A Supplementary Material

We show more interpretations and experiments in the supplementary material. In Section A.1, we introduce the input details of the video stream and audio stream in the model. In Section A.2, we verify the effectiveness of the AU classifier. In Section A.3, we show more visualization results.

A.1 Data Pre-processing

As shown in Fig. 1, each recording in the GRID [4] dataset and TCD-TIMIT dataset [5] is processed into a video stream and an audio stream. The video stream consists of consecutive cropped video frames. The audio stream consists of consecutive 12-dimensional MFCC features. We use the first frame of each video as the input identity information. For audio input, we slide a fixed window containing 280ms audio information on the audio stream in Fig. 1. The corresponding video frame in the middle of the audio sliding window is the ground truth. So we compare the image generated by each audio frame with the corresponding ground truth.

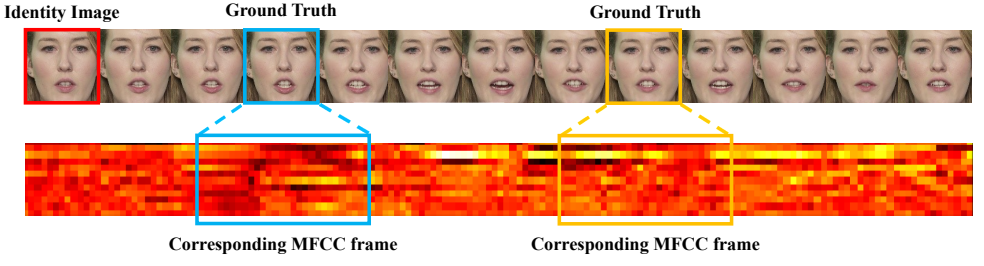

Figure 1: The input details of video stream and audio stream in the model.

A.2 AU Detection with the Pre-trained AU Classifier

In order to enable the AU classifier to judge whether the generated image contains the correct AU information, we use the GRID dataset to pre-train it. The training data are the images and corresponding speech-related AU labels. We use Openface [1] to extract the speech-related AU labels of each image.

AU Number	Training Set		Test Set	
	F1 score	Accuracy	F1 score	Accuracy
AU10	0.971	96.81%	0.875	88.60%
AU14	0.896	95.54%	0.752	91.07%
AU20	0.955	95.13%	0.896	89.85%
AU25	0.983	97.16%	0.979	96.38%
AU26	0.920	91.99%	0.854	82.18%
Average	0.945	95.33%	0.871	89.62%

Table 1: Results of our pre-trained AU classifier on real images of the GRID training set and test set.

method	GRID		TCD-TIMIT	
	Avg. F1 score	Avg. Accuracy	Avg. F1 score	Avg. Accuracy
AU classifier	0.746	80.77%	0.507	88.56%
Openface[1]	0.751	80.92%	0.590	84.92%

Table 2: The AU detection results of the images generated by our proposed method on AU classifier and Openface respectively. Because the ground truth AU labels of the real images are extracted by Openface, the Openface detection results are accurate.

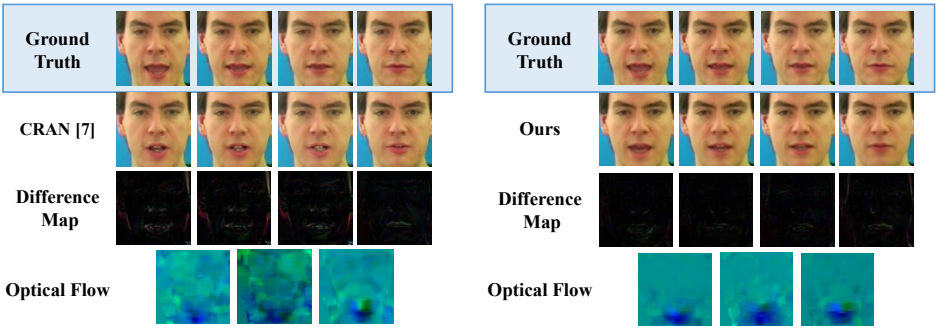

Figure 2: Difference map and optical flow of the images generated by CRAN [7] and our proposed method.

The results of our pre-trained AU classifier are shown in Table 1. The average F1 score of AUs is 0.945 and the average accuracy of AUs is 95.33% on the training set. This proves that our AU classifier can effectively capture the AU information. On the test set, it achieves 0.871 on average F1 score and 89.62% on average accuracy, which shows that our pre-trained AU classifier has strong generalization ability.

To further verify the effectiveness of the AU classifier, we use it to detect the speech-related AUs of the images generated by our proposed method. The detection results are shown in Table 2. Because the ground truth AU labels are extracted by Openface [1], the Openface detection results are accurate. Since the AU classifier is pre-trained on the GRID dataset, when experimenting on the TCD-TIMIT dataset, we need to use the new dataset to refine it to adapt to the new domain. The detection results of the AU classifier on the GRID test set are very close to that of Openface. On the TCD-TIMIT data set, the AU classifier detection results on average F1 score is lower than Openface detection, but accuracy is higher. The main reason is that the distribution of AUs on the TCD-TIMIT dataset is unbalanced, so the prediction of unbalanced AU tends to be absent. Therefore, the wrong results may be predicted as correct by the AU classifier, resulting in lower F1 score and higher accuracy. In general, the AU classifier can judge whether the generated image contains the correct AU information.

A.3 More Visualization Results

To compare the difference between the generated image and the ground truth image more clearly, we show the difference map in Fig. 2. There are obvious differences in the mouth and the edge of the face in the difference map of Song et al. [7]. The difference map of ours only has a small difference in the mouth region. We also use optical flow to represents the

motion between the generated frames. The video generated by Song et al. [7] is accompanied by jitter, so the optical flow is chaotic. The changes of our video frames are concentrated in the mouth, so the optical flow is very clear.

Figure 3: Example of generated videos driven by silent audio with no one speaking and only slight noise. The images generated by our method is consistent with the input image, which obviously suppresses the lip movement.

Figure 4: More example results produced using our proposed model on the TCD-TIMIT test set. We use the same audio clip that corresponds to the word "wash" for different speakers to generate talking head videos.

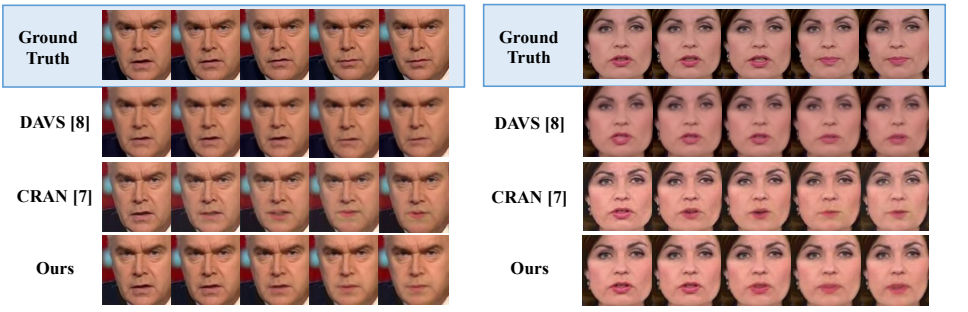

Figure 5: Example of generated frames produced by our proposed model and other methods on the LRW dataset. DAVS [8] is trained on the LRW dataset. CRAN [7] and our model are trained on the GRID dataset.

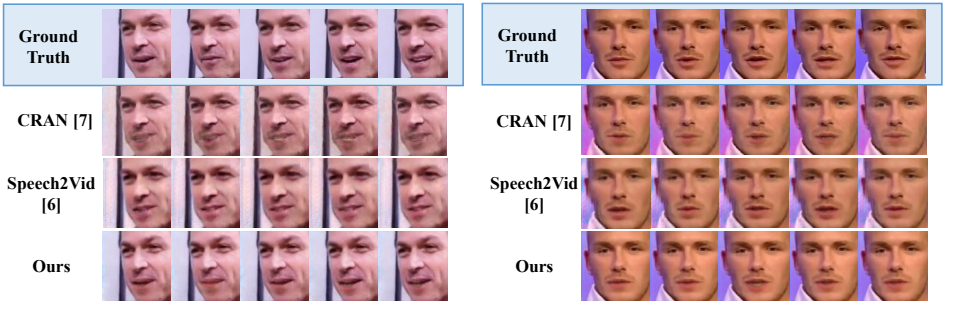

Figure 6: Example of generated frames produced by our proposed model and other methods on the VoxCeleb2 dataset. All of the methods are trained on the GRID dataset.

When no one is speaking, the mouth should not perform any movement and be consistent with the input. Therefore, we test the effect of our model without anyone speaking. The audio clip only has slight noise. Fig. 3 shows the generation results of three different subjects using silent audio. The speakers are taken from the GRID test set. The mouth of the first subject is closed in the input image, and the last two subjects are open. We can clearly see that whether the mouth of the input is closed or open, the video generated by our method is consistent with the input image. However, the video generated by Song et al. [7] can not suppress mouth movement, as shown in the red box.

We also show the results of generated videos for different speakers using the same audio in Fig. 4. The speakers are taken from the TCD-TIMIT test set. The content of the audio clip is the word "wash". We can see that the mouth movements generated by our model match the word very well.

To test the generation effect of our proposed model when cross-database, we conduct experiments on the LRW dataset [2]. The example of generated frames can be seen in Fig. 5. Our proposed model and CRAN [7] are trained on the GRID dataset, and DAVS [8] are trained on the LRW dataset. It can be seen that the images generated by DAVS are blurry, and the face texture of the images generated by CRAN are not well preserved. Our method can still generate high-quality talking face images when cross-database.

We also try to use our model trained on the GRID dataset to experiment on the VoxCeleb2 dataset [3]. The example of generated results can be seen in Fig. 6. All of the

methods are pre-trained on the GRID dataset. Since most of the videos in the VoxCeleb2 dataset have large head movements, they are not suitable for the model trained on the GRID dataset. Therefore, we only choose the examples of the faces with small head movement in the VoxCeleb2 test set. From the Fig. 6 we can clearly see that the mouths generated by song et al. [7] are very blurry, and the image quality generated by Jamaludin et al. [6] is very low, especially the background. The images generated by our method are superior to others in both image quality and lip-sync accuracy.

References

[1] Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[2] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103, 2016.

[3] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1086–1090, 2018.

[4] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[5] Naomi Harte and Eoin Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015.

[6] Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision*, 127(11):1767–1779, 2019.

[7] Yang Song, Jingwen Zhu, Dawei Li, Andy Wang, and Hairong Qi. Talking face generation by conditional recurrent adversarial network. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 919–925, 2019.

[8] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 9299–9306, 2019.