2020 25th International Conference on Pattern Recognition (ICPR)
Milan, Italy, Jan 10-15, 2021

Talking Face Generation via Learning Semantic and Temporal Synchronous Landmarks

Aihua Zheng[1], Feixia Zhu[1], Hao Zhu[1], Mandi Luo[2,3] and Ran He[2,3,4,*]
[1]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation,
School of Computer Science and Technology, Anhui University, Heifei, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR), CASIA, Beijing, China
[4]Center for Excellence in Brain Science and Intelligence Technology, CAS, Beijing, China
Emails: ahzheng214@ahu.edu.cn, emmazfx@163.com, haozhu96@gmail.com, luomandi2019@ia.ac.cn, rhe@nlpr.ia.ac.cn

***Abstract*—Given a speech clip and facial image, the goal of talking face generation is to synthesize a talking face video with accurate mouth synchronization and natural face motion. Recent progress has proven the effectiveness of the landmarks as the intermediate information during talking face generation. However, the large gap between audio and visual modalities makes the prediction of landmarks challenging and limits generation ability. This paper proposes a semantic and temporal synchronous landmark learning method for talking face generation. First, we propose to introduce a word detector to enforce richer semantic information. Then, we propose to preserve the temporal synchronization and consistency between landmarks and audio via the proposed temporal residual loss. Lastly, we employ a U-Net generation network with adaptive reconstruction loss to generate facial images for the predicted landmarks. Experimental results on two benchmark datasets LRW and GRID demonstrate the effectiveness of our model compared to the state-of-the-art methods of talking face generation.**

I. Introduction

Talking face generation aims to generate realistic talking face video with lip synchronization and smooth facial motion based on a given audio clip and facial image. It has gained more attention recently in both research and industrial communities due to its wild applications prospect in virtual computer games, speech comprehension, and teleconferencing, etc. With the blossom of Generative Adversarial Networks (GANs) [1], many works use this idea to improve the performace of other tasks [2], [3]. In talking face generation, GAN-based methods [4]–[9] have made great progress to talking face generation compared to the HMM-based traditional methods [10], [11].

Recently, to mitigate the cross-modal heterogeneity, Jalalifar *et al.* [6] and Chen *et al.* [8] introduce the landmarks, as the intermediate information to guide taking face generation. Note that landmarks refer to a set of coordinate points to locate the contours of key parts. These methods separate the audio-face generation into two steps, i.e., audio to landmarks prediction and landmarks to face generation. Although the two-step strategy can further improve the synthesis results of talking face, there is still challenge to predict landmarks from audio directly and limitation in generation ability.

* corresponding author

Particularly, the precision of landmark learning plays a crucial role and affects the face generation step. For example, Jalalifar *et al.* [6] and Chen *et al.* [8] learn the landmarks via applying RNN on audio clips without paying enough attention to semantic information and the mutual information between two modalities, resulting in the performance limitation of face generation. Besides, Jalaifar *et al.* [6] focus on generating talking face for specific person (President Barak Obama) without typical evaluation metrics. To address these issues, this paper focuses on landmark learning from audio to obtain more reasonable and synchronous landmarks to guide talking face generation.

Inspired by the fact that semantic information has been widely studied as high-level information to boost the machine learning tasks [12]–[14], we propose to utilize the high-level word information existing in the audio clip as the semantic supervision to compensate the conventional low-level L_1 or L_2 reconstruction supervision. By introducing a word detector, which judges whether the predicted landmarks sequence contains the word information or not, the module of landmark prediction will be optimized and express richer semantic information into predicted landmarks and make the lip motion more realistic.

In addition, temporal synchronization is a key issue to realize a smooth transition between frames. In this cross-modality task, we argue the change of adjacent audio and landmarks to be synchronized for the mutual information inter-modality correspondence. Therefore, we tend to explore the temporal residual correlation between audio and landmarks domains to better preserve the temporal synchronization. In detail, inspired by [15], we develop a MI estimator introducing a novel constraint named temporal residual loss, which is derived from the mutual information of the audio and landmarks domain.

According to the above discussion, the landmarks prediction module is jointly implemented by the word semantic information and the temporal synchronous constraint to further enhance the talking face generation. After obtaining the robust landmarks, we project them into heatmaps with a differentiable function, then a variant of U-Net [16] is leveraged to generate

978-1-7281-8808-9/20/$31.00 ©2020 IEEE

2020 25th International Conference on Pattern Recognition (ICPR) | 978-1-7281-8808-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICPR48806.2021.9412425

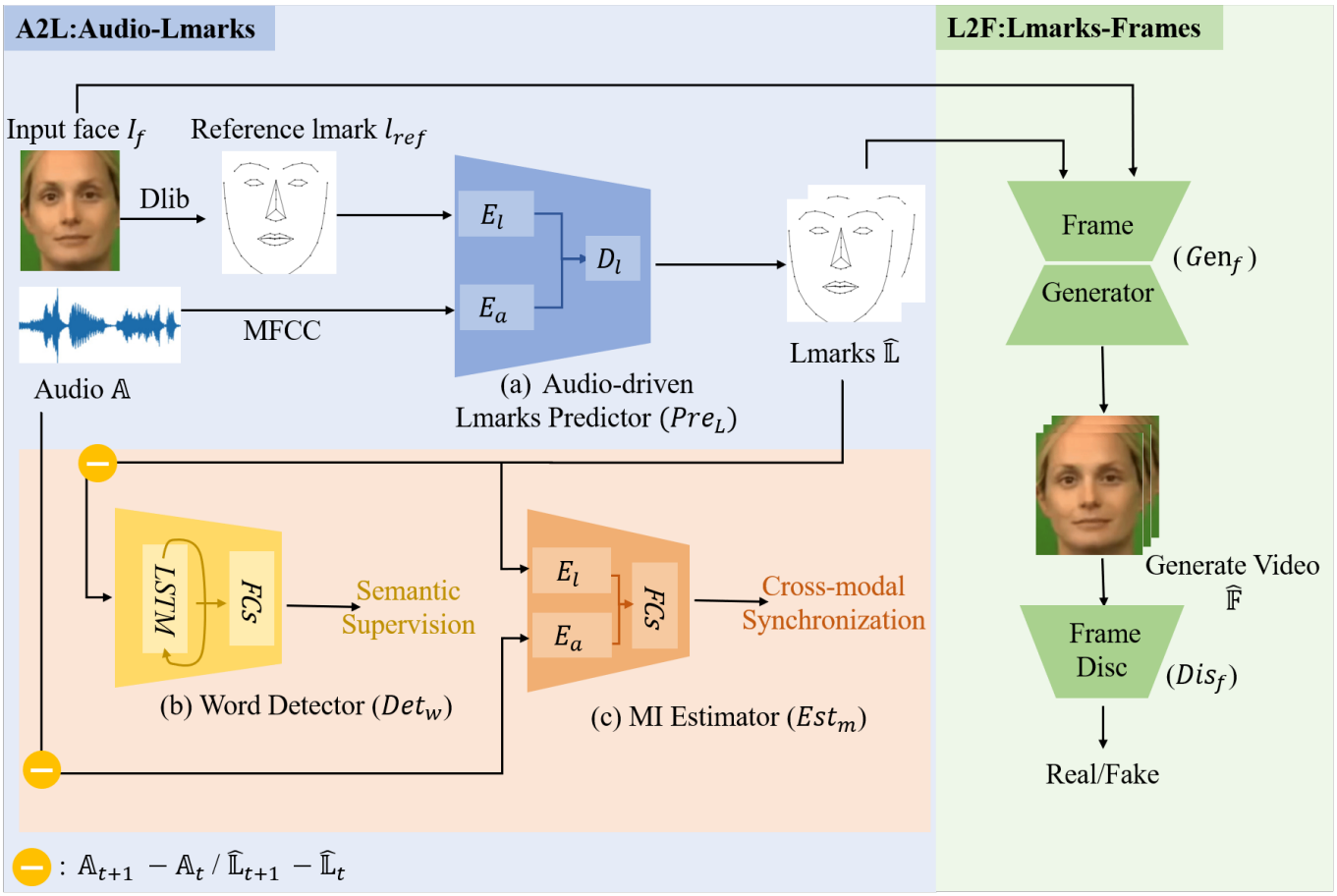


Fig. 1. The overview of our network. The blue part is the process of audio-to-landmarks (A2L). The green part is the process of landmarks-to-frames (L2F). The landmarks are more robust through the modules of Word Detector and MI Estimator which shown in orange block.

the final talking face. The main contributions of this work can be summarized as follow:

- We propose to introduce high-level semantic supervision for audio-landmark learning. Specifically, a word detector is developed to detect whether the predicted landmarks sequence contains the target word.
- We propose to enforce the temporal consistency between audio and landmarks domains with the newly introduced temporal residual loss, which is derived by estimating the mutual information between the adjacent audio and landmarks.
- Extensive experiments on two benchmark talking face datasets LRW and GRID demonstrate the effectiveness of our method, which yield promising talking face generation results comparing to the state-of-the-art methods.

II. Related Work

A. Audio-landmark Prediction

The landmarks, serving as the coordinate points marking the key parts, are widely used in many tasks such as face reconstruction [17]–[19], and human pose estimation [20]–[22]. Considering the positioning function of landmarks, researchers explore many interesting tasks, including auto-dance [23]–[26], audio-to-body [27] and audio-to-lip [28]. These tasks all belong to audio-landmark prediction which aims at predicting landmarks of body or lip from the input audio automatically. Specifically, Alemi *et al.* [23] propose to predict dance movement driven by music based on GroooveNet. Lee *et al.* [24] and Tang *et al.* [24] respectively utilize an auto-regressive encoder-decoder network and LSTM auto-encoder to achieve automatic choreography. Yalta *et al.* [26] propose a weakly supervised method based on LSTM for automatic choreography. Shlizerman *et al.* [27] propose to predict skeleton motion driven by audio of violin or piano based on LSTM. Eskimez *et al.* [28] utilize LSTM to predict lip motion correspond to audio.

B. Talking face generation

At the earliest stage, some researchers use traditional ways [10], [11], such as Hidden Markov Models (HMMs) [29] to capture the correspondence between audio and talking face sequence. Later, with the development of deep learning, works are mainly based on deep neural networks. For instance, Suwajanakorn *et al.* utilize RNN to generate talking face

video. Chung *et al.* [30] propose an encoder-decoder CNN structure to learn the embedding of audio and frames. Furthermore, recent works are fostered by the success in image generation based on GANs, such as [4], [5], [31]. Specifically, Wiles *et al.* [5] realize the generation of faces by mapping source modalities to motion space. Chen *et al.* [4] propose a correlation loss which calculates by lip optical flow and audio derivative to capture the correlation between two modalities. Vougioukas *et al.* [31] propose to utilize temporal GAN with three discriminators to jointly optimize the final generation of talking face video.

Different from the aforementioned methods which dig the relationship between audio and visual modalities directly, some works [6], [8] explore the mapping between audio and facial or lip landmarks as the bridge of talking face generation. Specifically, Jalalifar *et al.* [6] select mouth landmarks as a condition to help the face frames generation via Conditional Generative Adversarial Nets (C-GAN) [32]. Furthermore, Chen *et al.* [8] divide talking face generation into two stages, audio to facial landmark and landmark to face generation, and they focus on the latter stage via attention mechanism and regression-based discriminator. As we discussed before, the precision of landmarks makes important contributions and affects the next faces generation. However, these works only use L_1 or L_2 as reconstruction loss to guide the landmark prediction which may not obtain robust landmarks. Thus, we pay more attention to the stage of facial landmarks from audio. Different from Zhou *et al.* [7], which learn in the purely data-driven manner and propose to disentangle video into word-space and person-space via introducing word label and person id, our work introduces a word detector to detect predicted landmarks whether contain target word and capture more semantic feature when optimizing the model of audio to landmarks. Also, we propose a novel temporal residual loss, which calculates the Mutual Information (MI) between audio and landmarks residual in the timing axis, to contribute the temporal consistency between landmarks and audio clips.

III. Approaches

In this paper, we propose a word semantic supervised and cross-modal temporal synchronization guided landmark learning approach for talking face generation. As shown in Fig. 1, our network consists of two phases: 1) Audio-to-Landmarks (A2L), to predict facial landmarks by a landmark predictor via an LSTM based network supervised by the word semantic and audio-landmark cross-modal synchronization. 2) Landmarks-to-Face (L2F), to generate talking face image from the predicted sequential landmarks of A2L based on a simple variant of U-Net. We shall elaborate on these two phases in the following two sections.

A. A2L: Audio To Landmarks

To obtain landmarks from audio, we design an Audio-driven Landmarks Predictor (Pre_L) to predict landmarks by feeding an audio clip and the reference landmarks as the basic information. Considering the high correlation between the words and either audio or lip motion during the talking, we firstly utilize the words information as semantic supervision to enhance the semantic consistency in audio-landmarks pairs via a Word Detector Det_w. Second, to further emphasize the audio-landmarks cross-modal synchronization, we propose a temporal residual loss. It maximizes the mutual information between the change of adjacent audio and landmarks via an MI estimator Est_m. By jointly optimizing the word detector and the temporal residual loss, our method can predict more meaningful and synchronized landmarks, which can better guide the next landmarks to face generation.

Audio-driven Landmark Predictor (Pre_L). Audio-driven landmark predictor (Pre_L) is first fed by T segments audio clip $\mathbb{A} = \{ a_1, a_2, ..., a_T\}$, and the reference landmarks l_{ref} of the input identity face image I_f detected via Dlib [33] toolkit. The audio $\mathbb{A}$ and reference landmarks l_{ref} provide the movement and identity information respectively for the landmarks prediction. The landmark predictor Pre_L consists of three components, including Landmarks Encoder E_l, Audio Encoder E_a, and Landmarks Decoder D_l. First, we feed the Mel-scale Frequency Cepstral Coefficients (MFCC) as the audio feature to the Audio Encoder E_a, and code the reference landmarks l_{ref} via the Landmark Encoder E_l. Second, we concatenate the speech and the reference landmarks features into Landmark Decoder D_l to obtain the final sequential landmarks.

The sequential landmarks $\hat{\mathbb{L}} = \{ \hat{l}_1, \hat{l}_2, ..., \hat{l}_T \}$ prediction process can be simplified as,

$$\hat{\mathbb{L}} = D_l(E_a(A), E_l(l_ref)), \tag{1}$$

where we impose L_2 loss as the reconstruction constraint,

$$\mathcal{L}^l_{rec} = \frac{1}{T \times K} \sum_{t=0}^{T} \sum_{k=0}^{K} \parallel \hat{l}_t - l_t \parallel_2, \tag{2}$$

where $\mathbb{L} = \{ l_1, l_2, ..., l_T \}$ indicate the ground truth landmarks and K denotes the number of landmarks per face image.

Word Detector based Semantic Supervision. The words shared in both audio and lip motion can be regarded as the semantic information, to bridge the audio and visual/landmarks modalities. By taking words semantic information into consideration, the predicted landmarks will reflect closer words content to the corresponding audio. Based on this intuitive motivation, we propose to introduce a word detector Det_w to distinguish whether the predicted landmarks sequence $\hat{\mathbb{L}}$ contains the word semantic information existed in the given audio clip. In addition, we use the word information provided by the dataset as the semantic supervision to bridge the gap between the audio and landmarks, therefore to capture more meaningful and realistic landmarks.

We train the word detector on the training set of large-scale in-the-wild dataset LRW [34], containing about 500,000 videos with 1.16 seconds per video, to detect the word the in sequential landmarks. Specifically, we feed a landmark sequence of 25 frames to the detector, and output a one-hot vector representing the probability of the predicted word.

TABLE I
THE DETECTION ACCURACY OF THE WORD DETECTOR ON DATASETS LRW [34] AND GRID [35].

Datasets	LRW	GRID
Accuracy	98.24%	96.73%

The word detector is pretrained as,

$$\mathcal{L}_{Det_w} = log(Det_w(\mathbb{L}) - v_w), \tag{3}$$

where the structure of Det_w is based on LSTM and two fully connected layers v_w is a 500 dimension one-hot vector indicating the target word. We incorporate the pretrained word detector into A2L training via the following classification loss,

$$\mathcal{L}_{wd} = log(Det_w(\hat{\mathbb{L}}) - v_w). \tag{4}$$

By minimizing the Eq. (4), the predicted sequential landmarks are endowed with words semantic information and make it more reasonable to guide the next stage L2F generation.

To evaluate the performance of the word detector, we test the detection accuracy of the word detector on both LRW [34] and GRID [35] datasets. The high accuracy as reported in Table I ensures the effectiveness of word detector in A2L to guarantee the next face generation.

Temporal Residual Loss based Cross-modal Synchronization. Temporal consistency is important for the authenticity of the smooth transition between frames in sequence. Different from the existing methods that only consider the temporal consistency in the visual domain [6]–[8], we propose to enforce the cross-modal synchronization between the residual of both adjacent audio segments $(\mathbb{A}_{t+1} - \mathbb{A}_t)$ and their corresponding predicted landmarks $(\hat{\mathbb{L}}_{t+1} - \hat{\mathbb{L}}_t)$. In particular, we propose to maximize the mutual information (MI) to optimize the temporal residual loss. Motivated by the idea of [15], we first estimate the MI via a two-stream MI Estimator Est_m, based on the three convolution layers and two fully connected layers. In the same manner as the word detector, MI Estimator Est_m is pretrained on ground truth landmarks,

$$\mathcal{L}_{Est_m} = -\frac{1}{T}\sum_{t=0}^{T} Est_m[(\mathbb{A}_{t+1} - \mathbb{A}_t)), (\mathbb{L}_{t+1} - \mathbb{L}_t)]. \tag{5}$$

Then, we incorporate the MI Estimator Est_m into A2L training via the following temporal residual loss,

$$\mathcal{L}_{tr} = -\frac{1}{T}\sum_{t=0}^{T} Est_m[(\mathbb{A}_{t+1} - \mathbb{A}_t)), (\hat{\mathbb{L}}_{t+1} - \hat{\mathbb{L}}_t)]. \tag{6}$$

By maximizing the mutual information between audio and the predicted landmarks transitions, the predicted sequential landmarks can better preserve the smooth transition in cross-modal synchronization corresponding to the given audio clip.

B. L2F: Landmarks to Face

After obtaining the semantic and cross-modal temporal synchronized landmarks, we utilize GAN [1] to implement the landmark to face generation (L2F). In the L2F phase, the main components consist of Frame Generator (Gen_f) and Frame Discriminator (Dis_f). We employ a variation of U-Net as the generator G_f, which is widely used due to its promising performance in the image to image translation. To provide more spatial information during generation, we first translate the predicted landmarks into heatmaps. Then we concatenate the heatmaps with the identity face image and feed into the Frame Generator G_f to obtain the generated video $\hat{\mathbb{F}}$ = { $\hat{f}_1$, $\hat{f}_2$, ..., $\hat{f}_T$ }. We impose L_1 reconstruction loss to optimize G_f,

$$\mathcal{L}^f_{rec} = \| f_i - \hat{f}_i \|_1 . \tag{7}$$

Due to the larger variation of the lips, we superimpose an extra adaptive reconstruction loss which focuses on lip movement according to predicted lip landmarks,

$$\mathcal{L}^{lip}_{rec} = \| Crop(f_i) - Crop(\hat{f}^{lip}_i) \|_1, \tag{8}$$

where $Crop(\cdot)$ indicates the operation of cropping the lip area based on landmarks.

To distinguish the authenticity of the generated face, we employ the adversarial loss to optimize G_f and Dis_f.

$$\mathcal{L}^f_{adv} = \mathbb{E}[Dis_f(f_i)] + \mathbb{E}[\log(1 - D_f(G_f(I_f, \hat{l}_i)))], \tag{9}$$

where f_i indicates the i-th frame in the video ground truth.

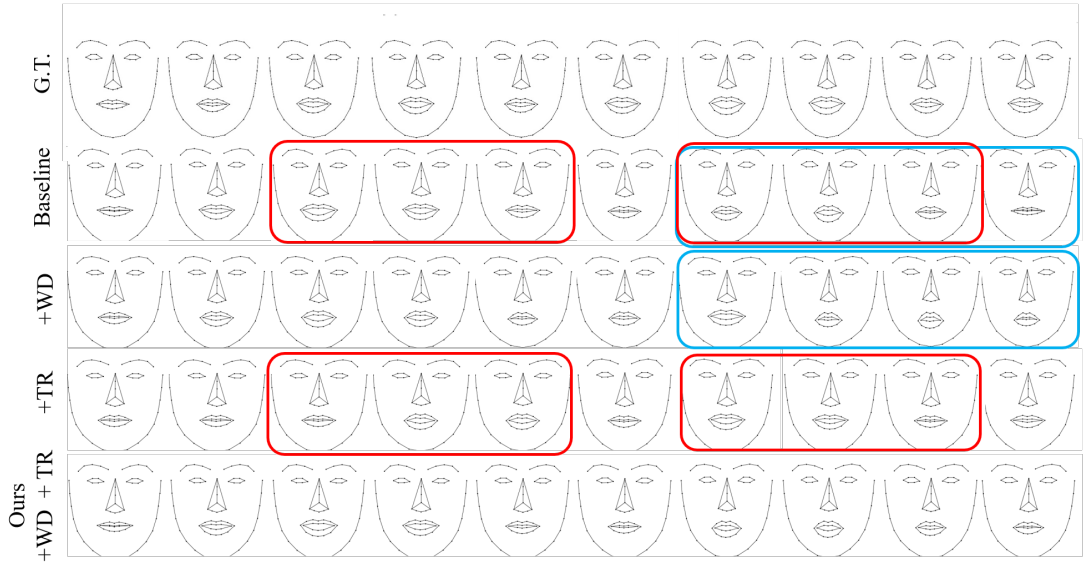
IV. EXPERIMENTS

To verify the validity of the proposed methods, we evaluate our model on two benchmark datasets LRW [34] and GRID [35] comparing to the state-of-the-art methods Chung *et al.* [30], Wiles *et al.* [5], Zhou *et al.* [7], and Chen *et al.* [8].

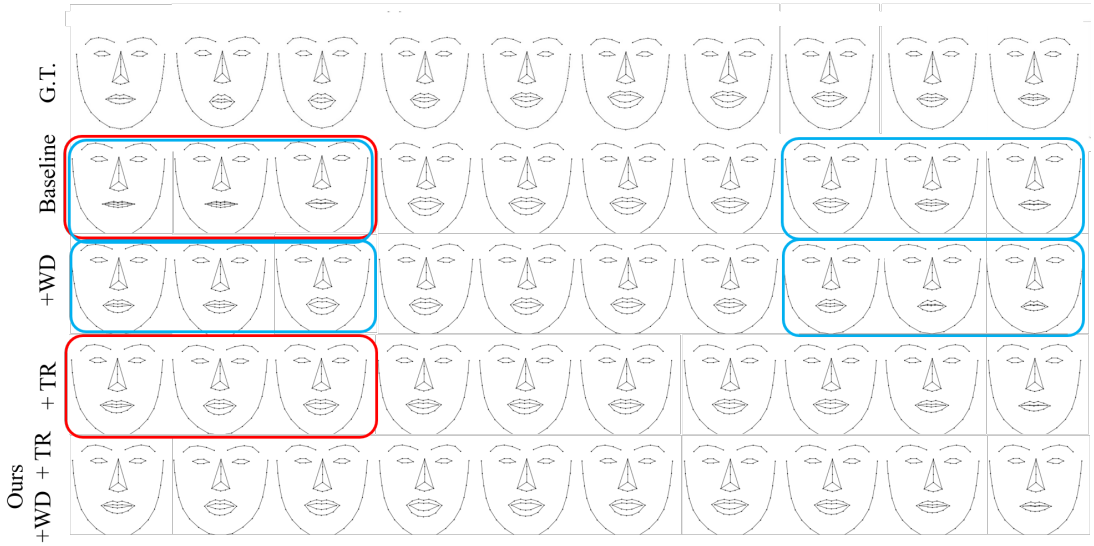
A. Dataset

a) LRW dataset: LRW dataset is a large in-the-wild dataset audio-visual lip-reading database from BBC TV broadcasts which consists of more than 1000 utterances of 500 different words. The length of each video is 29 frames and the target word is in the middle of the video. Our model is trained on the training set of the LRW dataset.

b) GRID dataset: GRID dataset contains 1000 short videos with simple and syntactically identical phrases spoken by 33 different speakers in the constrained environments. To demonstrate the generalization of our model, we use the model pretrained on the LRW dataset for the cross-dataset evaluation on the GRID dataset.



(a) The landmarks of 'BETTER'



(b) The landmarks of 'WORD'

Fig. 2. Predicted landmarks examples of our method adding word detector and temporal residual loss gradually in the test set of LRW dataset [34]. (a) Results from the word of 'BETTER' contained in the input audio clip. (b) Results from the word of 'WORD' contained in the input audio clip. The paired blue boxes represent the difference between landmarks in word semantic pronunciation and the paired red boxes represent the difference between landmarks in temporal consistency and smoothness.

B. Experimental Setup

a) Preparing: Frames are firstly extracted from raw video files and aligned by the RSA algorithm [36]. As our model can generate 256 $\times$ 256 resolution images, we then resize all the frames to this resolution. Subsequently, the Dlib [33] is used to detect facial landmarks composed of 64 key points, which act as the landmarks ground truth. For audio clips, following the operation of [8], we firstly extract MFCC at the window size of 10 ms in the audio segment extracted from the raw video and align to the center image frame and then remove the first coefficient from the original MFCC vector and finally obtain a 28 $\times$ 12 MFCC feature for each audio clip.

b) Metrics: In this work, we use the common reconstruction metrics, such as the peak signal-to-noise ratio (PSNR) and the structural assessment (SSIM) index to evaluate the generated videos. For PSNR and SSIM, a larger score corresponds to more realistic generated results. In addition, to evaluate the quality of the predicted facial landmarks from the audio clip. we employ the Landmark Distance (LMD) [4] to calculate the Euclidean distance between the pseudo facial landmark labels and the generated landmarks from audio.

Fig. 3. The examples of generated talking faces from test set of LRW dataset [34] comparing with Zhou *et al.* [7] and Chen *et al.* [8]. The results show that our model can better synchronize with the ground truth. More visualized demonstration please refer to the demo video[1].

C. Evaluation on LRW

We evaluate our model on the LRW dataset to verify the effectiveness, as shown in Table II and Fig. 3. Firstly, as reported in Table II: 1) Our method achieves the best scores on LMD and SSIM metrics, which verifies the effectiveness of our landmark learning. 2) The value of PSNR metric of ours is slightly overshadowed than Chen *et al.* [8]. However, our method significantly outperforms Chen *et al.* [8] on LMD and SSIM, which keeps a better balance on the three metrics. Then, Fig. 3 demonstrates two examples of the generated talking faces comparing to the most recent methods Chen *et al.* [8] and Zhou *et al.* [7] which architectures are based on Audio-Landmark-Face and Audio-Face respectively. It is clear that the generated facial frames in Chen *et al.* and Zhou *et al.* [7] are not well synchronized with the ground truth especially in the shape of lip motion. While in the second example, the results of the Zhou *et al.* [7] tend to have blurry lips due to the heterogeneous gap between audio and visual modality. Note that Chen *et al.* [8] can only generate 128×128 resolution images which are much blurrier than our 256×256 resolution images. In contrast, our approach can generate more synchronized and realistic results comparing to the ground truth, suggesting the effectiveness of the proposed landmark learning approach. More visualized demonstration is provided in the video[1].

TABLE II
QUANTITATIVE COMPARISON RESULTS AGAINST STATE-OF-THE-ART METHODS ON LRW DATASET [34].

Method	Evaulation on LRW		
	LMD	PSNR	SSIM
Chung *et al.* [30]	1.35	29.36	0.74
Zhou *et al.* [7]	–	26.80	0.88
Wiles *et al.* [5]	1.60	29.82	0.75
Chen *et al.* [8]	1.37	**30.27**	0.78
Ours	**1.09**	30.15	**0.90**

TABLE III
QUANTITATIVE COMPARISON AGAINST STATE-OF-THE-ART METHODS ON GRID DATASET [35]. NOTE THAT OUR METHOD IS TRAINED ON LRW DATASET [34].

Method	Evaulation on GRID		
	LMD	PSNR	SSIM
Chung *et al.* [30]	1.44	29.87	0.76
Wiles *et al.* [5]	1.48	29.39	0.80
Chen *et al.* [8]	1.29	**32.15**	0.83
Ours	**0.88**	31.50	**0.96**

D. Cross-dataset Evaluation on GRID

The GRID dataset is constrained in a lab-controlled environment, which is easy for a model to fit on the training set and produce high-quality results. Therefore, we verify the effectiveness of our method on GRID with the model trained on LRW. Table III reports the cross-dataset evaluation results comparing with the state-of-the-art methods, which are directly trained on GRID. As shown in Table III, our model achieves the best scores on LMD and SSIM metrics, which ensures the generalization of our model. Note that our model is slightly overshadowed by Chen *et al.* on PSNR, but still competitive while training on the LRW dataset.

E. Ablation Study

To evaluate the contribution of each component, we further conduct an ablation study on the Word detector and temporal residual loss. As reported in Table IV: 1) By progressively introducing the Word Detector (WD) and the Temporal Residual Loss (TR), the results are constantly improved on all the metrics during both A2L and L2F stages. 2) Introducing the word detector has slightly reduced the SSIM, but can significantly improve the LMD. The main reason is that inaccurate landmarks may not affect the sharpness of the generated faces.

[1]https://drive.google.com/file/d/1y201oPzd5g8b8emyDF5jt3kLsG8UlP2O

TABLE IV
ABLATION STUDY ON TWO KEY COMPONENTS OF THE PROPOSED METHOD, WORD DETECTOR (WD), AND TEMPORAL RESIDUAL LOSS (TR) ON LRW DATASET [34].

	A2L	L2F		
	LMD	LMD	PSNR	SSIM
Baseline	3.43	1.41	29.43	0.83
+ WD	3.08	1.16	29.61	0.82
+ TR	3.27	1.28	30.04	0.87
+ WD + TR (Ours)	**3.03**	**1.09**	**30.15**	**0.90**

3) Introducing the temporal residual loss consistently improves the performance on all the three metrics.

We observe that the LMD of A2L is larger than it of L2F. The explanation is that LMD of L2F is calculated between landmarks detected in ground truth face image and generated image by Dlib. The Dlib has a prior on face image that small deformation will be ignored, while the Audio2Landmark predictor do not share the same prior.

Fig. 2 presents two examples of the predicted landmarks while speaking "BETTER" and "WORD" to further verify the effectiveness. Compared with baseline, the landmarks highlighted by the blue boxes of '+WD' (Word Detector) present more synchronously with word semantic pronunciation. The landmarks highlighted by the red boxes of '+TR' (Temporal Residual Loss) present more temporal consistency. After jointly introducing 'TR' and the '+ WD', the sequential landmarks of 'Ours' are more synchronously with the word pronunciation and transit smoother on lip motion which can refer to the ground truth.

F. Landmarks Analysis

To evaluate the influence of the number of reference landmarks in our model, we train our method on five different settings of reference landmarks as shown in Table V (a) by removing part of the landmarks in corresponding settings. As reported in Table V, our method is robust to the landmarks of the jaw, eyes and nose while more sensitive to the landmarks in the lip area, which contains more crucial information for talking face generation.

Furthermore, we randomly select 1430 frontal and 1430 profile face videos from test set of LRW dataset to explore the influence the reference landmarks in terms of face poses as shown in Table V (b). Generally speaking, "profile pose" achieves comparable performance to "frontal face", which evidences the robustness of our model with different face poses. "frontal face" outperforms the "profile pose" on all metrics due to more precise landmarks (less LMD) which can further enforce the L2F phase for better talking face generation.

V. CONCLUSION

In this paper, we have presented a novel method via focusing on learning robust landmarks from audio to better guide the talking face generation. For obtaining robust landmarks to guarantee the quality of generated talking face, we consider the semantic information contained in landmarks and cross-modal temporal synchronization between the change of adjacent audio and landmarks. The former implements via a word detector to capture richer semantic information in sequential landmarks that depend on words semantic supervision and the latter is achieved by learning synchronous relationship between temporal residual landmarks via a mutual information estimator. Experimental results on benchmark datasets validate the effectiveness of our contributions.

TABLE V
EXPERIMENTS ON DIFFERENT SETTINGS OF REFERENCE LANDMARK EVALUATED ON LRW DATASET [34].

	Settings (Predicted Lmarks)	A2L	L2F		
		LMD	LMD	PSNR	SSIM
(a)	w/o eyes & nose (47)	3.09	**1.12**	**30.24**	**0.89**
	w/o jaw (51)	3.13	1.13	29.97	0.87
	w/o part of lip (58)	3.31	1.15	29.71	0.87
	w/o up lip (60)	**3.01**	1.86	28.33	0.74
	w/o right lip (60)	3.05	2.03	28.16	0.71
(b)	frontal only (68)	**2.92**	**1.02**	**30.27**	**0.93**
	profile only (68)	3.19	1.21	30.01	0.87
	frontal + profile (68)	3.06	1.12	30.14	0.90

ACKNOWLEDGMENT

This research is supporte d in part by the National Natural Science Foundation of China (61976002), the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2019A0033), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (201900046).

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[2] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 637–647, 2017.

[3] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, "Learning a high fidelity pose invariant model for high-resolution face frontalization," in *Advances in Neural Information Processing Systems*, 2018, pp. 2867–2877.

[4] L. Chen, Z. Li, R. K Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance," in *European Conference on Computer Vision*, 2018, pp. 520–535.

[5] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *European Conference on Computer Vision*, 2018, pp. 670–686.

[6] S. A. Jalalifar, H. Hasani, and H. Aghajan, "Speech-driven facial reenactment using conditional generative adversarial networks," *arXiv preprint arXiv:1803.07461*, 2018.

[7] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *AAAI Conference on Artificial Intelligence*, 2019, pp. 9299–9306.

[8] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generationwith dynamic pixel-wise loss," *arXiv preprint arXiv:1905.03820*, 2019.

[9] H. Zhu, A. Zheng, H. Huang, and R. He, "High-resolution talking face generation via mutual information approximation," *arXiv preprint arXiv:1812.06589*, 2018.

[10] A. Simons, "Generation of mouthshape for a synthetic talking head," *Proc. of the Institute of Acoustics*, 1990.
[11] L. Xie and Z.-Q. Liu, "A coupled hmm approach to video-realistic speech animation," *Pattern Recognition*, vol. 40, no. 8, pp. 2325–2340, 2007.
[12] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1253–1260.
[13] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8798–8807.
[14] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan, "Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation," in *Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2417–2426.
[15] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.
[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2015, pp. 234–241.
[17] F. Liu, D. Zeng, Q. Zhao, and X. Liu, "Joint face alignment and 3d face reconstruction," in *European Conference on Computer Vision*, 2016, pp. 545–560.
[18] J. K. Pontes, C. Kong, S. Sridharan, S. Lucey, A. Eriksson, and C. Fookes, "Image2mesh: A learning framework for single image 3d reconstruction," in *Asian Conference on Computer Vision*, 2018, pp. 365–381.
[19] X. Yuan and I. K. Park, "Face de-occlusion using 3d morphable model and generative adversarial network," *arXiv preprint arXiv:1904.06109*, 2019.
[20] A. Martínez-González, M. Villamizar, O. Canévet, and J.-M. Odobez, "Real-time convolutional networks for depth-based human pose estimation," in *International Conference on Intelligent Robots and Systems*, 2018, pp. 41–47.
[21] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2500–2509.
[22] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in *International Conference on Computer Vision*, 2017, pp. 3706–3714.
[23] O. Alemi, J. Françoise, and P. Pasquier, "Groovenet: Real-time music-driven dance movement generation using artificial neural networks," *networks*, vol. 8, no. 17, p. 26, 2017.
[24] J. Lee, S. Kim, and K. Lee, "Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network," *arXiv preprint arXiv:1811.00818*, 2018.
[25] T. Tang, J. Jia, and H. Mao, "Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis," in *ACM Multimedia Conference on Multimedia Conference*, 2018, pp. 1598–1606.
[26] N. Yalta, S. Watanabe, K. Nakadai, and T. Ogata, "Weakly-supervised deep recurrent neural networks for basic dance step generation," in *International Joint Conference on Neural Networks*, 2019, pp. 1–8.
[27] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman, "Audio to body dynamics," in *Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7574–7583.
[28] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," in *International Conference on Latent Variable Analysis and Signal Separation*, 2018, pp. 372–381.
[29] L. R. Rabiner and B.-H. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.
[30] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" *arXiv preprint arXiv:1705.02966*, 2017.
[31] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with gans," *arXiv preprint arXiv:1906.06337*, 2019.
[32] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.
[33] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.
[34] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016, pp. 87–103.
[35] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.
[36] Y. Liu, H. Li, J. Yan, F. Wei, X. Wang, and X. Tang, "Recurrent scale approximation for object detection in cnn," in *International Conference on Computer Vision*, 2017, pp. 571–579.