DINet: Deformation Inpainting Network for Realistic Face Visually Dubbing on High Resolution Video

Zhimeng Zhang,[1*] **Zhipeng Hu,**[1 3*] **Wenjin Deng,**[2*] **Changjie Fan,**[1] **Tangjie Lv,**[1] **Yu Ding**[1 3*]

[1] Virtual Human Group, Netease Fuxi AI Lab
[2] Xiamen University
[3] Zhejiang University
{zhangzhimeng, zphu, fanchangjie, hzlvtangjie, dingyu01}@corp.netease.com
dengwenjin@stu.xmu.edu.cn

Abstract

For few-shot learning, it is still a critical challenge to realize photo-realistic face visually dubbing on high-resolution videos. Previous works fail to generate high-fidelity dubbing results. To address the above problem, this paper proposes a Deformation Inpainting Network (DINet) for high-resolution face visually dubbing. Different from previous works relying on multiple up-sample layers to directly generate pixels from latent embeddings, DINet performs spatial deformation on feature maps of reference images to better preserve high-frequency textural details. Specifically, DINet consists of one deformation part and one inpainting part. In the first part, five reference facial images adaptively perform spatial deformation to create deformed feature maps encoding mouth shapes at each frame, in order to align with the input driving audio and also the head poses of the input source images. In the second part, to produce face visually dubbing, a feature decoder is responsible for adaptively incorporating mouth movements from the deformed feature maps and other attributes (i.e., head pose and upper facial expression) from the source feature maps together. Finally, DINet achieves face visually dubbing with rich textural details. We conduct qualitative and quantitative comparisons to validate our DINet on high-resolution videos. The experimental results show that our method outperforms state-of-the-art works.

Introduction

Talking head generation tasks, including one-shot talking face (Chung, Jamaludin, and Zisserman 2017; Chen et al. 2018; Zhou et al. 2019; Vougioukas, Petridis, and Pantic 2020; Chen et al. 2019; Song et al. 2018; Das et al. 2020; Chen et al. 2020; Zhou et al. 2020, 2021; Zhang et al. 2021b; Wang et al. 2021, 2022a), person-specific talking face (Lahiri et al. 2021; Guo et al. 2021; Zhang et al. 2021a; Wu et al. 2021; Fried et al. 2019; Song et al. 2022; Thies et al. 2020; Ji et al. 2021) and few-shot face visually dubbing (KR et al. 2019; Prajwal et al. 2020; Xie et al. 2021; Park et al. 2022), have attracted growing research attention due to broad applications in media production, film industry, etc.

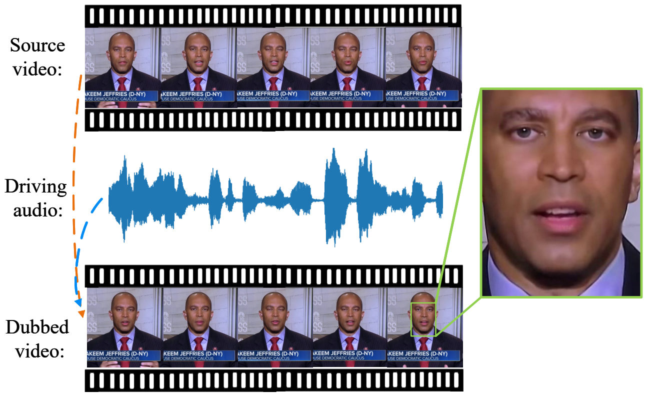

Figure 1: Our method dubs a high-resolution source video according to a driving audio.

Face visually dubbing, as shown in Figure 1, aims to synchronize mouth shape in a source video according to an input driving audio, while keeping identity and head pose in consistent with the source video frame. Existing few-shot face visually dubbing works (KR et al. 2019; Prajwal et al. 2020; Xie et al. 2021; Park et al. 2022; Liu et al. 2022) make great efforts on synthesizing realistic faces. They utilize convolutional networks with multiple up-sample layers to directly generate mouth pixels from latent embeddings. For example, Wav2Lip (Prajwal et al. 2020) designs face decoder with six de-convolutional layers. TRVD (Xie et al. 2021) combines up-sample layers with AdaIN (Huang and Belongie 2017) in face decoder. However, these straightforward methods always face trouble on producing high-resolution videos, despite using high-resolution training data (see the results of Wav2Lip (Prajwal et al. 2020) in Figure 5)

One main reason is that, under a few-shot condition, mouth textural details have few correlations with driving audios, making it challenging to generate high-frequency textural details directly. For networks, it is difficult to learn a complex mapping of audio-lip synchronization from amount of identities with various speaking styles, thus they easily ignore correlated textural details. Furthermore, compared with person-specific dubbing works (Fried et al. 2019; Thies et al. 2020), this reason makes few-shot works always generate

[*]Equal contribution. Yu Ding is the corresponding author.

more obvious blur results, when the person-specific dubbing works also rely on straightforward generation methods.

To solve the above problem, we develop spatial deformation on feature maps of reference facial images to inpaint mouth pixels. Specifically, our spatial deformation is able to synchronize a mouth shape with a driving audio and align a head pose with a source face. Deformation operation moves pixels into appropriate locations rather than generation from scratch, thus it nearly preserves all textural details.

In this paper, we propose a Deformation Inpainting Network (DINet) for realistic face visually dubbing on high-resolution videos. The framework is shown in Figure 2. DINet consists of two parts: a deformation part and an inpainting part. The deformation part first encodes the features of head pose and speech content from the source face and the driving audio respectively, and then utilizes these features to deform the reference faces. The inpainting part merges features of source face and deformed results by convolutional layers to inpaint the pixels in source mouth region. With the combination of deformation and inpainting, our DINet achieves more realistic face visually dubbing than direct-generation based methods.

Our contributions are summarized as follows:

- We develop and validate a novel Deformation Inpainting Network (DINet) to achieve face visually dubbing on high-resolution videos. Our DINet is able to produce accurate mouth movements but also preserve textual details.
- We conduct qualitative and quantitative experiments to evaluate our DINet, and experimental results show that our method outperforms state-of-the-art works on high-resolution face visually dubbing.

Related Work

Talking Face Generation

Talking face generation aims to synthesize facial images according to a driving audio or text. It consists of three main directions: one-shot talking face, few-shot face visually dubbing and person-specific talking face.

One-shot talking face. One-shot talking face focus on driving one reference facial image with synchronic lip movements, realistic facial expressions and rhythmic head motions. Some works utilize latent embeddings to generate talking face. They first encode a reference image and a driving audio into latent embeddings, and then use networks to decode the embeddings into a synthetic image. Extra losses, like deblurring loss (Chung, Jamaludin, and Zisserman 2017), audio-visual correlation loss (Chen et al. 2018), audio-visual disentangled loss (Zhou et al. 2019, 2021; Liang et al. 2022) and spatial-temporal adversarial loss (Song et al. 2018; Vougioukas, Petridis, and Pantic 2020) are used to improve lip synchronization and visual quality.

Other works leverage explicit intermediate representations, including unsupervised keypoints (Wang et al. 2021, 2022a; Ji et al. 2022), facial landmarks (Chen et al. 2019; Das et al. 2020; Zhou et al. 2020) and 3DMM (Chen et al. 2020; Zhang et al. 2021b) to synthesize facial images. They split a pipeline into two separated parts: one audio-to-animation part and one animation-to-video part. Two parts are trained independently to alleviate pressures of networks, thus they generate more realistic results.

Few-shot face visually dubbing. Few-shot face visually dubbing focus on repairing mouth region in source face according to driving audio. Existing works (KR et al. 2019; Prajwal et al. 2020; Xie et al. 2021; Park et al. 2022; Liu et al. 2022) use a similar face decoder, multiple up-sample layers, to directly generate pixels of mouth region in source face. To improve visual quality, (Xie et al. 2021) utilize facial landmarks as intermediate structural representations. (Liu et al. 2022) use effective SSIM loss. To improve lip-synchronization, (Prajwal et al. 2020) add one sync loss supervised by a pre-trained syncnet. (Park et al. 2022) use one audio-lip memory to accurately retrieve synchronic lip shape. However, these direct-generation based methods fail to realize face visually dubbing on high resolution videos. Our method replaces direct-generation fashion with deformation to achieve more realistic results.

Person-specific talking face. Person-specific talking face requires identity appears in training data. Early work (Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017) needs 16 hours of footage to learn audio-lip mapping. (Fried et al. 2019) reduce the training data to less than 1 hour by rule-based selection. (Song et al. 2022; Thies et al. 2020; Guo et al. 2021) utilize shared animation generators or Nerf (Mildenhall et al. 2020) to make their methods only require several minutes footage. (Lahiri et al. 2021) use lighting normalization to handle the extreme condition of changeable light. (Ji et al. 2021; Zhang et al. 2021a) pay attention to realistic facial expressions, e.g., emotion and eye blink.

Spatial Deformation

In deep learning based works, there are two main ways to realize spatial deformation: affine transformation and dense flow. In affine transformation based works, they first compute coefficients of affine transformations, then they do deformation on image feature maps by combining all affine transformations. (Jaderberg et al. 2015; Lin and Lucey 2017) first import affine transformation in CNN networks. They compute one global affine transformation in each feature layer. To improve the complexity of deformation, (Siarohin et al. 2019; Wang, Mallya, and Liu 2021) increase the number of affine transformations by computing different coefficients in different 2D or 3D regions. (Zhang and Ding 2022) propose one AdaAT operator to simulate a more complex deformation. They compute channel specific coefficients, increasing the number of affine transformations to hundreds.

In dense flow based works, they directly use networks to compute a complete dense flow, then they warp feature maps with dense flow to achieve spatial deformation. The dense flow can be computed from graph convolutional networks (Yao et al. 2020), encoder-decoder networks (Ren et al. 2021; Doukas, Zafeiriou, and Sharmanska 2021) and weight demodulation decoder (Wang et al. 2022b). In our work, we utilize AdaAT operator to realize spatial deformation because it synthesizes the best results.

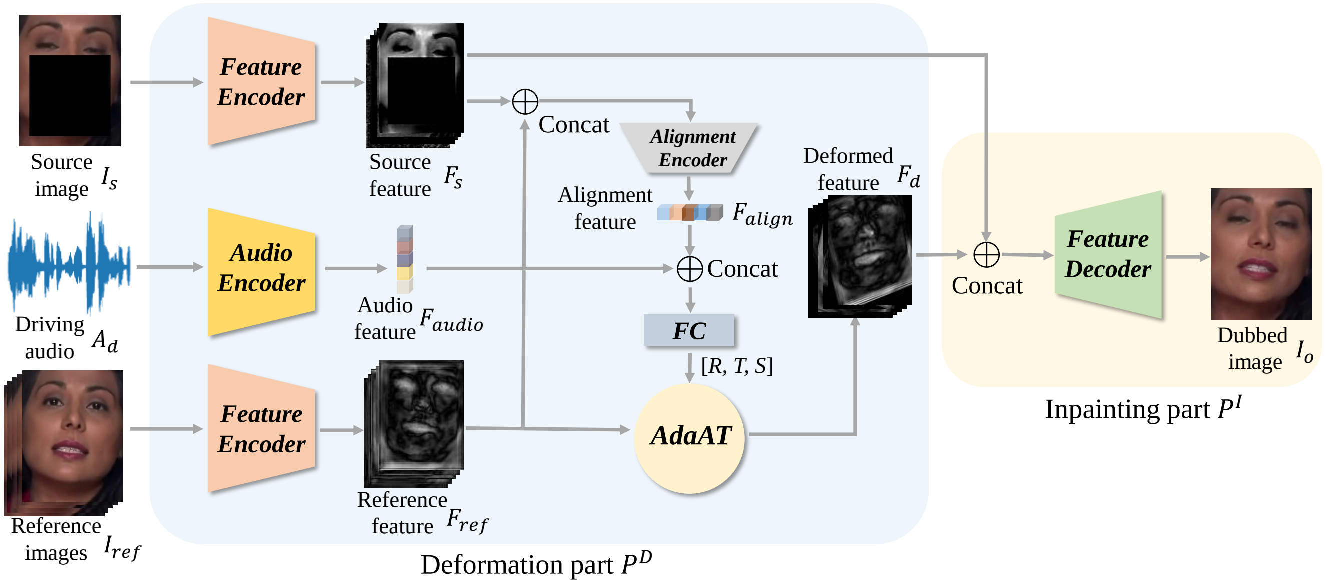

Figure 2: Illustration of our DINet framework. DINet consists of one deformation part P^D and one inpainting part P^I. In P^D, the feature maps of reference images are deformed spatially to synchronize mouth shape with the driving audio and align head pose with the source image. In P^I, deformed feature maps are used to inpaint the mouth region in the source face.

Method

We propose one DINet to achieve high resolution face visually dubbing. The structural details of DINet are shown in Figure 2. DINet consists of one deformation part P^D and one inpainting part P^I. The deformation part P^D focus on deforming the feature maps of the reference images spatially to synchronize the mouth shape with the driving audio and align the head pose with the source image. The inpainting part P^I focus on utilizing the deformed results to repair the mouth region in the source face. We detail these two parts in the following.

Deformation Part

The blue rectangle in Figure 2 illustrates the structure of P^D. Given one source image $I_s \in R^{3\times H\times W}$, one driving audio $A_d \in R^{T\times 29}$ (29 is a dimension of audio feature, we use deepspeech feature (Hannun et al. 2014) in our method) and five reference images $I_{ref} \in R^{15\times H\times W}$, P^D aims to produce deformed features $F_d \in R^{256\times\frac{H}{4}\times\frac{W}{4}}$ that have synchronous mouth shape with A_d and have aligned head pose with I_s. To realize this purpose, A_d is first input into one audio encoder to extract audio feature $F_{audio} \in R^{128}$. F_{audio} encodes the speech content of A_d. Then, I_s and I_{ref} are input into two different feature encoders to extract source feature $F_s \in R^{256\times\frac{H}{4}\times\frac{W}{4}}$ and reference feature $F_{ref} \in R^{256\times\frac{H}{4}\times\frac{W}{4}}$. Next, F_s and F_{ref} are concatenated and input into one alignment encoder to compute alignment feature $F_{align} \in R^{128}$. F_{align} encodes the aligned information of head pose between I_s and I_{ref}. Finally, F_{audio} and F_{align} are used to spatially deform F_{ref} into F_d.

In our method, we borrow the AdaAT operator (Zhang and Ding 2022) instead of dense flow to realize spatial deformation. The main reason is that, compared with dense flow, AdaAT can deform feature maps with misaligned spatial layouts by doing feature channel specific deformations. AdaAT operator computes different affine coefficients in different feature channels. In our P^D, fully-connected layers are used to compute coefficients of rotation $R = \{\theta^c\}_{c=1}^{256}$, translation $T_x = \{t_x^c\}_{c=1}^{256}$/ $T_y = \{t_y^c\}_{c=1}^{256}$ and scale $S = \{s^c\}_{c=1}^{256}$. Then, these affine coefficients are used to do affine transformations on F_{ref}, as written in

$$\begin{bmatrix}\hat{x}_c\\ \hat{y}_c\end{bmatrix} = \begin{bmatrix}s^c cos(\theta^c) & s^c(-sin(\theta^c)) & t_x^c\\ s^c sin(\theta^c) & s^c cos(\theta^c) & t_y^c\end{bmatrix}\begin{bmatrix}x_c\\ y_c\\ 1\end{bmatrix}, \quad (1)$$

where x_c/y_c and $\hat{x}_c$/$\hat{y}_c$ denote the pixel coordinates before and after affine transformation respectively. $c \in [1, 256]$ represents c_{th} channel in F_{ref}. After AdaAT operator, F_{ref} is deformed into F_d.

Inpainting Part

The yellow rectangle in Figure 2 illustrates the structure of P^I. P^I aims to produce dubbed image $I_o \in R^{3\times H\times W}$ from F_s and F_d. To realize this purpose, F_s and F_d are first concatenated in feature channel. Then, one feature decoder with convolutional layers is used to inpaint the masked mouth and generate I_o. More structural details are in supplementary materials.

Loss Function

In training stage, we use three kind of loss functions to train DINet, including perception loss (Johnson, Alahi, and Fei-Fei 2016), GAN loss (Mao et al. 2017) and lip-sync loss (Prajwal et al. 2020).

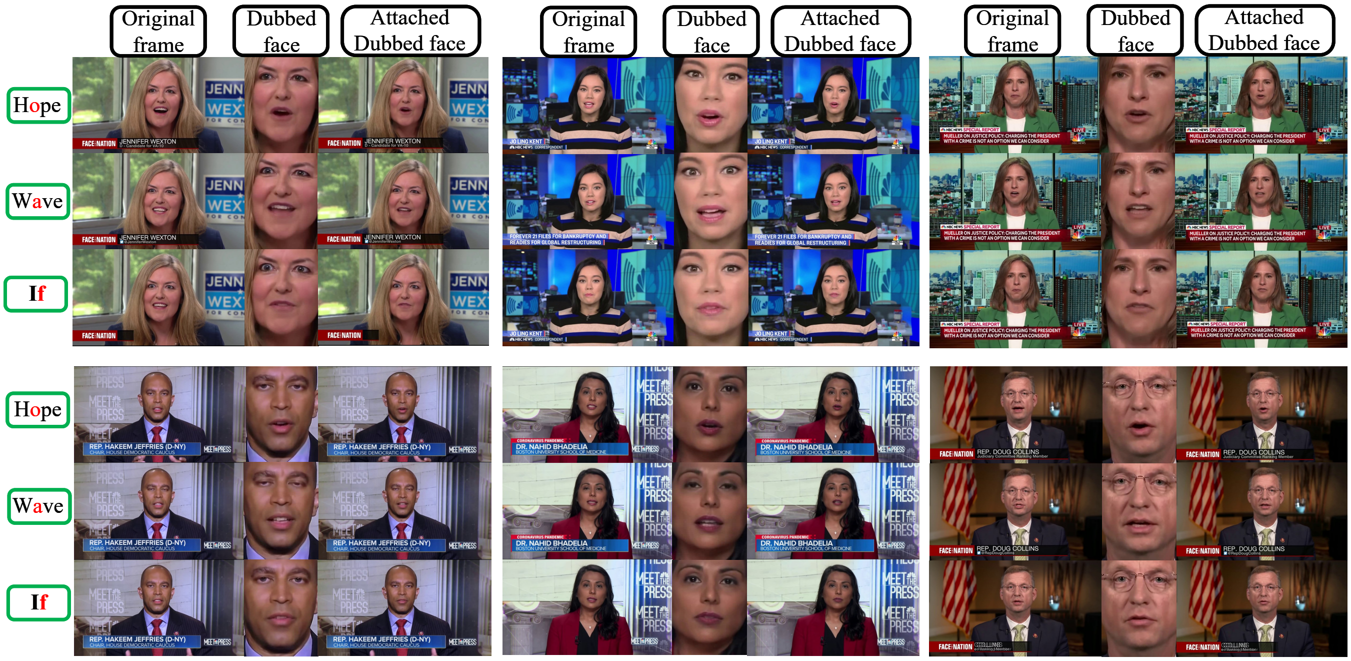

Figure 3: The dubbed results of our DINet (Please zoom in for more details). Our method achieves realistic face visually dubbing on 1080P videos. More results can be observed in the demo video.

Perception loss. Similar to (Zhang and Ding 2022), we compute perception loss in two image scale. Specifically, we downsample dubbed image $I_o \in R^{3\times H\times W}$ and real image $I_r \in R^{3\times H\times W}$ into $\hat{I}_o \in R^{3\times \frac{H}{2}\times \frac{W}{2}}$ and $\hat{I}_r \in R^{3\times \frac{H}{2}\times \frac{W}{2}}$. Then, paired images $\{I_o, I_r\}$ and $\{\hat{I}_o, \hat{I}_r\}$ are input into one pretrained VGG-19 network (Simonyan and Zisserman 2014) to compute perception loss. The perception loss is written as

$$\mathcal{L}_p = \sum_{i=1}^{N} \frac{\|V_i(I_o) - V_i(I_r)\|_1 + \|V_i(\hat{I}_o) - V_i(\hat{I}_r)\|_1}{2NW_iH_iC_i}, \tag{2}$$

where $V_i(.)$ represents i_{th} layer in VGG-19 network. $W_iH_iC_i$ is the feature size in i_{th} layer.

GAN loss. We use effective LS-GAN loss (Mao et al. 2017) in our method. The GAN loss is written as

$$\mathcal{L}_{GAN} = \mathcal{L}_D + \mathcal{L}_G, \tag{3}$$

where

$$\mathcal{L}_D = \frac{1}{2}E(D(I_r) - 1)^2 + \frac{1}{2}E(D(I_o) - 0)^2 \tag{4}$$

$$\mathcal{L}_G = E(D(I_o) - 1)^2. \tag{5}$$

G represents DINet and D denotes discriminator. We use GAN loss on both single frame and five consecutive frames. The structural details of D are in supplementary materials.

Lip-sync loss. As similar in (Prajwal et al. 2020), we add a lip-sync loss to improve the synchronization of lip movements in dubbed videos. We replace audio spectrogram with deep speech feature and re-train the sycnet. The structural details of sycnet are in supplementary materials. The lip-sync loss is written as

$$\mathcal{L}_{sync} = E(sycnet(A_d, I_o) - 1)^2. \tag{6}$$

We sum above losses as final loss $\mathcal{L}$, which is written as

$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_{sync}\mathcal{L}_{sync} + \mathcal{L}_{GAN}. \tag{7}$$

where λ_p and λ_{sync} denote the weights of $\mathcal{L}_p$ and $\mathcal{L}_{sync}$. We set $\lambda_p = 10$ and $\lambda_{sync} = 0.1$ in our experiment.

Experiment

In this section, we first detail datasets and implementation details in our experiment. Then, we show synthetic results of our method. Next, we carry out qualitative and quantitative comparisons with other state-of-the-art works. Next, we conduct ablation studies. Finally, an online user study is conducted to furthermore validate our method.

Datasets

We conduct experiments on two common high-resolution talking face datasets: HDTF dataset (Zhang et al. 2021b) and MEAD dataset(Wang et al. 2020).

HDTF dataset. HDTF dataset contains about 430 in-the-wild videos collected with 720P or 1080P resolution. We randomly select 20 videos for testing in our experiments.

MEAD dataset. MEAD dataset records around 40 hours emotional in-the-lab videos at 1080P resolution. In our work, we do not focus on emotional face visually dubbing, so we select a total of 1920 videos with neutral emotion and frontal view as **MEAD-Neutral** dataset. In MEAD-Neutral dataset, we randomly select 240 videos of 6 identities for testing.

Table 1: Quantitative comparisons with the state-of-the-art methods on talking face generation.

	HDTF					**MEAD-Neutral**				
	SSIM↑	PSNR↑	LPIPS↓	LSE-D↓	LSE-C↑	SSIM↑	PSNR↑	LPIPS↓	LSE-D↓	LSE-C↑
ATVG (Chen et al. 2019)	0.7315	20.5430	0.1104	8.4222	7.0611	0.8325	23.9723	0.0869	8.8908	5.8337
Wav2Lip-96 (Prajwal et al. 2020)	0.9078	29.2875	0.0576	6.8714	8.2908	0.8994	28.5387	0.0779	**6.0218**	**8.9587**
Wav2Lip-192	0.8487	27.6561	0.1208	8.0912	6.9509	0.8036	25.6863	0.1302	7.8426	6.8515
Wav2Lip-384	0.8963	28.4695	0.0760	11.5373	3.5162	0.8415	25.9716	0.1166	9.8241	3.8001
MakeitTalk (Zhou et al. 2020)	0.5969	19.8602	0.1592	11.4913	3.0293	0.8283	25.2988	0.1037	11.4256	2.5128
PC-AVS (Zhou et al. 2021)	0.6383	20.6301	0.1077	**6.6137**	**8.8550**	0.7754	23.6950	0.0960	6.5035	8.6240
Grount Truth	1.0000	N/A	0.0000	6.4989	8.9931	1.0000	N/A	0.0000	6.2937	9.6747
DINet (Ours)	**0.9425**	**30.0082**	**0.0289**	8.3771	6.8416	**0.9204**	**29.1177**	**0.0459**	7.5157	7.2603

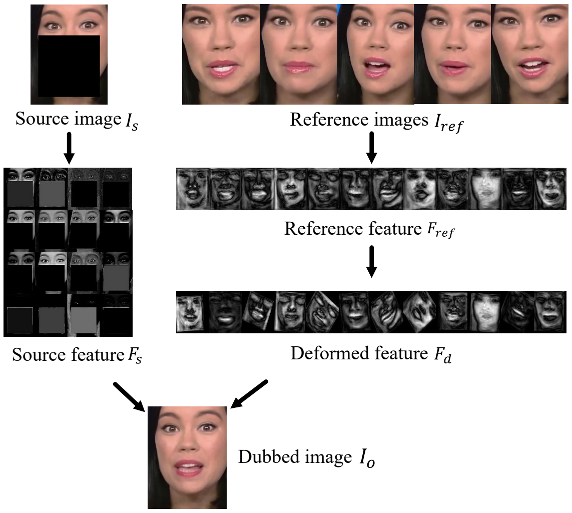

Figure 4: Visualization of intermediate results in DINet, including the source image I_s, the reference images I_{ref}, the source feature maps F_s, the reference feature maps F_{ref}, the deformed feature maps F_d and the dubbed image I_o.

Implementation Details

In data processing, all videos are resampled in 25 fps. We crop face region according to 68 facial landmarks of openface (Baltrušaitis, Robinson, and Morency 2016) and resize all faces into 416×320 resolution. The mouth region covers 256×256 resolution in resized facial image. More details about how we crop face are in supplementary materials. Considering that HDTF dataset and MEAD dataset have limited subjects, we use pre-trained deepspeech model (Hannun et al. 2014) to extract audio features of 29 dimention to improve generalization. The audio feature is aligned with video in 25 fps.

In training stage, DINet inputs one source frame $I_s \in R^{3\times416\times320}$, one driving audio $A_d \in R^{5\times29}$ and five reference facial images $I_{ref} \in R^{15\times416\times320}$. Syncnet inputs 5 frames of mouth images (256×256) and corresponding deep speech features. We use Adam optimizer (Kingma and Ba 2014) with default setting to optimize DINet and syncnet. The learning rate is set to 0.0001. The batch size is set to 3 in DINet and 20 in syncnet on four A30 gpu.

Synthetic Results

Figure 3 shows the dubbed results of our method. We display the original 1080P frames, the synthetic results of DINet and the original frames with attached synthetic face of six identities. Our method realizes realistic face visually dubbing on 1080P videos.

Figure 4 visualizes the intermediate results in DINet, including the source image I_s, the reference images I_{ref}, the source feature maps F_s, the reference feature maps F_{ref}, the deformed feature maps F_d and the dubbed image I_o. It indicates that F_s encodes spatial features both inside and outside the mask region. F_{ref} encodes different spatial features of each frame in I_{ref}, so F_{ref} is spatially misaligned. AdaAT performs different spatial deformation being specific to feature channels in F_{ref} to generate F_d. The deformation is rich and contains scaling, rotation and translation, etc.

Comparisons with State-of-the-art Works

We compare our method with state-of-the-art one$-$/few$-$ shot talking head works that is open sourced, including ATVG (Chen et al. 2019), Wav2Lip (Prajwal et al. 2020), MakeitTalk (Zhou et al. 2020) and PC-AVS (Zhou et al. 2021). Wav2Lip is the most relevant to our method, but their original model is trained on 96×96 resolution. For a fair comparison, we retrain their framework in 192×192 and 384×384 resolution. We denote trained models in two resolutions as Wav2Lip-192 and Wav2Lip-384 respectively.

Qualitative comparisons. We first compare with state-of-the-art works in qualitative comparisons. Figure 5 illustrates the results. ATVG only generates 128×128 resolution videos, while our DINet can generate more realistic videos with 416×320 resolution. Wav2Lip synthesizes face with 96×96 resolution, and their mouth region becomes blurry when attaching the face into 1080P video. Wav2Lip-192 and Wav2Lip-384 still synthesize blurry results, although they are trained on high resolution videos. The main reason is that wav2lip utilizes networks to dierctly generate pixels in mouth region, making networks easily neglect textural details. Our DINet deforms existing textural details into appropriate locations of mouth region, thus achieves more realistic results.

MakeitTalk tends to generate inaccurate mouth shape (see the red circle). This is caused by two main reasons: their intermediate facial landmarks are not accurate enough, and facial landmarks are too sparse to describe lip motion details. In contrast, our DINet generates more accurate lip motions

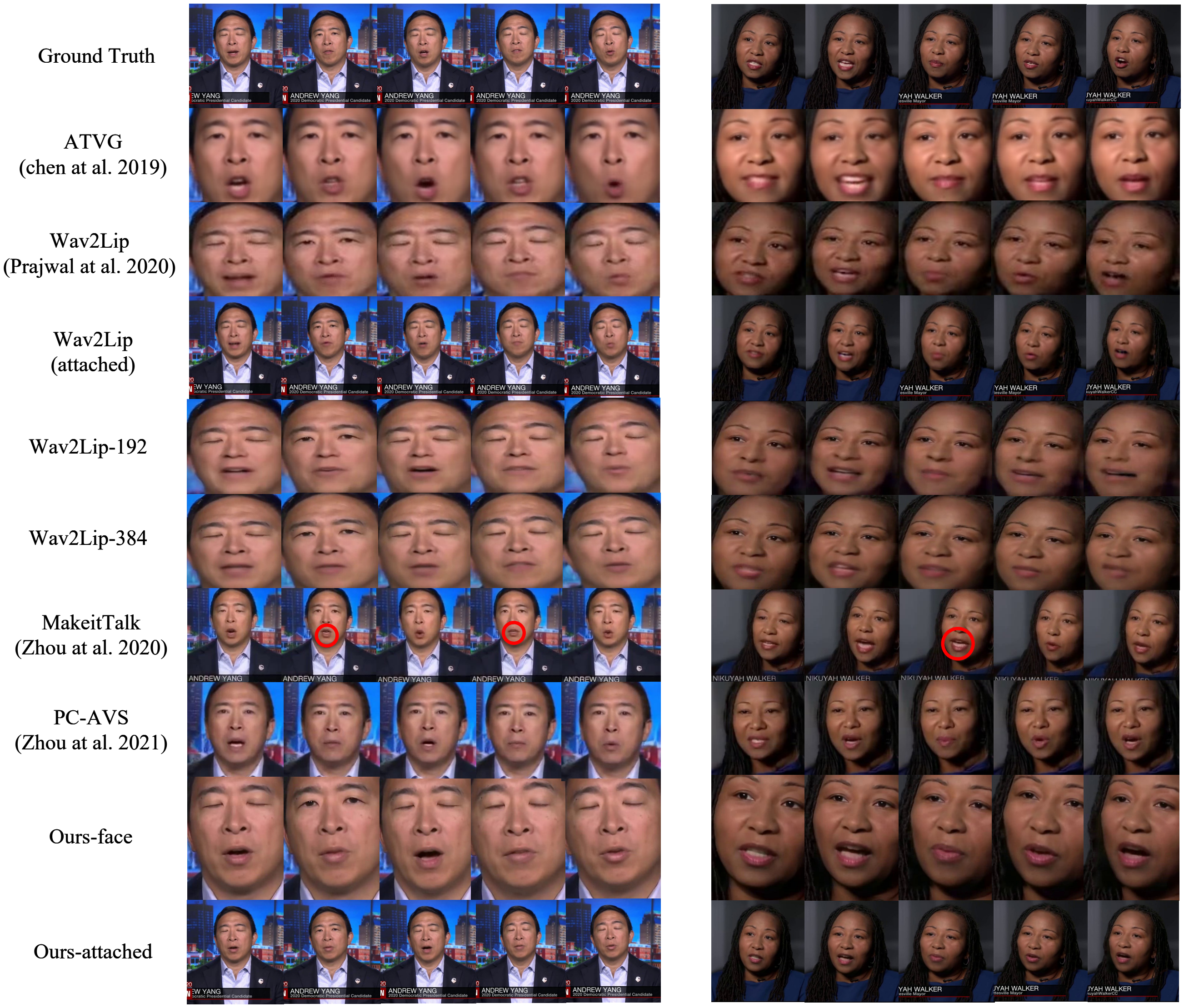

Figure 5: Qualitative comparisons with the state-of-the-art works (Please zoom in for more details). Our method achieves high resolution face visually dubbing.

by directly learning a mapping between audio and mouth image. PC-AVS generates 224 × 224 resolution videos. Their resolution is limited due to direct-generation fashion. This limitation is statemented in their following work (Liang et al. 2022). Instead of directly generating pixels, our DINet utilizes operations of deformation and inpainting to realize high resolution face visually dubbing.

Quantitative comparisons. Table 1 shows results of quantitative comparisons. To evaluate the visual quality, we compute the metrics of Structural Similarity (SSIM) (Wang et al. 2004), Peak Signal to Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). To evaluate the audio-visual synchronization, inspired from (Prajwal et al. 2020), we compute the metrics of Lip Sync Error Distance (LSE-D) and Lip Sync Error Confidence (LSE-C). In the evaluation of visual quality, our DINet gets the best results on all metrics. In the evaluation of audio-visual synchronization, our method is worse than Wav2Lip and PC-AVS. One possible reason is that metrical syncnet is trained on LRS2 dataset, while our model is trained from scratch.

Ablation Study

We conduct ablation experiments to validate each component in our DINet. Specifically, we set 5 conditions: (1) *Ours w/o deformation*: we remove the AdaAT operation in DINet and inject F_{audio} and F_{align} into F_s with AdaIN operation (Huang and Belongie 2017). (2) *Ours w/o inpainting*: we directly generate I_o from I_d without fusing F_s. (3) *Ours w attention*: we replace AdaAT operation with Attention (Vaswani et al. 2017) to select features from F_{ref} instead of deformation. (4) *Ours w dense flow*: we replace the

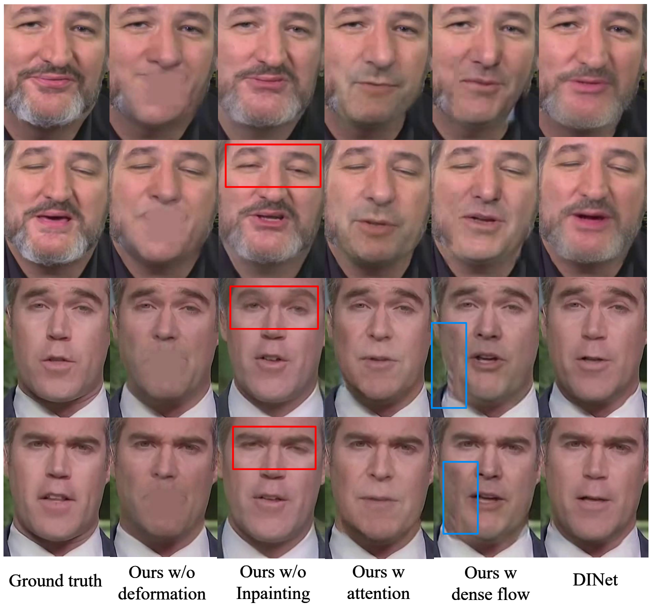

Figure 6: Qualitative results of ablation study.

AdaAT operation with dense flow to realize spatial deformation on F_{ref}. (5) *Ours*: proposed DINet.

Figure 6 illustrates the qualitative results of ablation experiments. In the condition of *Ours w/o deformation*, the synthetic facial images have blurry mouth. It indicates that it is difficult for networks to directly generate pixels in mouth region on high resolution videos. In the condition of *Ours w/o inpainting*, the synthetic facial images have vivid textural details, including beard and nasolabial fold, due to effective deformation of AdaAT. However, without F_s providing the information out of mouth region, upper face easily has incorrect performance, e.g., mismatched eyeblink and wrong sight line (see the red rectangle in Figure 6) .

In the condition of *Ours w attention* and *Ours w dense flow*, synthetic facial images lose more texture details. One possible reason is that ”attention” and ”dense flow” do ”pixel-level deformation” while AdaAT operation do ”region-level deformation”. In ”pixel-level deformation”, each pixel is deformed flexibly into appropriate location and there is no regularization on the deformation. In ”region-level deformation”, region in a whole feature map do same affine transformation and the deformation is regularized by the number of feature channel. On one hand, too flexible deformation makes ”attention” and ”dense flow” are easily overfitting on the dataset. On the other hand, under a few-shot condition, mouth shape has higher correlations with driving audio than textural details, leading to networks focus on synthesizing synchronized lip motions instead of textural details. AdaAT has a strong regularization on the deformation, thus has a better performance on preserving textural details and facial structures than ”attention” and ”dense flow”.

Besides, *Ours w dense flow* generate images with apparent artifacts in facial region (see blue rectangle in Figure 6). One possible reason is that dense flow requires feature maps have similar spatial layouts. However, reference feature map F_{ref} has different layouts due to the different head pose in reference images. Same deformation on misaligned feature maps may cause the generation of artifacts.

Table 2: Quantitative results of ablation study on the HDTF dataset.

Method	SSIM↑	PSNR↑	LPIPS↓
Ours w/o deformation	0.9147	26.1665	0.0519
Ours w/o inpainting	0.8691	26.0071	0.0561
Ours w attention	0.9153	27.2451	0.0433
Ours w dense flow	0.9001	23.5656	0.0656
Ours	**0.9425**	**30.0082**	**0.0289**

Table 2 shows the quantitative results of ablation experiments. Our DINet gets the best visual quality according to the metrics of SSIM, PSNR and LPIPS.

User Study

One user study is conducted for subjective evaluation. We randomly download five 1080P videos and one driving audio from internet. In one-shot works, we select the first frame as reference facial image and generate videos according to driving audio. In face visually dubbing works, we loop video frames if the length of video is shorter than driving audio. 20 volunteers are invited to rate realism of each synthetic video from 1 (pretty bad) to 5 (pretty good). The rating values are ATVG (2.9), Wav2Lip(2.4), MakeitTalk(3.1), PC-AVS (3.3) and ours (3.4). We get the best rating score.

Limitations

Our method achieves high-resolution face visually dubbing, yet we still suffer from several challenging conditions. Our DINet deforms reference images to inpaint mouth region in source face, so it can not handle conditions of changeable lighting, dynamic background, pendulous earrings, flowing hair and camera movements, and may generate artifacts out of face if mouth region covers background. The training videos in HDTF and MEAD-Neutral only have frontal view, so our method is restricted to limited head pose. Sometimes, the synthetic facial images are sensitive to the selection of reference images.

Conclusion

In this paper, we propose a Deformation Inpainting Network (DINet), including one deformation part and one inpainting part, to realize high-fidelity face visually dubbing. The deformation part is designed to perform spatial deformation of feature maps of reference images to synchronize mouth shape with the driving audio and also align head pose with the source image. The inpainting part is designed to merge the deformed feature maps and the source feature map, to inpaint the mouth region in source image. With the combination of deformation and inpainting, our DINet preserves more textural details than the existing methods. Extensive qualitative and quantitative experiments have validated the performance of our method on high resolution face visually dubbing. In the future, we will make great efforts on solving above limitations.

Acknowledgments

This work is supported by the 2022 Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054) and the Key Research and Development Program of Zhejiang Province (No. 2022C01011).

References

Baltrušaitis, T.; Robinson, P.; and Morency, L.-P. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10. IEEE.

Chen, L.; Cui, G.; Liu, C.; Li, Z.; Kou, Z.; Xu, Y.; and Xu, C. 2020. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, 35–51. Springer.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 520–535.

Chen, L.; Maddox, R. K.; Duan, Z.; and Xu, C. 2019. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7832–7841.

Chung, J. S.; Jamaludin, A.; and Zisserman, A. 2017. You said that? *arXiv preprint arXiv:1705.02966*.

Das, D.; Biswas, S.; Sinha, S.; and Bhowmick, B. 2020. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *European conference on computer vision*, 408–424. Springer.

Doukas, M. C.; Zafeiriou, S.; and Sharmanska, V. 2021. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14398–14407.

Fried, O.; Tewari, A.; Zollhöfer, M.; Finkelstein, A.; Shechtman, E.; Goldman, D. B.; Genova, K.; Jin, Z.; Theobalt, C.; and Agrawala, M. 2019. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Guo, Y.; Chen, K.; Liang, S.; Liu, Y.-J.; Bao, H.; and Zhang, J. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5784–5794.

Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; et al. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. *Advances in neural information processing systems*, 28.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. *arXiv preprint arXiv:2205.15278*.

Ji, X.; Zhou, H.; Wang, K.; Wu, W.; Loy, C. C.; Cao, X.; and Xu, F. 2021. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14080–14089.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

KR, P.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; and Jawahar, C. 2019. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*, 1428–1436.

Lahiri, A.; Kwatra, V.; Frueh, C.; Lewis, J.; and Bregler, C. 2021. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2755–2764.

Liang, B.; Pan, Y.; Guo, Z.; Zhou, H.; Hong, Z.; Han, X.; Han, J.; Liu, J.; Ding, E.; and Wang, J. 2022. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3387–3396.

Lin, C.-H.; and Lucey, S. 2017. Inverse compositional spatial transformer networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2568–2576.

Liu, J.; Zhu, Z.; Ren, Y.; Huang, W.; Huai, B.; Yuan, N.; and Zhao, Z. 2022. Parallel and High-Fidelity Text-to-Lip Generation.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 405–421. Springer.

Park, S. J.; Kim, M.; Hong, J.; Choi, J.; and Ro, Y. M. 2022. SyncTalkFace: Talking Face Generation with Precise Lip-syncing via Audio-Lip Memory. In *36th AAAI Conference on Artificial Intelligence (AAAI 22)*. Association for the Advancement of Artificial Intelligence.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. 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*, 484–492.

Ren, Y.; Li, G.; Chen, Y.; Li, T. H.; and Liu, S. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13759–13768.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, L.; Wu, W.; Qian, C.; He, R.; and Loy, C. C. 2022. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17: 585–598.

Song, Y.; Zhu, J.; Li, D.; Wang, X.; and Qi, H. 2018. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13.

Thies, J.; Elgharib, M.; Tewari, A.; Theobalt, C.; and Nießner, M. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, 716–731. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vougioukas, K.; Petridis, S.; and Pantic, M. 2020. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5): 1398–1413.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation. In *ECCV*.

Wang, S.; Li, L.; Ding, Y.; Fan, C.; and Yu, X. 2021. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*.

Wang, S.; Li, L.; Ding, Y.; and Yu, X. 2022a. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2531–2539.

Wang, T.-C.; Mallya, A.; and Liu, M.-Y. 2021. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10039–10049.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022b. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. *arXiv preprint arXiv:2203.09043*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, H.; Jia, J.; Wang, H.; Dou, Y.; Duan, C.; and Deng, Q. 2021. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1478–1486.

Xie, T.; Liao, L.; Bi, C.; Tang, B.; Yin, X.; Yang, J.; Wang, M.; Yao, J.; Zhang, Y.; and Ma, Z. 2021. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1739–1747.

Yao, G.; Yuan, Y.; Shao, T.; and Zhou, K. 2020. Mesh guided one-shot face reenactment using graph convolutional networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1773–1781.

Zhang, C.; Zhao, Y.; Huang, Y.; Zeng, M.; Ni, S.; Budagavi, M.; and Guo, X. 2021a. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3867–3876.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; and Ding, Y. 2022. Adaptive Affine Transformation: A Simple and Effective Operation for Spatial Misaligned Image Generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1167–1176.

Zhang, Z.; Li, L.; Ding, Y.; and Fan, C. 2021b. 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*, 3661–3670.

Zhou, H.; Liu, Y.; Liu, Z.; Luo, P.; and Wang, X. 2019. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 9299–9306.

Zhou, H.; Sun, Y.; Wu, W.; Loy, C. C.; Wang, X.; and Liu, Z. 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4176–4186.

Zhou, Y.; Han, X.; Shechtman, E.; Echevarria, J.; Kalogerakis, E.; and Li, D. 2020. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6): 1–15.