Expressive Talking Head Generation with Granular Audio-Visual Control

Borong Liang[1*] Yan Pan[2,3*] Zhizhi Guo[1†] Hang Zhou[1†] Zhibin Hong[1]
Xiaoguang Han[2,3] Junyu Han[1] Jingtuo Liu[1] Errui Ding[1] Jingdong Wang[1]
[1]Department of Computer Vision Technology (VIS), Baidu Inc.,
[2]SSE, CUHK-Shenzhen, [3]FNii, CUHK-Shenzhen
{liangborong,zhouhang09,guozhizhi,hongzhibin,hanjunyu,liujingtuo,dingerrui,wangjingdong}@baidu.com,
{yanpan@link.,hanxiaoguang@}cuhk.edu.cn.

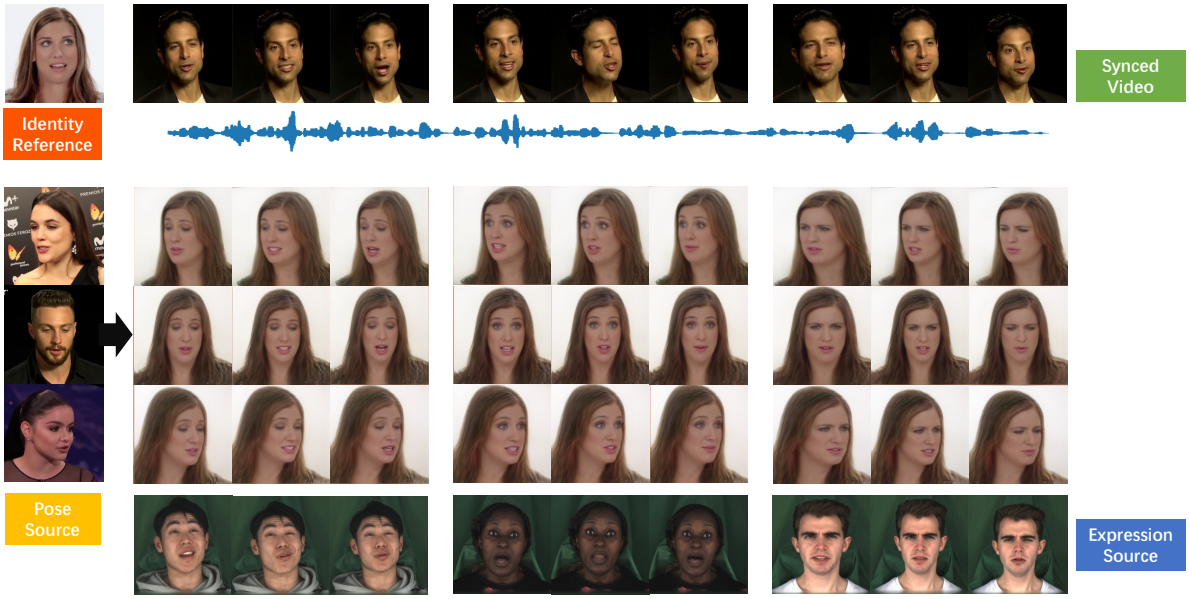

Figure 1. Example animations generated by our **Granularly Controlled Audio-Visual Talking Heads (GC-AVT)**. Given a reference identity frame, **GC-AVT** generates audio-visual driven talking head video with other emotional expression source and pose source video frames independently. The mouth shapes of driven results are matched with the synced video (on top row), and the expressions of driven results are matched with the expression source (on bottom row) while the poses are matched with the pose source (left column).

Abstract

Generating expressive talking heads is essential for creating virtual humans. However, existing one- or few-shot methods focus on lip-sync and head motion, ignoring the emotional expressions that make talking faces realistic. In this paper, we propose the ***Granularly Controlled Audio-Visual Talking Heads (GC-AVT)****, which controls lip movements, head poses, and facial expressions of a talking head in a granular manner. Our insight is to decouple the audio-visual driving sources through prior-based pre-processing designs. Detailedly, we disassemble the driving image into three complementary parts including: 1) a cropped mouth that facilitates lip-sync; 2) a masked head that implicitly learns pose; and 3) the upper face which works corporately and complementarily with a time-shifted mouth to contribute the expression. Interestingly, the encoded features from the three sources are integrally balanced through reconstruction training. Extensive experiments show that our method generates expressive faces with not only synced mouth shapes, controllable poses, but precisely animated emotional expressions as well.*

1. Introduction

With the rapid development of automatic video generation technology, the task of audio-driven talking head gen-

*Equal contribution.
†Corresponding authors.

eration has drawn much attention due to its extensive real-world applications such as creating virtual anchors, digital avatars, and animated movies. In order to achieve convenient deployment with a generalized model, researchers have proposed to drive only a single or a few frames to talk with audios [8, 10, 46, 48, 50, 51]. While accurate lip sync has been almost realized, the ability to control the facial expression, which is crucial for creating human-like talking heads, has not been fully explored.

A great number of previous methods focus only on the lip-sync accuracy with audios [7, 10, 27, 30, 48]. More recently, researchers propose to generate rhythmic [6, 42] or changeable head poses [50] along with talking heads. However, their methods cannot change detailed expressions such as eyebrows. On the other hand, methods that generate emotional dynamics [18, 21, 35] are basically person-specific, *i.e.*, one model has to be trained for one specific person. Moreover, their models rely on labeled emotional data, thus can only cover limited expressions.

In real-world scenes, people could speak the same content with fixed stress and intonation but flexible expressions and head motions. Inspired by this observation, we argue that the generation of emotional expressions can be divided from mouth movements and poses, that the three of them could be controlled independently. This is technically challenging for nearly all existing models. 1) For methods predicting intermediate structural representation such as 2D or 3D landmarks [6, 8, 51], the above information is inherently entangled. Even mainstream 3D face models, such as 3D Face Morphable Model (3DMM) [1], represent mouth movement and facial expression within the same parameter. Besides, the accuracy of intermediate representations will be compromised under extreme cases. 2) For latent feature learning methods [3, 10, 48, 50], the expression information can hardly be individually distilled, and current works [3, 50] do not support disentangled expression and mouth control.

In this work, we propose **Granularly Controlled Audio-Visual Talking Heads (GC-AVT)**, which drives a portrait head from a higher level of granularity. Avoid using any intermediate representation, our method is pure learning-based without specific emotion labels. The most intriguing property of our model is the independent facial control from three complementary perspectives: *speech content*, *head pose*, and *emotional expression*, which makes our talking head more expressive. As shown in Fig 1, while the head pose and expression information are derived from visual sources, the mouth movement can be decided by either audio or visual information.

Our insight is to *explicitly divide the driving information into granular parts through delicate pre-processing designs*. Different from previous methods that learn *non-identity representation* in a holistic view [3, 50], we argue that all information can be separately extracted in a complementary manner. We analyze the *key-factors* that affect each desired facial area and adopt different types of masking and augmentation schemes. Three functional inputs are thus formulated. *Audio input* associates explicitly with the mouth shapes, thus the temporal alignment between speech and cropped mouths is leveraged to account for the speech content information. Then we expect that the *emotional expressions* could be driven by additional visual sources. In particular, we factorize the emotion of a whole face into an upper-face and a time-shifted mouth. The two of them are seamlessly collaborated together to provide precise expressions. Finally, an implicit *pose* code is devised from the whole face. Three encoders are leveraged for the individual information extraction, and a style-based generator processes them through reconstruction training. Experiments demonstrate that our method manages to generate an expressive talking head with the precise mouth shape, head pose, and emotional expression control.

The contributions of this work are summarized as follows: (1) We propose the **Granularly Controlled Audio-Visual Talking Heads (GC-AVT)** System, which generates expressive portrait videos from the granular control of pose, audio, and an expression video. (2) We identify three delicate pre-processing procedures for handling the three different control sources. (3) By integrating audio-visual synchronization, our system generates accurate mouth movements that can be driven by either audio or video.

2. Related work

Audio-Driven Talking Head Generation. The task of animating virtual humans [5, 22, 23, 50, 52] from arbitrary speech sequence has drawn considerable attention in both computer vision and graphics, among which talking face generation is particular important. Earlier works [32, 36, 37] require a large number of video footage of a target person by modeling the mouth area through either retrieval or graphics-based methods. With the develop of deep learning, a number of works leverage structural information within GAN-based pipelines [18, 25, 29, 33] to generate person-specific high-quality results. Other researchers tend to seek speaker-independent settings that can address all identities through one or more framework references [8, 10, 30, 48, 50, 51]. Chung *et al.* [10] firstly propose an end-to-end reconstruction-based network in an image-to-image translation manner based on audios. Then [48] uses adversarial training to further separate identity from word. Wav2Lip [27] particularly proposes to inpaint the mouth areas. The basic idea behind these reconstruction-based methods is to synchronize mouth motion in video with speech content in audio. Facial expressions and head poses, on the other hand, are neglected.

More recently a few methods [6, 18, 31, 50, 51] have been

proposed not only to solve the problem of synchronization but add extra components to create a more vivid talking head. Zhou *et al.* [48] and Yi *et al.* [51] models rhythmic head motions with 3D representations. PC-AVS [50] leverage another pose source video to control head poses while driving talking faces with audio sequences. [46] produces the animation parameters of mouth, eyebrow and head pose simultaneously and synthesizes talking face videos from dense flow. Particularly, Wang *et al,* [35] and Ji *et al.* [18] propose to alter emotions, but one model has to be trained on one person. Controlling different attributes of the portrait video simultaneously in the one-shot manner has been proven to be difficult.

Visually Driven Face Reenactment. The task of sace reenactment aims to generate talking head videos by transferring the facial dynamics from a different actor's video. Most techniques rely on structural information such as landmarks [16, 40, 43, 44] or 3D models [4, 14, 19, 20, 34, 47]. Deep Video Portraits [20] is capable of producing high-quality photo-realistic dubbing results. It keeps the target actor's identity and pose while capturing the source actor's facial emotions, but should be trained per target video. Recently, FReeNet [44] utilize a unified landmark converter to transfer facial expressions between identities. Moreover, latent pose descriptors based on the reconstruction losses [3, 24] are proposed for cross-person reenactment. These works aim to handle multi-identity face reenactment, and our work expands the task's complexity by involving granularly control.

3. Our Approach

In this section, we describe our **Granularly Controlled Audio-Visual Talking Heads (GC-AVT)** system, which encodes the head pose, speech content, emotional expression, together with person identity into latent spaces and generates the driven talking head with either audio or video. First, we briefly introduce the pipeline of our approach in Sec. 3.1. Next, we introduce the prior-based face preprocessing which is crucial for devising independent granular control sources (Sec. 3.2). Finally, we introduce the learning process of the pipeline 3.3.

3.1. Overall Formulation

The whole pipeline of our method is illustrated in Fig. 2. We adopt the typical cross-frame self-reconstruction [10, 50] setting for training, and expect the driving information of speech content, pose, and expression could originate from completely different videos during inference.

Given a pre-processed video clip with N frames $V = (I_1, \ldots, I_N)$ and its corresponding audio spectrograms $A = (a_1, \ldots, a_N)$, we sample a set of K frames $\{I_{i1}, I_{i2}, ..., I_{iK}\}$ from V randomly as the representatives for identity information. This representation is supervised by a simple identity loss [3]. Then we randomly sample one frame I_k from V as the source of all driving conditions (*i.e.* expression, pose, and speech content). Our goal is to recover I_k based on the corresponding audio spectrogram a_k and the desired information from I_k. This is inherently difficult for two reasons: (1) The input source I_k is also the target, the network might take a shortcut during the reconstruction. (2) The granular information desired is entangled together and difficult to discriminate and extract.

To this end, we propose that each desired driving part can be specifically identified from the input image domain. Specifically, I_k is decomposed into three complementary parts through delicate prior-based pre-processing. As the identity information also requires modeling, a total of four visual encoders independently encode the identity, head pose, emotional expression, and speech content (mouth shape) information into latent features named f_{id}, f_p, f_e, f_c^v respectively. Specifically, f_c^v is further leveraged to assist the learning of the audio feature f_c^a. The two features should lie in a same latent space. At last, we expect that one generator G is capable of handling all information. The features can be assembled together as the overall audio-based feature $f_{all}^a = \{f_{id}, f_p, f_e, f_c^a\}$ or visual-based feature $f_{all}^v = \{f_{id}, f_p, f_e, f_c^v\}$. They are sent into G for reconstructing $I_k^{a\prime}$ and $I_k^{v\prime}$.

The detailed pre-processing steps are described in Sec. 3.2 and the learning objectives are illustrated in Sec. 3.3.

3.2. Prior-based Pre-Processing

As stated above, three particular types of pre-processing paradigms are designed based on the prior knowledge of different functional areas of a face. Each of them corresponds to a driving source, representing disentangled information.

While detailed pre-processing procedures are different, identity information should be removed from all sources. Specifically, it is achieved by *pixel-wise augmentation* consisting of color transfer, blurring, sharpening, and JPEG compression. This augmentation is applied to all three pre-processing steps.

On the other hand, masking is widely applied in our implementations, where the landmarks of I_k and the foreground segmentation map are detected. The segmentation map is also used for wiping out background interference. Note that we do not leverage the landmarks as an intermediate representation. They are used only as guidance for data pre-processing, thus we do not suffer from the error accumulation problem caused by inaccurate predictions.

Pre-processing for Expression. The extraction of expression information alone without the semantics on mouth shapes has rarely been achieved before. One plausible way is to mask out the mouth based on landmarks around it. This

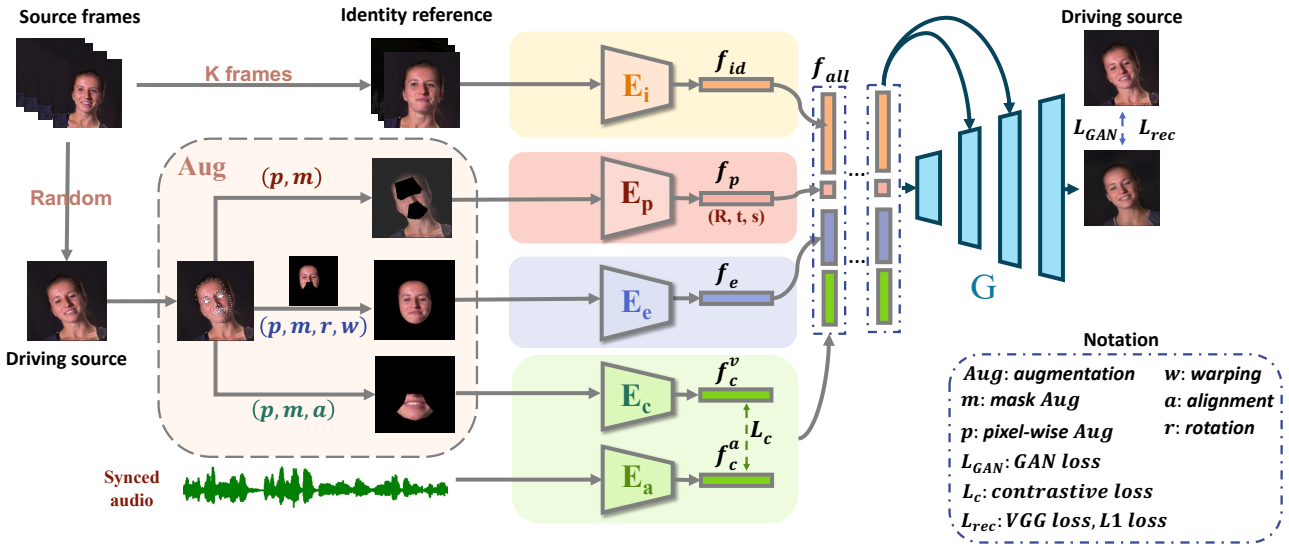

Figure 2. The proposed architecture of our Granularly Controlled Audio-Visual Talking Heads (GC-AVT). The data pre-processing sample $K+1$ frames from a video frame sequence, one of the selected frames is used to generate training data for pose encoder (E_p), emotional expression encoder (E_e), and content encoder (E_c) through different data augmentation methods, which will be described in Sec. 3.2. The rest K frames are used as input to the identity encoder (E_i) and will be encoded into latent feature f_{id}. The pose encoder and the emotional expression encoder encode the corresponding augmented images into f_p and f_e respectively. To encode the speech content information, we design a visual-audio synchronization network (E_c and E_a) that encode the visual frame and the audio spectrum into latent feature f_c^v and f_c^a. The features are assembled together and fed to the Generator. The learning of the pipeline is described in Sec. 3.3.

is to maintain the expressions on the upper face. However, the influence of the mouth cannot be directly ignored. Emotional information also has effects on the mouth, *e.g.*, we can infer that a person is smiling and talking at the same time by looking at the mouth movements only.

Our method is built upon the observation that the semantics in mouth shapes change much more rapidly than emotion. For example, a person rarely changes the emotion and even the head pose within one second but could speak several syllables. Thus we argue that a shortly time-shifted frame I_{k+i} could possess the same emotional but different semantic information in mouth shapes with I_k.

Specifically, the mouth areas are cropped out from I_{k+i}. When i is reasonably small, the time-shifted mouth can be seamlessly blended to I_k. In this way, the precise expression and emotion information on the mouth are preserved. Furthermore, an additional *random rotation* is applied for erasing the pose information.

Pre-processing for Speech Content. The encoding of speech content information from visual modality is intended as a particular type of guidance for audio information encoding. Specifically, researchers have verified that the intrinsic temporal audio-visual synchronization lies around the mouths [13, 27]. Thus we leverage a cropped out mouth of I_k. The *random rotation* is also applied to the speech content processing.

Pre-processing for Pose. It is simple and safe to mask out the facial organs on a talking head to represent the head pose information. We also devise the latent pose space as a dimension of 12 and rely solely on networks for learning the implicit pose information in a fully reconstruction-based network as performed in [50].

3.3. Learning Procedures

Except for the simple learning objective on the identity features, other learning constraints are designed from two perspectives: 1) The constraints on speech content features which synchronizes audio to the visual modality; 2) the constraints on the reconstructed frames $I_k^{a\prime}$ and $I_k^{v\prime}$ (uniformly denoted as $I_k{}'$) that implicitly balance the information within all embeddings.

Learning Audio-Visual Synchronization. It has been verified that learning audio-visual synchronization benefits audio-visual cross-generation tasks generation [27, 31, 48–50], and it would be easier to learn mouth shapes from the visual domain [48].

Thus in order to stabilize the training, we prevent the synchronization loss from affecting the visual branch and update the audio branch alone. Detailedly, we adopt softmax contrastive loss. The distances between two features are measured as $\mathcal{D}(f_c^v, f_c^a) = \frac{f_c^{v\mathbf{T}} * f_c^a}{|f_c^v| \cdot |f_c^a|}$, where f_c^v and f_c^a are timely assembled visual and audio features from consecutive frames. Supposing a total of M^- negative samples

are leveraged, the contrastive learning is formulated as:

$$\mathcal{L}_c = -\log[\frac{\exp(\mathcal{D}(f_c^v, f_c^a))}{\exp(\mathcal{D}(f_c^v, f_c^a)) + \sum_{j=1}^{M^-} \exp(\mathcal{D}(f_{c(j)}^{v-}, f_c^a))}], \tag{1}$$

where $f_{c(j)}^{v-}$ denotes the jth negative sample.

Reconstruction Objectives. We directly borrow the generator structure from [3], which relies on the AdaIN [17]. Note that the same set of losses are applied to both audio and visual reconstructed images as $I_k^{a'}$ and $I_k^{v'}$. The reconstruction training is generally supervised by pixel-wise comparing L_1 distances between I_k' and I_k. Two VGG-19 models, one pre-trained on ImageNet classification and one on face recognition are leveraged in the perceptual loss manner [26, 38], where a total of N_{vgg} feature maps are leveraged. The three loss functions can be written as:

$$\mathcal{L}_{L_1} = \|I_k - I_k'\|_1, \tag{2}$$

$$\mathcal{L}_{\text{vgg}} = \sum_{i=1}^{N_{vgg}} \|\text{VGG}_i(I_k) - \text{VGG}_i(I_k')\|_1 + \sum_{i=1}^{N_{vgg}} \|\text{VGG}_i^{Face}(I_k) - \text{VGG}_i^{Face}(I_k')\|_1. \tag{3}$$

To further improve the generation quality, a multi-scale discriminator D with N_D layers is involved with the generative adversarial loss:

$$\mathcal{L}_{\text{GAN}} = \min_{\text{G}} \max_{\text{D}} \sum_{n=1}^{N_D} (\mathbb{E}_{I_k}[\log \text{D}_n(I_k)] + \mathbb{E}_{f_{all(k)}}[\log(1 - \text{D}_n(I_k'))], \tag{4}$$

The overall constraints during training can be summarized as:

$$\mathcal{L}_{all} = \mathcal{L}_{\text{GAN}} + \lambda_1 \mathcal{L}_{L_1} + \lambda_2 \mathcal{L}_{\text{vgg}} + \lambda_3 \mathcal{L}_c, \tag{5}$$

where the λs are coefficients.

Notably, we not only constrain the embedding space of audio and visual speech content features but also use both of them for reconstruction training. Thus our method supports talking face generation with mouth shapes driven by both an audio clip or a mouth sequence.

4. Experiments

4.1. Experimental Settings

Dataset. Our method is trained on VoxCeleb2 [11] and evaluated on both Voxceleb2 and MEAD [35].

- **VoxCeleb2** [11] is an audio-visual dataset which is popularly used in the area of talking head generation. We use the URLs provided by VoxCeleb2 to download the original videos, collecting roughly 2,000 speaker identities for training and 100 for evaluation.
- **MEAD** [35] is a high-quality emotional audio-visual dataset with over 30 available actors/actresses and eight emotion categories at three different intensity levels. The frontal-view videos in this dataset are leveraged only for testing.

Implementation Details: All videos are processed at 25 frames per second. For each frame, we detect the face with S^3FD detector [45], then enlarge the bounding box by 80% to keep the face in the center. The final cropped images are of size 256×256. We apply the Graphonomy [15] model to get background segmentation and mask out the background in the pre-processing. We retrain a FAN model [2] to get landmarks for each image. Similarly to [50], we process the audios to 16kHz, then convert them to mel-spectrograms with FFT window size 1280, hop length 160 and 80 Mel filter-banks. For each video frame, 0.2s mel-spectrogram with the target frame time-step in the middle are sampled as condition.

In our method, the ID encoder is a ResNeXt-50 [41] structure. We set $K = 8$ for the input of the identity encoder and 512 dimension for the identity embedding output. Both of the pose encoder and emotional expression encoder are the MobileNetV2 [28] structure. The pose and emotional expression embedding sizes are 12 and 256 respectively. The content encoder and audio encoder are ResNetSE34 borrowed from [9], each generating a 256-dimension embedding. We train our model for 80 epoch with a minibatch of 16 samples on 32 GB Tesla V100 GPUs. We pretrain the visual-audio synchronization with the contrastive loss L_c then joint train the whole pipeline end-to-end.

Comparing Methods: Our method focuses on audio-driven talking head generation, thus we mainly compare the audio-driven results of **Ours (audio)** with state-of-the-arts audio-driven works [3, 27, 50]. **Wav2Lip** [27] is a reconstruction-based method that focuses on producing accurate lip movements; **MakeitTalk** [51] is based on 3D landmarks for learning personalized head movements under the audio-driven setting. **PC-AVS** [50] is also a reconstruction-based framework and can generate lip synchronization while controlling pose implicitly. Note that our model could also adopt the visual-driven setting, thus we compare the visual-driven results of **Ours (video)** with **LPD** [3], a head reenactment system. We compare all the results generated by non-fine-tuned models directly for fairness.

4.2. Quantitative Evaluation

Evaluation Metrics: To quantitatively evaluate different methods, we compute four evaluation metrics under the

Table 1. The comparison of quantitative results on Voxceleb2 [11] and MEAD [35]. For LMD and LMD_m the lower the better, and the higher the better for other metrics. Note that in this comparison the **PC-AVS** [50] fails on some frames because of the landmark detecting failure and the results of it are just for reference.

Method	MEAD				VoxCeleb2			
	SSIM ↑	LMD ↓	LMD_m ↓	Sync_{conf} ↑	SSIM ↑	LMD ↓	LMD_m ↓	Sync_{conf} ↑
Ground Truth	1.000	0.000	0.000	4.770	1.000	0.000	0.000	5.543
Wav2Lip	**0.747**	3.543	4.014	**4.674**	0.704	4.139	3.662	5.218
MakeItTalk	0.618	4.102	4.249	3.926	0.624	5.358	4.689	4.887
PC-AVS	0.605	3.963	4.334	3.248	0.606	5.101	4.654	4.986
ours (audio)	0.659	**2.764**	**3.252**	3.730	**0.710**	**3.025**	**3.356**	**5.250**
LPD	0.669	2.762	2.966	3.355	0.707	4.176	4.035	**5.213**
ours (video)	**0.671**	**2.483**	**2.349**	**3.435**	**0.739**	**2.757**	**2.811**	5.149

self-driven setting on the test set of VoxCeleb2. They are: **SSIM** [39] for generation quality; **LMD** for mean distance of all landmarks and $\textbf{LMD}_m$ for landmarks around the mouths. We also borrow the confidence score $\textbf{Sync}_{conf}$ from SyncNet [12] to evaluate the precision of lip synchronization.

Evaluation Results: We use a similar experimental setting with PC-AVS [50]. Specifically, we select the first frame of each test video as the identity reference. Then the rest frames are used as the sources of pose, emotional expression and speech information. The audios are used as driving conditions to generate audio-driven results. We calculate the numerical metrics between the generated results and the ground truth.

The results are shown in Table 1. In this comparison, our **GC-AVT** achieves comprehensively better results on both datasets. Note that audio-driven methods and visual-driven ones are not directly comparable, so we analysis them separately, and focus more on the audio-driven setting. In terms of the lip sync accuracy, our audio setting achieves a better $\textbf{LMD}_m$ than other methods, which proves that we can generate good lip sync quality from one perspective. Though we do not possess the highest confidence score ($\textbf{Sync}_{conf}$). Our results are close to the ground truth, which show competitive performance. Note that Wav2Lip directly uses SyncNet in its loss function, thus naturally leading to better results on this metric. Benefited from the pose control and expression manipulation ability, our method is naturally better on the general LMD metric. The SSIM score is suitable for Wav2Lip as their only inpaint missing areas. As for the visual-driven setting, we observe several failure cases in the LPD results, making their LMD and LMD_m results lower than ours.

4.3. Qualitative Evaluation

Comparisons with Other Methods. The comparing methods do not support granular control, therefore it's unfair to set too detailed sources. Since LPD [3] and PC-AVS [50] can control the head pose of generated video, here we assign the pose source, speech content source, and expression source all as one single video denoted as *driving source* on the Figure 3. Note that MakeitTalk [51] can neither control pose nor generate accurate mouth shapes, thus we neglect its results here.

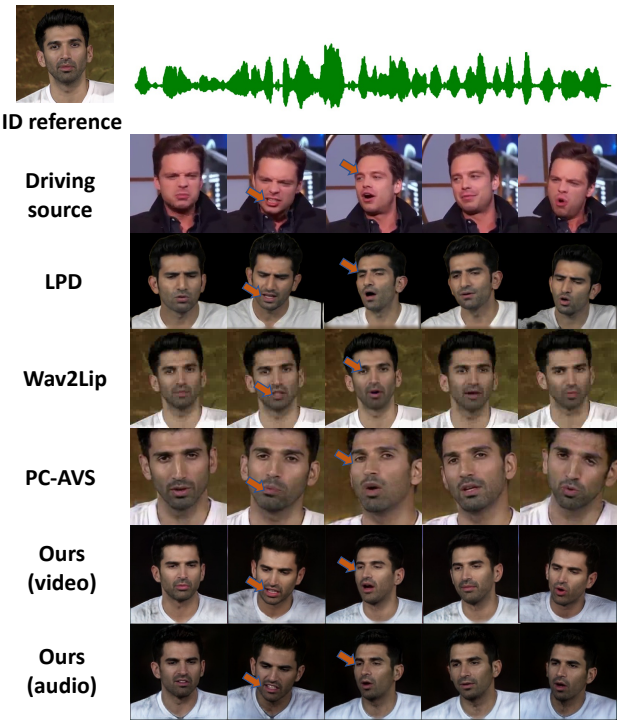

Figure 3. Qualitative evaluation results. The driving source frames are listed in first row. **Wav2Lip** [27] fails to generate the frames with head pose similar to the driving source. The **PC-AVS** [50] can generate most images with similar head pose to the driving source but the result in second column is not quite accurate. Both **LPD** [3] and **our GC-AVT** can generate driven results with accurate head pose. The expression driven results are better than **LPD** [3].

We can see that Wav2Lip [27] can only leverage the pose of the original video. Its background will be fixed still when its input is a single image (see demo video). While PC-

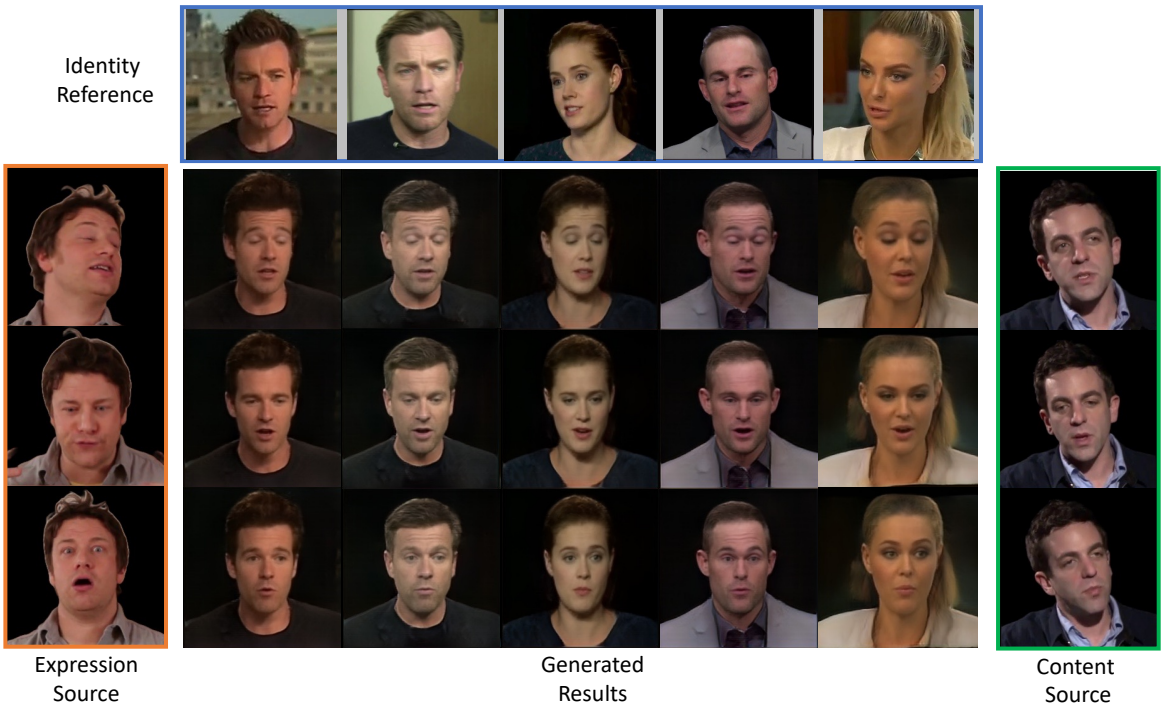

Figure 4. The qualitative results with different driving expression source and content source. The first row lists the identity reference while the expression and content source frames are listed in the left column and right column respectively. Our **GC-AVT** can generate the vivid driven results with the corresponding expression source and content source.

AVS [50] can mimic the pose of the driving source, its results are not quite precise on certain cases. Both of them can only generate neutral expressions. The pose driving results of LPD [3] are quite close to ours. Both generated results of LPD [3] and Ours (video) have precise head pose with the driving source. It can also be seen from the second column that, our pre-processing scheme enables the successful emotional expression transfer from the source video to our results. While such information is neglected in LPD. In terms of the lip sync accuracy, we can see that both our visual- and audio-driven results generate high-fidelity mouth shapes which are aligned with the driving source and outperforms the results of PC-AVS.

Evaluation on Emotional Expression Control. A remarkable feature of our **GC-AVT** is that we can control the emotional expression independently from the semantic mouth shapes and head poses. We visualize the independent control of emotional expression, and speech content in Figure 4. We frontalize all the generated results. As can be seen that in the process of independent control, the emotional expression and the speech content can be well decoupled.

User Study. To further verify the quality of *audio-driven* results by organizing a user study of 20 participants for their opinions on 50 videos. Specifically, we randomly sample 5 videos as the driving source videos and 10 identity reference images from Voxceleb2 dataset. Then we generate the 50 videos with the same setting as we described in Sec. 4.1. The comparing methods are **Wav2Lip** [27], **MakeitTalk** [51], **PC-AVS** [50] and our **GC-AVT** respectively. The evaluation of user study is developed on three dimensions for users: (1) Lip Sync Quality; (2) Expression Realness and Richness. (3) Overall Fidality and Quality. The widely used Mean Opinion Scores (MOS) is adopted with rating scores from 1 to 5.

The rating results of our user study are listed in Table 2. Our GC-AVT outperforms previous methods on the expression realness and richness by a large margin, which verifies the effectiveness of our method in handling emotional expressions. And our results are apparently more vivid than others. Although we do not score the best in lip sync quality, the results between the three methods are very close and can be regarded as comparable.

4.4. Ablation Study

In this section, we study the effects of the losses setting and the necessity of time-shift operation. Note that the experiments are carried out on the VoxCeleb2 dataset with our **audio-driven** setting.

For loss setting, we study the effects of VGG loss, VGGFace loss and Contrastive loss. The results are listed in Table 3, where w/o VGG means without both VGG loss and VGGFace loss. The contrastive loss is used for audio-visual synchronization. In order to verify the effects of contrastive loss, we test the LMD, LMD_m and Sync_{conf} in Voxceleb2

Table 2. User study on audio-driven methods, the evaluations are conducted on lip synchronization, the naturalness of facial expression and video quality.

Method	Wav2Lip [27]	MakeItTalk [51]	PC-AVS [50]	GC-AVT (Ours)
Lip Sync Quality	**3.92**	2.85	3.90	3.91
Expression Realness and Richness	2.65	2.68	3.16	**4.21**
Overall Fidelity and Quality	3.33	3.06	3.69	**3.95**

Table 3. Ablation study on Voxceleb2 [11].

Method	SSIM	LMD	LMD_m	Sync_{conf}
w/o VGG	0.662	4.753	4.212	4.586
w/o Contrastive	0.692	4.890	4.311	4.066
w/o time-shift	0.684	4.311	3.704	4.760
Ground Truth	1.000	0.000	0.000	5.543
ours (audio)	**0.710**	**3.025**	**3.356**	**5.250**

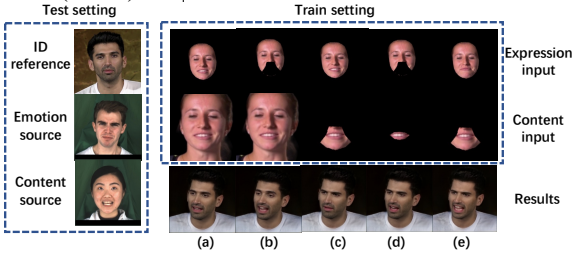

Figure 5. Ablation study of the masking areas.

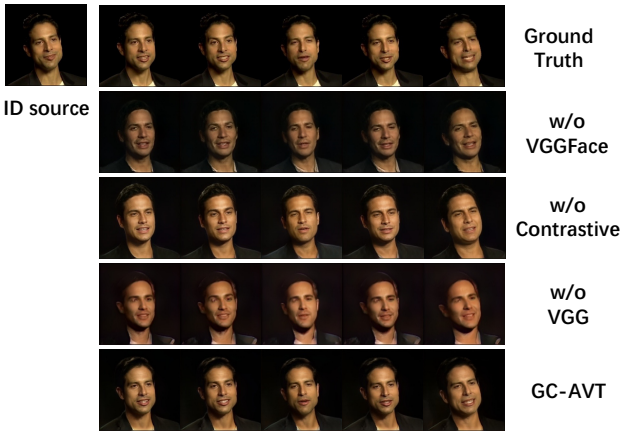

Figure 6. Ablation study for losses setting with visual results. As shown in second row and the fourth row, without either VGGFace loss or VGG loss, the quality of generated results decreased significantly. The speech content driving results are affected without the contrastive loss as shown in the third row.

testset. As demonstrated in Table 3, the performances of LMD, LMD_m and Sync_{conf} all get worse prominently. Besides, we visualize the results of the ablation studies in Figure 6. Without the perceptual losses such as VGG loss and VGGFace loss, the quality of generated images are obviously poor, and the performance of attribute control is also worse than the results of our complete setting. The speech content driving results are affected when we remove the contrastive loss. The speech driven results are not synced with the driving source. Without the time-shift operation the speech driven results is affected but the quality of the generated image is hardly affected.

We further show the ablation studies on the mask designs. Experiments are carried out on the following settings shown in Fig. 5: **(a)** no masks are applied; **(b)** no mask on mouth; **(c)** no mask on expression; **(d)** smaller mouth area; and **(e)** time-shifted mouth on expression. Setting **(a)**, **(b)**, **(c)** would confuse the training procedure of the networks, which eventually leads to the loss of the speech content control ability. The results of setting **(a)** - **(f)** are shown in the figure below. Our setting **e)** achieves the best results. The qualitative and quantitative comparisons will be added to the final version.

5. Conclusion and Discussion

Conclusion. In this paper, we propose the **Granularly Controlled Audio-Visual Talking Heads (GC-AVT)** pipeline. By explicitly divide the driving information into granular parts through delicate pre-processing designs, **GC-AVT** supports talking head generation controlled from the perspectives of speech-content, pose, expressions. To the best of our knowledge, such property has rarely been achieved before. Moreover, it supports accurate lip sync from both audio and visual inputs, which enlarges applications of our system.

Limitations. One of most important limitations is that the backgrounds are masked our in our method, thus we cannot handle sophisticated background changes. Moreover, our method cannot generate high resolution results.

Ethical Statements. Although animating talking heads has extensive applications, it might be misused for deepfake creation and media manipulation. We will restrict the usage of our model and share it with the deepfake detection community.

Acknowledgement. The work was supported in part by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen-HK S&T Cooperation Zone, National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, and by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104.

References

[1] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999. 2

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 5

[3] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 5, 6, 7

[4] Lele Chen, Chen Cao, Fernando De la Torre, Jason Saragih, Chenliang Xu, and Yaser Sheikh. High-fidelity face tracking for ar/vr via deep lighting adaptation, 2021. 3

[5] Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201*, 2020. 2

[6] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. *European Conference on Computer Vision (ECCV)*, 2020. 2

[7] Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[8] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[9] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. In defence of metric learning for speaker recognition. *arXiv preprint arXiv:2003.11982*, 2020. 5

[10] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? In *BMVC*, 2017. 2, 3

[11] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018. 5, 6, 8

[12] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, 2016. 6

[13] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV*, 2016. 4

[14] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[15] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7450–7459, 2019. 5

[16] Po-Hsiang Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Learning identity-invariant motion representations for cross-id face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[17] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 5

[18] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chan Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[19] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 2019. 3

[20] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 2018. 3

[21] Lincheng Li, Suzhen Wang, Zhimeng Zhang, Yu Ding, Yixing Zheng, Xin Yu, and Changjie Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2

[22] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[23] Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. Semantic-aware implicit neural audio-driven video portrait generation. *arXiv preprint arXiv:2201.07786*, 2022. 2

[24] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)*, 37(4):1–13, 2018. 3

[25] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021. 2

[26] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[27] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. 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 (ACMMM)*, 2020. 2, 4, 5, 6, 7, 8

[28] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[29] Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. Everybody's talkin': Let me talk as you want. *arXiv preprint arXiv:2001.05201*, 2020. 2

[30] Yang Song, Jingwen Zhu, Dawei Li, Xiaolong Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. *IJCAI*, 2019. 2

[31] Yasheng Sun, Hang Zhou, Ziwei Liu, and Hideki Koike. Speech2talking-face: Inferring and driving a face with synchronized audio-visual representation. In *IJCAI*, 2021. 2, 4

[32] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 2017. 2

[33] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[34] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobaltzhang2020freenet, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[35] 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. 2, 3, 5, 6

[36] Lijuan Wang, Wei Han, and Frank K Soong. High quality lip-sync animation for 3d photo-realistic talking head. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012. 2

[37] Lijuan Wang, Xiaojun Qian, Wei Han, and Frank K Soong. Synthesizing photo-real talking head via trajectory-guided sample selection. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010. 2

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004. 6

[40] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[41] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[42] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with natural head pose. *arXiv preprint arXiv:2002.10137*, 2020. 2

[43] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[44] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[45] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 5

[46] 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 (CVPR)*, 2021. 2, 3

[47] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[48] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2, 3, 4

[49] Hang Zhou, Ziwei Liu, Xudong Xu, Ping Luo, and Xiaogang Wang. Vision-infused deep audio inpainting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 4

[50] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4, 5, 6, 7, 8

[51] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: Speaker-aware talking head animation. *SIGGRAPH ASIA*, 2020. 2, 3, 5, 6, 7, 8

[52] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)*, 2018. 2