|---|---|
| Method | FID↓ | CSIM↑ | UMTN↑ | UAPP↑ | FID↓ | CSIM↑ | UMTN↑ | UAPP↑ |
| MetaPortrait [35] | 77.5 | 0.64 | * | * | 78.8 | 0.66 | * | * |
| NOFA [32] | 69.2 | 0.65 | * | * | 69.8 | 0.66 | 3.8 | 8.8 |
| FOMM [22] | 84.4 | 0.56 | 2.4 | 0.3 | 85.9 | 0.55 | 1.8 | 0.5 |
| UVA [17] | 78.7 | 0.68 | 0.9 | 6.4 | 79.5 | 0.68 | 0.6 | 9.4 |
| StyleHEAT [31] | 71.9 | 0.63 | 3.7 | 3.3 | 72.4 | 0.61 | 2.6 | 3.8 |
| MegaPortraits [9] | 61.1 | 0.73 | 22.7 | **46.6** | 62.8 | **0.73** | 16.7 | **42.8** |
| Our (EMOPortraits) | **59.6** | **0.74** | **70.3** | 43.4 | **60.2** | 0.70 | **74.6** | 34.7 |

Table 3. Our method notably outperforms others in the FID score and strongly leads in the user preference metric for face expression translation (UMTN). It excels in reliably translating strong and extreme expressions from the driving image, distinguishing it as the most capable among the compared methods. Additionally, our approach maintains the source image's shape and appearance (measured by CSIM and UAPP metrics) on par with MegaPortraits [9] when using strong and regular expressions from the MEAD dataset as the driving signal. However, we observe a minor deviation in identity preservation metrics for extreme emotions (using the FEED dataset). This difference may arise because our method attempts to transfer asymmetric facial expressions, while MegaPortraits might modify the source image less, or not at all, with a challenging driver (as depicted in Fig. 5, case 1), thereby achieving better identity preservation. UVA was excluded from the MEAD-driven user study due to the authors were able to provide only part of the test data. Also, MetaPortrait was excluded from both user studies because of late data provision by authors, offering only the basic model results, without their super-resolution module outputs.

| Method | FID↓ | CSIM↑ | UMTN↓ | UAPP↓ | $\text{AUC}_{\mathbf{z}}$↓ |
|---|---|---|---|---|---|
| MegaPortraits fine-tuned | 62.3 | 0.63 | 8.5 | 12.4 | 0.89 |
| Ours w/o $\mathcal{L}^n_{CV}$ | 61.4 | 0.67 | 9.2 | 17.9 | 0.88 |
| Ours w/o $\mathcal{L}_{sdm}$ | 61.8 | 0.59 | 11.3 | 6.7 | 0.85 |
| Ours $dim(\mathbf{z}) = 512$ | 63.4 | 0.44 | 25.6 | 1.1 | **0.77** |
| Ours | **59.6** | **0.74** | **45.4** | **61.9** | 0.80 |

Table 4. Our ablation study examines key aspects of the image-driven mode. After fine-tuning MegaPortraits on the FEED dataset, we observed a decline in identity preservation due to overfitting on FEED identities, without any improvement in emotion translation. The lack of $\mathcal{L}^n_{CV}$ primarily impacts the ability to translate extreme expression, as indicated by UMTN. Additionally, not using $\mathcal{L}_{sdm}$ or setting $dim(\mathbf{z}) = 512$ results in identity leakage, particularly evident in the latter scenario. For this ablation, we use the part of our driven by MEAD dataset and described in Sec. 6.1

facial landmarks from outputs of all methods, frontalized, and normalized to position the mouth edges at (-1, 0) and (1, 0). We compute mean absolute error (MAE) for both predicted mouth landmark positions ($M_P$) and velocities ($M_V$), along with errors in facial geometry ($F_P$) and velocity ($F_V$). For our evaluation, we use 100 randomly selected video sequences from the HDTF dataset [37], each up to 30 seconds. As the results show in Tab. 5, our speech-driven model performs on par with leading methods, excelling in realism and facial dynamics. SyncNet [6] is used to assess lip sync quality, providing offset (temporal misalignment between the audio and video) and confidence (certainty of audio-visual alignment) scores.

**Source** **Driver** **FOMM** [22] **UVA** [17] **StyleHEAT** [31] **MetaPortrait** [35] **NOFA** [32] **MegaPortraits** [9] **Ours**

Figure 9. A qualitative comparison of head avatar systems in cross-reenactment scenario.

| Method | FID↓ | $F_P$↓ | $F_V$↓ | $M_P$↓ | $M_V$↓ | Sync. offset ↓ | Sync. conf. ↑ |
|---|---|---|---|---|---|---|---|
| StyleHEAT [31] | 62.4 | 1.34 | 0.41 | 2.98 | 1.67 | 0.41 | 7.22 |
| PC-AVS [39] | 71.5 | 2.78 | 0.51 | 2.56 | **1.04** | 3.79 | **8.42** |
| SadTalker [36] | 55.6 | 0.91 | 0.35 | **2.12** | 1.17 | 0.27 | 7.58 |
| MakeIrTalk [40] | 63.3 | 2.15 | 0.38 | 3.11 | 1.59 | 2.98 | 5.17 |
| Ours (EMOPortraits) | **28.5** | **0.82** | **0.33** | 2.53 | 1.38 | **0.07** | 5.77 |

Table 5. Our model matches other leading speech-driven models, excelling in realism (FID), facial geometry and velocity, but is slightly less accurate in mouth landmark positions and velocities. We also score best in audio-video sync offset, but lower in confidence. We believe that while SyncNet metrics are still helpful, they may not fully represent true synchronization quality, as generated videos can sometimes score higher than ground truth without necessarily being better synchronized. In our experimentation, ground truth videos scored an offset of 1.13 (which is worse than our method, SadTalker, and StyleHEAT) and a confidence score of 8.35 (less than PC-AVS). For qualitative analysis, please refer to our supplementary materials.

| Method | FID↓ | $M_P$↓ | $M_V$↓ | $F_P$↓ | $F_V$↓ |
|---|---|---|---|---|---|
| Ours w/o disentanglement | 44.7 | 0.98 | 0.45 | 5.78 | 2.55 |
| Ours w/o $\mathcal{L}_{\text{PCA}}$ | 29.3 | 0.84 | 0.35 | 3.13 | 1.88 |
| Ours with MAE$(\mathbf{z}_i, \hat{\mathbf{z}}_i^{\text{speech}})$ | **28.2** | 0.84 | 0.34 | 2.61 | 1.45 |
| Ours | 28.5 | **0.82** | **0.33** | **2.53** | **1.38** |

Table 6. Ablation study of our speech-driven mode, where in the first line, we show that it notably underperforms without head pose - face expression disentanglement of the latent space (see Sec. 5.1). Removing $\mathcal{L}_{\text{PCA}}$ leads to a significant reduction in generating realistic mouth movements. Replacing $\mathcal{L}_{\text{PCA}}$ with MAE$_(\mathbf{z}_i, \hat{\mathbf{z}}_i^{\text{speech}})$ causes a noticeable, decline in mouth movement generation. This suggests that identifying mouth-related PCA components was beneficial in enhancing the outcome.

### 6.3. Ablation study

We conducted an extensive ablation study to evaluate the contributions of individual components within our method. For our main model, we present the evaluation of the importance of the proposed source-driver unmatch loss, canonical volume loss, and change of $dim(\mathbf{z})$ from 512 in original model to 128 Tab. 4. For our speech-driven mode, we show how vital the disentanglement of the latent space for speech-drive ability, how $\mathcal{L}_{\text{PCA}}$ is better than naive matching of PCA components from random $\mathcal{Z}$ and how it helps match speech and latent expression space Tab. 6. For more ablation analysis, plots and visual comparison, please refer to our supplementary.

## 7. Conclusion

In this paper, we introduce EmoPortraits, a novel method for creating neural avatars with superior performance in image-driven, cross-identity emotion translation. Our speech-driven mode makes it possible to drive the facial animation through multiple conditions (video, audio, head motion). We collected FEED dataset which, we believe, will be a valuable asset for researchers in diverse human-centered studies.

However, our method has some limitations. It doesn't generate the avatar's body or shoulders, limiting some use cases. We currently integrate our output with a source image body. Additionally, the model sometimes struggles with accurate expression translation and underperforms with extensive head rotation. These challenges are crucial for future enhancements and remain central to our ongoing research efforts.

## References


[1] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pages 279–283, 2016. 3

[2] Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, and Georgios Tzimiropoulos. Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces, 2023. 2

[3] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. 2

[4] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014. 3

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2018. 7

[6] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 7

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 4

[8] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing, 2021. 2

[9] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars, 2023. 2, 4, 5, 7, 8

[10] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction, 2020. 2

[11] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos, 2022. 2

[12] Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable expression, 2022. 3, 7

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 7

[14] Philip Jackson and SJUoSG Haq. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*, 2014. 3

[15] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars, 2022. 2

[16] Alexander C. Li, Alexei A. Efros, and Deepak Pathak. Understanding collapse in non-contrastive siamese representation learning, 2022. 4

[17] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar, 2023. 7, 8

[18] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 3

[19] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 ieee computer society conference on computer vision and pattern recognition-workshops*, pages 94–101. IEEE, 2010. 3

[20] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pages 5–pp. IEEE, 2005. 3

[21] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020. 3

[22] Aliaksandr Siarohin, Stéphane Lathuilière, S. Tulyakov, Elisa Ricci, and N. Sebe. First order motion model for image animation. *ArXiv*, abs/2003.00196, 2019. 2, 7, 8

[23] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Hsin-Ying Lee, Jian Ren, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. *arXiv preprint arXiv:2301.11326*, 2023. 7

[24] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation, 2023. 3

[25] Mani Kumar Tellamekala, Ömer Sümer, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. Are 3d face shapes expressive enough for recognising continuous emotions and action unit intensities?, 2023. 2

[26] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition, 2018. 5

[27] H. Wang, Yitong Wang, Zheng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jin Zhou, and Wenyu Liu. Cosface: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 6

[28] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 3

[29] 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 Conference on Computer Vision and Pattern Recognition*, 2021. 2

[30] Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. AvatarMAV: Fast 3d head avatar reconstruction using motion-aware neural voxels. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*. ACM, 2023. 2

[31] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan, 2022. 2, 3, 7, 8

[32] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, and Baoyuan Wu. Nofa: Nerf-based one-shot facial avatar reconstruction, 2023. 2, 7, 8

[33] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models, 2019. 2

[34] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars, 2020. 2

[35] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, Hsiang-Tao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation, 2023. 2, 7, 8

[36] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation, 2023. 3, 7, 8

[37] 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. 7

[38] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I m avatar: Implicit morphable head avatars from videos, 2022. 2

[39] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation, 2021. 2, 3, 7, 8

[40] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. MakeltTalk. *ACM Transactions on Graphics*, 39(6):1–15, 2020. 3, 7, 8