Identity-Preserving Talking Face Generation with Landmark and Appearance Priors

Weizhi Zhong[1]   Chaowei Fang[2]   Yinqi Cai[1]   Pengxu Wei[1]
Gangming Zhao[3]   Liang Lin[1]   Guanbin Li[1*]

[1]Sun Yat-sen University   [2]Xidian University   [3]The University of Hong Kong

{zhongwzh5, caiyq27}@mail2.sysu.edu.cn, chaoweifang@outlook.com

{weipx3, liguanbin, linlng}@mail.sysu.edu.cn, gangmingzhao@gmail.com

Abstract

Generating talking face videos from audio attracts lots of research interest. A few person-specific methods can generate vivid videos but require the target speaker's videos for training or fine-tuning. Existing person-generic methods have difficulty in generating realistic and lip-synced videos while preserving identity information. To tackle this problem, we propose a two-stage framework consisting of audio-to-landmark generation and landmark-to-video rendering procedures. First, we devise a novel Transformer-based landmark generator to infer lip and jaw landmarks from the audio. Prior landmark characteristics of the speaker's face are employed to make the generated landmarks coincide with the facial outline of the speaker. Then, a video rendering model is built to translate the generated landmarks into face images. During this stage, prior appearance information is extracted from the lower-half occluded target face and static reference images, which helps generate realistic and identity-preserving visual content. For effectively exploring the prior information of static reference images, we align static reference images with the target face's pose and expression based on motion fields. Moreover, auditory features are reused to guarantee that the generated face images are well synchronized with the audio. Extensive experiments demonstrate that our method can produce more realistic, lip-synced, and identity-preserving videos than existing person-generic talking face generation methods.

1. Introduction

Audio-driven talking face video generation is valuable in a wide range of applications, such as visual dubbing [19, 26, 37], digital assistants [32], and animation movies [46]. Based on the training paradigm and data requirement,
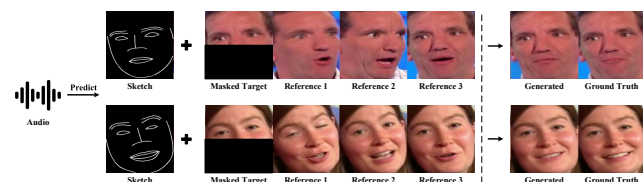
*Corresponding author is Guanbin Li.

Figure 1. This paper is targeted at generating a talking face video for a speaker which is coherent with input audio. We implement this task by completing the lower-half face of the speaker's original video. The outline of the mouth and jaw is inferred from the input audio and then used to guide the video completion process. Moreover, multiple static reference images are used to supply prior appearance information.

the talking face generation methods can generally be categorized as person-specific or person-generic types. Person-specific methods [11, 20, 21, 28, 32, 41] can generate photo-realistic talking face videos but need to be re-trained or fine-tuned with the target speaker's videos, which might be inaccessible in some real-world scenarios. Hence, learning to generate person-generic talking face videos is a more significant and challenging problem in this field. This topic also attracts lots of research attention [14, 19, 24, 26, 37, 45]. In this paper, we focus on tackling the person-generic talking face video generation by completing the lower-half face of the speaker's original video under the guidance of audio data and multiple reference images, as shown in Figure 1.

The main challenges of the person-generic talking face video generation include two folds: 1) How can the model generate videos having facial motions, especially mouth and jaw motions, which are coherent with the input audio? 2) How can the model produce visually realistic frames while preserving the identity information? To address the first problem, many methods [3, 5, 6, 17, 37, 46] leverage facial landmarks as intermediate representation when generating person-generic talking face videos. However, translation from audio to facial landmarks is an ambiguous task, considering the same pronunciation may correspond to mul-

tiple facial shapes. A few landmark-based talking face generation methods [3, 42] tend to produce results having the averaged lip shape of training samples, which may have remarkable differences with the lip shape of the speaker. A line of methods [6, 37] incorporate the prior information from a reference image's landmarks to generate landmarks consistent with the speaker's shape. However, they directly fuse the features of audio and landmarks with simple concatenation or addition operations without modeling the underlying correlation between them. For example, the relations between the reference landmarks and the audio clips from different time intervals are different. Those methods are not advantageous at capturing such kinds of differences. Moreover, the temporal dependencies are also valuable for predicting facial landmarks. Existing methods, such as [6, 17, 38, 46], depend on long short-term memory (LSTM) models to explore the temporal dependencies when transforming audio clips to landmark sequences. However, those models are limited in capturing long-range temporal relationships.

Since the input audio and intermediate landmarks do not contain visual content information intrinsically, it is very challenging to hallucinate realistic facial videos from audio and intermediate landmarks while preserving the identity information. A few existing methods, such as [37, 46], adopt a single static reference image to supply visual appearance and identity information. However, one static reference image is insufficient to cover all facial details, e.g., the teeth and the side content of cheeks. This makes these algorithms struggle to synthesize unseen details, which is unreliable and easily leads to generation artifacts. [3, 7] use multiple reference images to provide more abundant details. However, they simply concatenate the reference images without spatial alignment, which is limited in extracting meaningful features from reference images.

To cope with the above problems, we devise a novel two-stage framework composed of an audio-to-landmark generator and a landmark-to-video rendering network. The goal of our framework is to complete the lower-half face of the video with content coherent to the phonetic motions of the audio. Specifically, we use pose prior landmarks of the upper-half face and reference landmarks extracted from static face images as extra inputs of the audio-to-landmark generator. The access to the two kinds of landmarks helps to prevent the generator from producing results that deviate from the face outline of the speaker. Then, we build up the network architecture of the generator based on the multi-head self-attention modules [33]. Our design is more advantageous at capturing relationships between phonetic units and landmarks compared to simple concatenation or addition operations [6, 37]. It is also more helpful for modeling temporal dependencies than LSTM used in previous methods [6, 17, 38, 46]. Additionally, multiple static face

images are referred to extract prior appearance information for generating realistic and identity-preserving face frames. Inspired by [10], we set up the landmark-to-video rendering network with a motion field based alignment module and a face image translation module. The alignment module is targeted at registering static reference images with the face pose and expression delivered by the results of the landmark generator. This target is achieved by inferring a motion field for each static reference image and then warping the image and its features. The alignment module can decrease the difficulty in translating meaningful features of static reference images to the target image. The face image translation module produces the final face images by combining multi-source features from the inferred landmarks, the occluded original images, the registered reference images, and the audio. The inferred landmarks provide vital clues for constraining the facial pose and expression. Those images are paramount for inferring the facial appearance. Besides, the audio features are reused to guarantee that the generated lip shapes are well synchronized with the audio. Extensive experiments demonstrate that our method produces more realistic and lip-synced talking face videos and preserves the identity information better than existing methods. Our main contributions are summarized as follows:

- We propose a two-stage framework composed of an audio-to-landmark generator and a landmark-to-video rendering model to address the person-generic talking face generation task under the guidance of prior landmark and appearance information.

- We devise an audio-to-landmark generator that can effectively fuse prior landmark information with the audio features. We also make an early effort to construct the generator with multi-head self-attention modules.

- We design a landmark-to-video rendering model which can make full use of multiple source signals, including prior visual appearance information, landmarks, and auditory features.

- Extensive experiments are conducted on LRS2 [1] and LRS3 [2] dataset, demonstrating the superiority of our method over existing methods in terms of realism, identity preservation, and lip synchronization.

2. Related Work

2.1. Audio-Driven Talking Face Generation

Existing audio-driven talking face generation methods mainly include two types, person-specific and person-generic methods. With the help of 3D Morphable Model (3DMM) [4] and neural radiance fields (NeRF) [23], some person-specific methods [11, 20, 28, 32, 39, 41, 44] can synthesize high-fidelity talking face videos. For example, NVP

[32] and FACIAL [41] first predict 3DMM expression parameters from audio and then use a neural rendering network to produce videos. A few works [11, 20, 28] attempt to implement the audio-driven talking face video generation by controlling the dynamic neural radiance field with the audio and then rendering face images. Nevertheless, all these methods require videos of the target speaker for re-training or fine-tuning, which may be inaccessible in some real-world scenarios. It is more significant to devise person-generic methods which can synthesize talking face videos for unseen speakers. There exists a series of literature focusing on person-generic talking face generation, such as [6,16,24,26,34,37,45,46]. Typically, Wav2Lip [26] uses an encoder-decoder based generator to synthesize talking face videos under the guidance of a lip sync discriminator. Based on Wav2Lip, SyncTalkFace [24] proposes an audio-lip memory to provide extra visual information of the mouth region. PC-AVS [45] modularizes talking faces into feature spaces of speech content, head pose, and identity, respectively. These features are assembled to produce talking face videos. However, these methods are insufficient in generating highly realistic and lip-synced videos while preserving identity information.

2.2. Landmark-based Talking Face Generation

Many audio-driven talking face generation methods [3,5,6,17,30,31,37,38,42,46] use facial landmarks as intermediate representation. For example, Suwajanakorn *et al.* [30] use a recurrent neural network (RNN) to learn the mapping from audio input to mouth landmarks and then synthesize high-quality mouth texture. [6,17,21,31] rely on Long Short Term Memory (LSTM) models to learn the mapping from audio to landmark movements. MakeItTalk [46] combines LSTM and self-attention mechanisms to predict the landmark displacement from audio. APB2Face [42] uses linear layers to predict landmark geometry from the concatenated pose, blink, and audio features. Xie *et al.* [37] devise a two-stage framework where three encoders are used to extract the audio, pose, and reference embeddings in the first stage, and three embeddings are fused with trainable weights and fed into a decoder to predict the landmarks of the lip and jaw in the second stage. Our method is also composed of two stages, including audio-to-landmark generation and landmark-to-video rendering. It is distinct from existing methods in the following aspects. First, built upon Transformer modules, the audio-to-landmark generator is more advantageous at exploring prior landmark information to predict facial landmarks accurately. Secondly, the landmark-to-video rendering model can effectively combine multi-source features from prior appearance information, landmarks, and audio, achieving better performance in terms of realism, lip synchronization, and identity preservation than existing methods.

3. Proposed Method

Given an audio sequence and an initial input video, we aim to generate a lip-synced talking face video by completing the lower-half occluded face of the input video in a frame-by-frame manner. An overview of our method is depicted in Figure 2. Our framework is composed of two stages. The first stage takes the audio signal and prior landmarks of the speaker's face as input to predict the landmarks of the lip and jaw. The second stage consists of an alignment module and a translation module. Based on the motion fields, the alignment module registers reference images and their features with the target face pose and expression. The translation module synthesizes the full face image from landmarks under the guidance of auditory features and prior appearance information from the occluded target face and registered reference images.

3.1. Audio-To-Landmark Generation

In this stage, the network aims to generate landmarks of lip $\{\hat{l}_t^l \in \mathbb{R}^{2 \times n_l}\}_{t=1}^T$ and jaw $\{\hat{l}_t^j \in \mathbb{R}^{2 \times n_j}\}_{t=1}^T$ for $T = 5$ adjacent frames at a time, given the reference landmarks $\{l_i^r \in \mathbb{R}^{2 \times n_r}\}_{i=1}^{N_l}$, pose prior landmarks $\{l_t^p \in \mathbb{R}^{2 \times n_p}\}_{t=1}^T$, and audio input $\{m_t \in \mathbb{R}^{h \times w}\}_{t=1}^T$. N_l is the number of reference landmarks (measured in frames). n_l, n_j, n_r, and n_p is the number of landmarks representing lip, jaw, whole face, and pose, respectively. Following [26], for each video frame, we extract out the corresponding audio interval and process it to Mel-spectrogram with size of 16×80, *i.e.*, $h = 16$ and $w = 80$. The reference landmarks provide prior personalized facial outline information for landmark prediction.

3.1.1 Transformer-based Landmark Generator

We extract the audio embedding a_t from the audio Mel-spectrograms with an encoder module E_a similar to [26]. We also utilize 1D convolution layers to construct a pose encoder E_p and a reference encoder E_r which extract pose embedding p_t and reference embedding r_i from pose prior landmarks l_t^p and reference landmarks l_i^r, respectively. This process is formulated as follows:

$$a_t = E_a\left(m_t\right) \quad t = 1, 2, ..., T \quad (1)$$
$$p_t = E_p\left(l_t^p\right) \quad t = 1, 2, ..., T \quad (2)$$
$$r_i = E_r\left(l_i^r\right) \quad i = 1, 2, ..., N_l \quad (3)$$

where $a_t, p_t, r_i \in \mathbb{R}^d$ (d is the dimension of audio, pose, or reference embedding).

For differentiating embeddings calculated from there types of source signals, we introduce three learnable encoding vectors e_a^{type}, e_p^{type}, and $e_r^{type} \in \mathbb{R}^d$, indicating that the embedding is calculated from audio Mel-spectrogram, prior pose landmarks, and reference landmarks, respectively. The
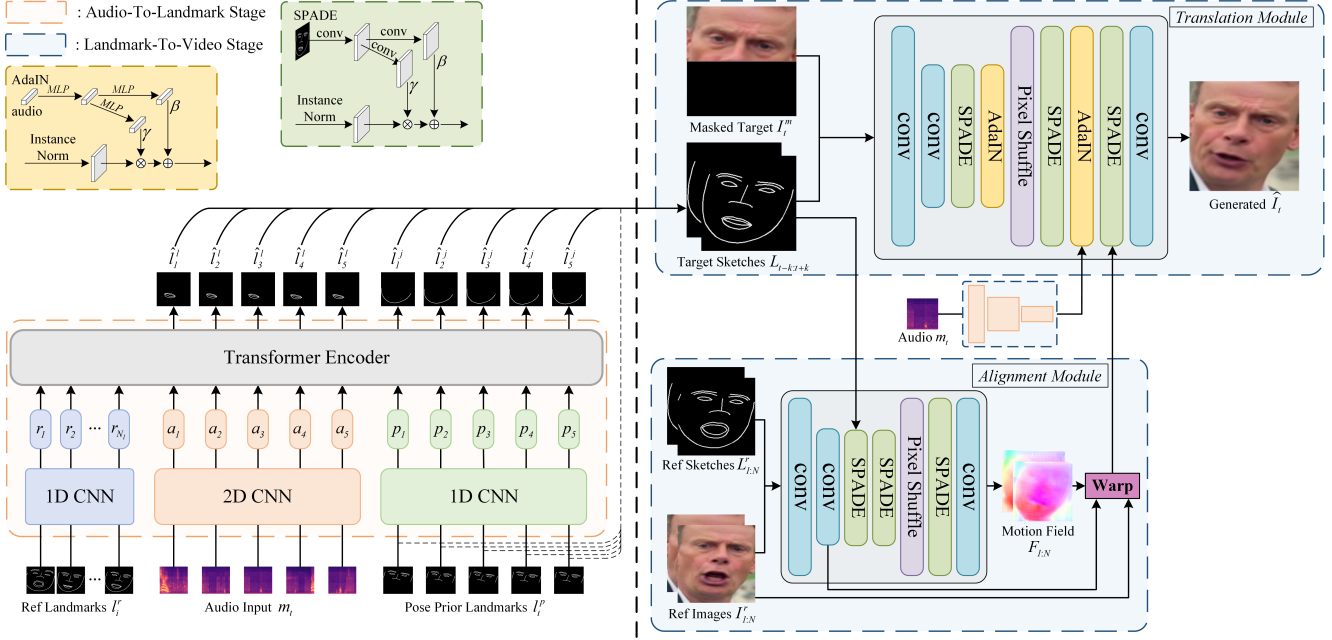
Figure 2. Overview of our framework. It can be divided into two stages: **(1) Audio-To-Landmark Generation** (left orange part). The transformer-base landmark generator takes the audio, reference landmarks, and pose prior landmarks as input to predict the landmarks of lip and jaw, which are then combined with pose prior landmarks to construct the target sketches. Positional encodings and modality encodings are omitted for simplicity. **(2) Landmark-To-Video Rendering** (right blue part). According to target sketches, the alignment module takes multiple reference images and their sketches as input to obtain the motion fields, which warp the reference images and their features to target head pose and expression. With the assistance of audio features and warped images and features, the translation module translates the target sketches concatenated with the lower-half masked target face to the resulted face image.

temporal positional encoding of the t-th frame is denoted as $e_t^{pos} \in \mathbb{R}^d$ which is calculated following the sinusoidal positional encoding. These encoding variables are added to the audio, pose, and reference embeddings as below,

$$\overline{a_t} = a_t + e_t^{pos} + e_a^{type} \qquad t = 1, 2, ..., T \qquad (4)$$

$$\overline{p_t} = p_t + e_t^{pos} + e_p^{type} \qquad t = 1, 2, ..., T \qquad (5)$$

$$\overline{r_i} = r_i + e_r^{type} \qquad i = 1, 2, ..., N_l \qquad (6)$$

Afterwards, we employ the Transformer modules [33] to capture both intra-type and inter-type relation among three types of embeddings. The initial tokens z^0 are formed by concatenating $\{\overline{r_i}\}_{i=1}^{N_l}$, $\{\overline{a_t}\}_{t=1}^{T}$, and $\{\overline{p_t}\}_{t=1}^{T}$. Practically, L Transformer encoder modules are adopted in our model, and each module is constituted by a stack of multi-head self-attention (MSA), layer normalization (LN), and MLP layers. The calculation process is summarized as below,

$$\begin{aligned} \overline{z}^\ell &= \text{MSA}(\text{LN}(z^{\ell-1})) + z^{\ell-1}, \quad \ell = 1 \dots L \\ z^\ell &= \text{MLP}(\text{LN}(\overline{z}^\ell)) + \overline{z}^\ell, \qquad \ell = 1 \dots L \end{aligned} \qquad (7)$$

where $z^\ell \in \mathbb{R}^{(N_l+2T) \times d}$ represents the output of the ℓ-th Transformer module. Let z_i^L be the representation of the i-th token in z^L. The second last T tokens are used to predict

lip landmarks, and the last T tokens are used to predict jaw landmarks:

$$\hat{l}_t^l = \text{MLP}(z_{t+N_l}^L) \qquad t = 1, 2, ..., T \qquad (8)$$

$$\hat{l}_t^j = \text{MLP}(z_{t+N_l+T}^L) \qquad t = 1, 2, ..., T \qquad (9)$$

where $\hat{l}_t^l \in \mathbb{R}^{2 \times n_l}$ and $\hat{l}_t^j \in \mathbb{R}^{2 \times n_j}$ denote the predicted landmarks of lip and jaw at the t-th frame, respectively.

3.1.2 Loss Function for Landmark Generation

During training, we first apply the L1 reconstruction loss L_1 to constrain the predicted landmarks:

$$L_1 = \frac{1}{T} \sum_{t=1}^{T} \left(\|\hat{l}_t^l - l_t^l\|_1 + \|\hat{l}_t^j - l_t^j\|_1 \right) \qquad (10)$$

where l_t^l and l_t^j denote the ground-truth landmarks of lip and jaw, respectively.

For sake of improving the temporal smoothness, we also adopt the following continuity regularization to constrain

the predicted landmarks,

$$L_c = \frac{1}{T-1} \sum_{t=1}^{T-1} \Big(\|(\hat{l}_{t+1}^l - \hat{l}_t^l) - (l_{t+1}^l - l_t^l)\|_2 \tag{11}$$
$$+ \|(\hat{l}_{t+1}^j - \hat{l}_t^j) - (l_{t+1}^j - l_t^j)\|_2 \Big)$$

The overall training loss for audio-to-landmark stage is defined as follows:

$$L = L_1 + \lambda_c L_c \tag{12}$$

where λ_c is a constant and is set to 1.

3.2. Landmark-To-Video Rendering

Inspired from [10], we design a rendering stage consisting of an alignment module G_a and a translation module G_r. At first, the predicted landmarks are assembled with the pose prior landmarks, forming a complete set of facial landmarks. Then we draw these landmarks on the image plane to get the target face sketches. We denote the target face sketch at the t-th frame as L_t. To achieve temporal continuity across frames, we select $2k+1$ target sketches around the t-th frame, $\{L_i \in \mathbb{R}^{3 \times H \times W}\}_{i=t-k}^{t+k}$ as inputs for predicting the t-th face image. Besides, to provide more appearance prior for the realistic rendering, multiple reference images $\{I_i^r \in \mathbb{R}^{3 \times H \times W}\}_{i=1}^N$ and their extracted sketches $\{L_i^r \in \mathbb{R}^{3 \times H \times W}\}_{i=1}^N$ are input to the alignment module for computing the motion fields $\{F_i \in \mathbb{R}^{2 \times H \times W}\}_{i=1}^N$. The reference images and their features are then warped to target head pose and expression by the motion fields. The translation module then translates the target sketches into the final face image $\hat{I}_t \in \mathbb{R}^{3 \times H \times W}$, under the guidance of auditory features, warped reference images and features, and the lower-half occluded target face.

3.2.1 Reference Images Warping

For each reference image I_i^r and its sketch L_i^r, the alignment module encodes the channel-wise concatenation of L_i^r and I_i^r with convolution layers to visual features in two spatial resolutions: $h_i^1 \in \mathbb{R}^{c_1 \times h_1 \times w_1}, h_i^2 \in \mathbb{R}^{c_2 \times h_2 \times w_2}$. To guide the prediction of motion field F_i, the channel-wise concatenation of $2k+1$ target sketches $L_{t-k:t+k}$ are injected into the alignment module by SPADE [25] layers, which modulate the visual features according to target sketches. Here, PixelShuffle layers [29] are utilized for up-sampling. Generally, the function of the alignment module can be formulated as:

$$F_i = G_a(L_i^r, I_i^r, L_{t-k:t+k}) \quad i = 1, 2, ..., N \tag{13}$$

Besides, for the aggregation of multiple reference images and corresponding visual features warped by the motion

fields, one more output layer is added to the alignment module to predict a 2D weight $w_i \in \mathbb{R}^{H \times W}$ for I_i^r. The aggregated warped image can be calculated as:

$$\overline{I^r} = \frac{\sum_{i=1}^N w_i F_i(I_i^r)}{\sum_{i=1}^N w_i} \tag{14}$$

where $F_i(I_i^r) \in \mathbb{R}^{3 \times H \times W}$ is the image warped by the motion field F_i. And the aggregated warped features in two spatial resolutions are $\overline{h}^1$ and $\overline{h}^2$, respectively, which are calculated as follows:

$$\overline{h}^s = \frac{\sum_{i=1}^N w_i F_i(h_i^s)}{\sum_{i=1}^N w_i} \quad s = 1, 2 \tag{15}$$

where $F_i(h_i^s) \in \mathbb{R}^{c_s \times h_s \times w_s}$ is the visual feature warped by the motion field F_i. Note that F_i and w_i are downsampled to match the size of h_i^s.

3.2.2 Sketch-To-Face Translation

In the translation module G_r, we aim to translate the target sketches concatenated with the masked target face I_t^m to the final face image $\hat{I}_t$. This is performed via the assistance of aggregated warped image $\overline{I^r}$ and feature $\overline{h}^1, \overline{h}^2$. Besides, to enhance the synthesized mouth detail and lip synchronization, the auditory feature $a_t \in \mathbb{R}^d$ extracted by an audio encoder similar to [26] is injected into the translation module through AdaIN layers [15]. The overall process can be formulated as:

$$\hat{I}_t = G_r(I_t^m, L_{t-k:t+k}, \overline{I^r}, \overline{h}^1, \overline{h}^2, a_t) \tag{16}$$

Specifically, I_t^m and $L_{t-k:t+k}$ are concatenated channelwise and fed into the convolution layers to obtain encoded features. $\overline{I^r}$, $\overline{h}^1$, and $\overline{h}^2$ are fused into the translation module through SPADE [25] layers to modulate the encoded features followed by AdaIN operation [15]. Meanwhile, up-sampling is implemented through PixelShuffle layers [29, 40] for better performance.

At last, the generated full face is pasted onto the original frame during inference. However, since the generated face may include small portions of background with artifacts, we composite the generated face with the background of the original frame through a Gaussian-smoothed face mask, as depicted in Figure 3.

Figure 3. Post-processing for generated face pasted onto the original frame through a Gaussian-smoothed face mask.

3.2.3 Loss Function for Rendering

Following [27], the alignment module, translation module, and audio encoder are trained jointly on the reconstruction task. The warping loss L_w, based on the perceptual loss of [18], between aggregated warped image $\overline{I^r}$ and the ground truth image I_t, is employed to constrain the alignment module for computing an accurate motion field.

$$L_w = \sum_i \left\| \phi_i(\overline{I^r}) - \phi_i(I_t) \right\|_1 \tag{17}$$

where ϕ_i is the activation output of the i-th layer in the VGG-19 network. The reconstruction loss L_r, which has a similar structure as that of L_w, and the style loss L_s, which computes the statistic error between activation output in the VGG-19 network, are applied to reduce errors between generated face $\hat{I}_t$ and ground truth I_t.

$$L_r = \sum_i \left\| \phi_i(\hat{I}_t) - \phi_i(I_t) \right\|_1 \tag{18}$$

$$L_s = \sum_i \left\| G_i^\phi(\hat{I}_t) - G_i^\phi(I_t) \right\|_1 \tag{19}$$

where G_i^ϕ is the Gram matrix derived from the aforementioned activation output ϕ_i. To further enhance the photorealism of the generated image, we also utilize patch GAN loss L_g and feature matching loss L_f from pix2pixHD [35]. To summarize, the training loss for the landmark-to-video rendering stage can be formulated as follows:

$$L = \lambda_w L_w + \lambda_r L_r + \lambda_s L_s + \lambda_g L_g + \lambda_f L_f \tag{20}$$

In experiments, we set $\lambda_w = 2.5$, $\lambda_r = 4$, $\lambda_s = 1000$, $\lambda_g = 0.25$, and $\lambda_f = 2.5$ based on cross-validation.

4. Experiments

4.1. Experimental Settings

Implementation Details. When generating landmarks from audio, we utilize the mediapipe tool [22] to detect facial landmarks from each video frame, where $n_l = 41$, $n_j = 16$, $n_r = 131$, and $n_p = 74$. Following [26], we calculate Mel-spectrograms from 16kHz audios using a window size of 800 and hop size of 200. N_l is set to 15, and d is 512. To render videos from landmarks, 128×128 face images (i.e., $H = W = 128$) are generated at 25 fps, and k is set to 2. The number of reference images N is set to 3 during training and $1/5$ of the video length during inference. More details about network architectures and hyperparameters are included in the supplementary document.

Dataset. Two audio-visual speech recognition datasets, LRS2 [1] and LRS3 [2], are used in our experiments.

1) **LRS2** [1]. The dataset consists of 48,164 video clips from outdoor shows on BBC television. Each video is accompanied by an audio corresponding to a sentence with up to 100 characters. The training, validation, and test sets are split based on broadcast date, including 45,839, 1,082, and 1,243 videos, respectively. We sample 45 videos from the test set for evaluating algorithms quantitatively.

2) **LRS3** [2]. This dataset consists of 151,819 videos from indoor shows of TED or TEDx. There exists no overlap between LRS2 and LRS3. The videos of the two datasets are distinct in shooting scenes, lighting conditions, actions, etc. Thus, we sample 45 videos from LRS3 to test the generalization ability of talking face generation methods.

Comparison Methods. We compare our method against state-of-the-art methods [6, 16, 26, 45, 46] on the person-generic audio-driven talking face generation. **EAMM** [16] generates talking face videos with emotion control based on unsupervised motion representations. **PC-AVS** [45] can generate pose-controllable talking face videos based on modularized audio-visual representation. **Wav2Lip** [26] uses an encoder-decoder model learned via adversarial training to produce talking face videos. **MakeItTalk** [46] leverages 3D landmarks to animate portrait images according to the input audio. **ATVGnet** [6] takes advantage of 2D landmarks to generate talking face videos from the input audio and an identity frame. For more comparison settings, please refer to our supplementary document.

4.2. Evaluation Metrics

Visual Quality Metrics. We use Peak Signal-to-Noise Ratio (**PSNR**) and Structured similarity (**SSIM**) [36] to measure the similarity between generated and ground-truth images. Learned Perceptual Image Patch Similarity (**LPIPS**) [43] and Fréchet Inception Distance (**FID**) [12] are employed to measure the feature-level similarity between generated and ground-truth images. We use the cosine similarity (**CSIM**) of identity vectors extracted by the face recognition network ArcFace [9] to evaluate the identity preservation ability.

Lip Synchronization Metrics. Following [37], we compute the normalized lip landmarks distance (**LipLMD**) between generated images and ground-truth images to evaluate the lip synchronization ability. When no ground-truth images exist, like in the video dubbing task, we employ SyncNet [8] to extract features from generated images and corresponding audio data and then calculate the **SyncScore** [8] between the two kinds of features. To a certain degree, this metric can convey how the lip shapes in generated images are coherent with the input audio.

Method	Dataset	Reconstruction						Dubbing
		PSNR↑	SSIM↑	LPIPS↓	FID↓	LipLMD↓	CSIM↑	SyncScore↑
ATVGnet [6]		11.55	0.3944	0.5575	223.26	0.27401	0.1020	3.51
Wav2lip [26]		27.92	0.8962	0.0741	43.46	0.02003	0.5925	3.86
MakeItTalk [46]	LRS2	17.25	0.5562	0.2237	76.57	0.05024	0.5799	2.68
PC-AVS [45]		15.75	0.4867	0.2802	110.60	0.07569	0.3927	**5.20**
EAMM [16]		15.17	0.4623	0.3398	91.95	0.15191	0.2318	3.01
Ours		**32.91**	**0.9399**	**0.0303**	**27.87**	**0.01293**	**0.6523**	4.49
ATVGnet [6]		10.90	0.3791	0.5667	190.49	0.30564	0.1176	4.43
Wav2lip [26]		28.45	0.8852	0.0683	49.60	0.02001	0.5909	4.26
MakeItTalk [46]	LRS3	17.78	0.5607	0.2788	97.99	0.08432	0.5465	3.04
PC-AVS [45]		15.60	0.4732	0.3321	115.25	0.10611	0.3537	**6.15**
EAMM [16]		15.37	0.4679	0.3812	108.83	0.17818	0.2689	3.30
Ours		**32.97**	**0.9222**	**0.0310**	**29.96**	**0.01353**	**0.6385**	5.63

Table 1. Quantitative comparison with state-of-the-art person-generic talking face generation methods on the reconstruction and dubbing setting. ↑ indicates higher is better while ↓ indicates lower is better.

ATVGnet   MakeItTalk   Wav2Lip   PC-AVS   EAMM   **Ours**   Ground Truth

Figure 4. Qualitative comparisons with state-of-the-art person-generic methods on LRS2 [1] and LRS3 [2] datasets. Our method produces realistic results with better identity preservation effect. More results are presented in the supplementary video.

4.3. Quantitative Comparison

We conduct a quantitative comparison of reconstruction and dubbing settings. In the reconstruction setting, we input the original audio to reconstruct the talking face videos. In the dubbing setting, the input audio comes from another video. Note that there is no ground truth for dubbed videos. Thus, quantitatively evaluating the generated results under the dubbing setting is challenging.

Talking Face Reconstruction. The quantitative evaluation results are reported in Table 1. It can be seen that our method achieves the best performance on all visual quality metrics. Especially in identity preservation, our method is significantly better than other methods by large margins on the CSIM metric. The CSIM of our method is 10.09% and 8.05% larger than that of the second-best method, Wav2lip, on LRS2 and LRS3, respectively. The reason is that our method can effectively leverage multiple reference images as well as landmark priors to provide more personalized facial attributes of a specific speaker, while other methods tend to generate the average facial attributes of the training dataset. Moreover, the LipLMD of our method is much

smaller than those of other methods. This means that our method can generate more accurate facial landmarks as intermediate representation and the lip movement in our results is better synchronized with the input audio.

Video Dubbing. As shown in Table 1, the SyncScore of our method is comparable to that of PC-AVS [45], which adopts specific contrastive learning strategies to pull close the visual and auditory features. Meanwhile, our method achieves a much higher SyncScore than other methods. This also validates that the lip shape generated by our method is well synchronized with the audio data.

4.4. Qualitative Comparison

Method	Image Quality	Lip Synchronization	Identity Preservation
ATVGnet [6]	1.78	2.42	2.22
Wav2lip [26]	2.74	4.14	3.68
MakeItTalk [46]	2.94	1.97	3.10
PC-AVS [45]	2.98	3.47	2.94
EAMM [16]	2.33	1.91	2.46
Ours	**4.40**	**4.56**	**4.54**

Table 2. User study about video generation quality.

User Study. To validate the effect of our method qualitatively, we conduct a user study where 25 participants are invited to evaluate the generated videos. Five videos are

selected from each dataset for evaluation. The input audio is chosen following the reconstruction setting. Each participant is asked to rate the generated video from 1 to 5 on three terms, including image quality, lip synchronization, and identity preservation. A higher score indicates a better result. The mean opinion scores (MOS) are presented in Table 2. As can be observed, our method receives a better evaluation from participants than other methods. Specifically, the MOS of our method is 47.6%, 10.1%, and 23.4% higher than that of the second best method on image quality, lip synchronization, and identity preservation, respectively.

Visualization of Generated Images. Three examples generated under the talking face reconstruction setting are presented in Fig. 4. Compared with other methods, our method can produce images that are visually closer to the ground truth. There also exist relatively less artifacts in our results. PC-AVS [45], Wav2Lip [26], MakeItTalk [46], and ATVGnet [6] fail to generate images that have consistent mouth shapes with the ground-truth images. EAMM [16] and PC-AVS tend to produce face images with moderate differences from ground truth images. For example, the eyes and wrinkles are apparently different from those in the ground truth images. Besides, the results of Wav2Lip and ATVGnet are a little blurry.

4.5. Ablation Study

In this section, we conduct ablation studies on the LRS2 [1] dataset to validate the effect of core components in our method and the performance gain derived by multiple reference images.

Method	Landmark Error
w/ LSTM	3.99
w/ Transformer	**3.04**

Table 3. Ablation study on Transformer modules.

Effectiveness of Transformer Encoder. To demonstrate the advantage of the transformer encoder used in landmark generator, we implement a variant of our method by replacing the transformer encoder with a commonly used bidirectional LSTM [13]. In this variant, reference embeddings are first averaged to form a global reference embedding as in [37]. Then, the LSTM module takes the concatenation of pose embedding, audio embedding, and global reference embedding as input to predict the landmarks. Landmark error is calculated to evaluate the accuracy of predicted landmarks (see supplementary material for more details). As shown in Table 3, our transformer-based landmark generator ('w/ Transformer') performs better than the LSTM-based landmark generator ('w/ LSTM'). This is because the transformer module is more advantageous at modeling the temporal dependencies and the relationships between landmark and audio features.

Effectiveness of Warping and Audio feature. We attempt to remove the warping mechanism in the alignment

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	LipLMD↓	CSIM↑
w/o warping	32.46	0.9377	0.0323	28.25	0.01357	0.6489
w/o audio	32.71	0.9383	0.0317	28.47	0.01334	0.6501
Ours	**32.91**	**0.9399**	**0.0303**	**27.87**	**0.01293**	**0.6523**

Table 4. Ablation study on audio features and reference image alignment in the landmark-to-video rendering model.

Reference Num	PSNR↑	SSIM↑	LPIPS↓	FID↓	CSIM↑
N=1	32.91	0.9390	0.0313	28.55	0.6485
N=5	32.96	0.9393	0.0312	28.35	0.6536
N=10	32.97	0.9394	0.0312	27.97	0.6529
N=25	**32.97**	**0.9394**	**0.0311**	27.93	**0.6546**

Table 5. Ablation study on the number of reference images.

module and inject the misaligned reference images and features into the translation module. This forms a variant of our method indicated by 'w/o warping'. As can be seen in Table 4, it derives deteriorated visual quality metrics. Besides, the LipLMD metric also gets worse because misaligned reference images have negative impacts on inferring the lip shape. Moreover, without using audio feature to enhance the lip synchronization and mouth details in the translation module (i.e. 'w/o audio'), both lip synchronization and visual quality metrics get worse.

Number of Reference Images. We try to vary the number of reference images N during inference to assess the performance gain derived by multiple reference images. The experimental results are provided in Table 5. We observe that using multiple reference images (5 to 25) gives rise to better metric values than using one reference image.

5. Conclusion

We propose a two-stage person-generic method for audio-driven talking face generation. First, we devise a novel transformer-based landmark generator to obtain accurate lip and jaw landmarks from audio. Then we align multiple reference images with the target expression and pose to provide more appearance prior for rendering face videos. Besides, acoustic features are utilized to enhance lip synchronization in the rendering stage. Extensive experiments show that our method can generate more realistic, lip-synced, and identity-preserving talking face videos than other person-generic methods.

6. Acknowledgments

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. In *arXiv:1809.02108*, 2018. 2, 6, 7, 8

[2] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. In *arXiv preprint arXiv:1809.00496*, 2018. 2, 6, 7

[3] Sandika Biswas, Sanjana Sinha, Dipanjan Das, and Brojeshwar Bhowmick. Realistic talking face animation with speech-induced head motion. In *Proceedings of the Twelfth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9, 2021. 1, 2, 3

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2

[5] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020. 1, 3

[6] 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/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019. 1, 2, 3, 6, 7, 8

[7] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017. 2

[8] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pages 251–263. Springer, 2016. 6

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 6

[10] Michail Christos Doukas, Evangelos Ververas, Viktoriia Sharmanska, and Stefanos Zafeiriou. Free-headgan: Neural talking head synthesis with explicit gaze control. *arXiv preprint arXiv:2208.02210*, 2022. 2, 5

[11] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 1, 2, 3

[12] 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. *Advances in neural information processing systems*, 30, 2017. 6

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 8

[14] Ricong Huang, Weizhi Zhong, and Guanbin Li. Audio-driven talking head generation with transformer and 3d morphable model. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7035–7039, 2022. 1

[15] 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

[16] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, 2022. 3, 6, 7, 8

[17] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021. 1, 2, 3

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[19] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1428–1436, 2019. 1

[20] 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. 1, 2, 3

[21] 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. 1, 3

[22] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 6

[23] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[24] Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. 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, 2022. 1, 3

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 5

[26] 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. 1, 3, 5, 6, 7, 8

[27] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 6

[28] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. *arXiv preprint arXiv:2207.11770*, 2022. 1, 2, 3

[29] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5

[30] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 3

[31] Jie Tang, Yiling Wu, Minglei Li, and Zhu Wang. Talking face generation based on information bottleneck and complementary representations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3443–3447, 2021. 3

[32] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, pages 716–731. Springer, 2020. 1, 2, 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 4

[34] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2531–2539, 2022. 3

[35] 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*, pages 8798–8807, 2018. 6

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[37] Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1739–1747, 2021. 1, 2, 3, 6, 8

[38] Ran Yi, Zipeng Ye, Ruoyu Fan, Yezhi Shu, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Animating portrait line drawings from a single face photo and a speech signal. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–8, 2022. 2, 3

[39] Ran Yi, Zipeng Ye, Zhiyao Sun, Juyong Zhang, Guoxin Zhang, Pengfei Wan, Hujun Bao, and Yong-Jin Liu. Predicting personalized head movement from short video and speech signal. *IEEE Transactions on Multimedia*, 2022. 2

[40] Jiutao Yue, Haofeng Li, Pengxu Wei, Guanbin Li, and Liang Lin. Robust real-world image super-resolution against adversarial attacks. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5148–5157, 2021. 5

[41] Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3867–3876, 2021. 1, 2, 3

[42] Jiangning Zhang, Liang Liu, Zhucun Xue, and Yong Liu. Apb2face: Audio-guided face reenactment with auxiliary pose and blink signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4402–4406. IEEE, 2020. 2, 3

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[44] Yuhan Zhang, Weihua He, Minglei Li, Kun Tian, Ziyang Zhang, Jie Cheng, Yaoyuan Wang, and Jianxing Liao. Meta talk: Learning to data-efficiently generate audio-driven lip-synchronized talking face with high definition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4848–4852. IEEE, 2022. 2

[45] 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/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 1, 3, 6, 7, 8

[46] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 1, 2, 3, 6, 7, 8