MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation

Kaisiyuan Wang[1], Qianyi Wu[1], Linsen Song[1,3,4], Zhuoqian Yang[2], Wayne Wu[1(✉)], Chen Qian[1], Ran He[3,4], Yu Qiao[5], and Chen Change Loy[6]

[1] SenseTime Research, Beijing, China
wuwenyan@sensetime.com
[2] Robotics Institute, Carnegie Mellon University, Pittsburgh, USA
[3] Center for Research on Intelligent Perception and Computing, CASIA, Beijing, China
[4] University of Chinese Academy of Sciences, Beijing, China
[5] Shenzhen Institutes of Advanced Technology, Chinese Academy of Science, Shenzhen, China
[6] Nanyang Technological University, Singapore, Singapore

Abstract. The synthesis of natural emotional reactions is an essential criterion in vivid talking-face video generation. This criterion is nevertheless seldom taken into consideration in previous works due to the absence of a large-scale, high-quality emotional audio-visual dataset. To address this issue, we build the **M**ulti-view **E**motional **A**udio-visual **D**ataset (MEAD), a talking-face video corpus featuring 60 actors and actresses talking with eight different emotions at three different intensity levels. High-quality audio-visual clips are captured at seven different view angles in a strictly-controlled environment. Together with the dataset, we release an emotional talking-face generation baseline that enables the manipulation of both emotion and its intensity. Our dataset could benefit a number of different research fields including conditional generation, cross-modal understanding and expression recognition. Code, model and data are publicly available on our project page ‡‡https://wywu.github.io/projects/MEAD/MEAD.html.

Keywords: Video generation · Generative adversarial networks · Representation disentanglement

1 Introduction

Talking face generation is the task of synthesizing a video of a talking face conditioned on both the identity of the speaker (given by a single still image) and the

K. Wang, Q. Wu, L. Song—Equal contribution.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58589-1_42) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12366, pp. 700–717, 2020.
https://doi.org/10.1007/978-3-030-58589-1_42

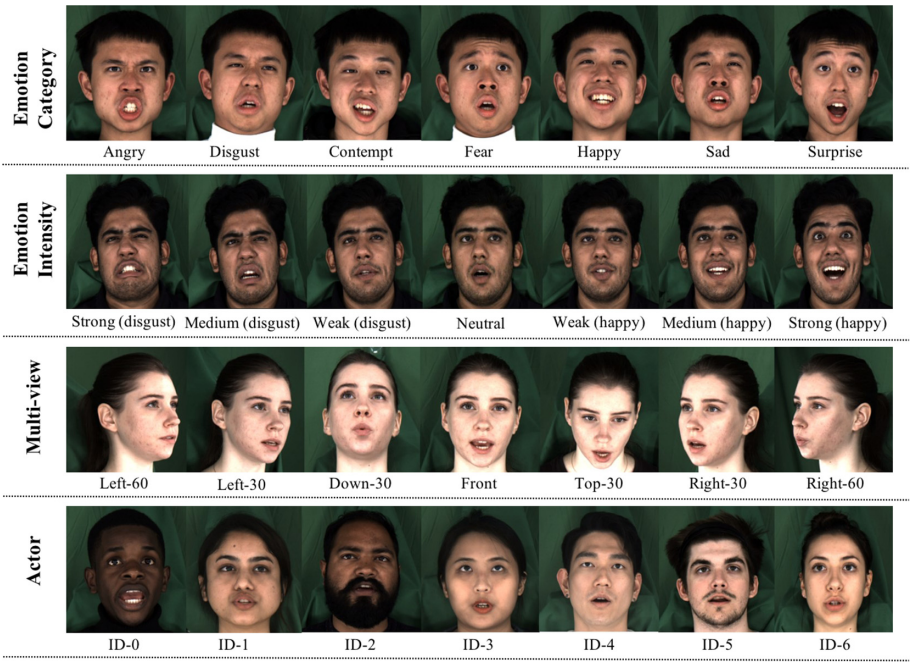

Fig. 1. MEAD overview. MEAD is a large-scale, high-quality audio-visual dataset with rich affective data, diversified speakers and multiple perspectives.

content of the speech (provided as an audio track). A major challenge in this task is constituted by the fact that natural human speech is often accompanied by several nonverbal characteristics, e.g. intonations, eye contact and facial expressions, which reflect the emotion of the speaker [11,54]. State-of-the-art methods are able to generate lip movements in perfect synchronization with the audio speech [6,58], but the faces in such videos are often emotionless and sometimes even impassive. Either the faces are devoid of any emotion or there is a distinct mismatch between the facial expression and the content of the audio speech.

A considerable number of recent advancements in the task of talking-face generation are deep learning based methods [6,52,57,58], where data has a significant influence on performance. We argue that the absence of a large-scale, high-quality emotional audio-visual dataset is the main obstacle to achieve vivid talking-face generation. As shown in Table 1, the available datasets are very limited in the diversity of the speakers, the duration of the videos, the number of the view-angles and the richness of the emotions. To address this issue, we build the **M**ulti-view **E**motional **A**udio-visual **D**ataset (MEAD), a talking-face video corpus featuring 60 actors talking with eight different emotions at three different intensity levels (except for neutral). The videos are simultaneously recorded at seven different perspectives in a strictly-controlled environment to provide high-quality details of facial expressions. About 40 hours of audio-visual clips are recorded for each person and view.

Table 1. Comparison of datasets. We compare with recent high-quality audio-visual datasets that are built in controlled environments. The symbol "#" indicates the number. The "Avg Dura" means average audio/video duration time per speaker.

Dataset	Avg Dura	#Actor	#Emo	#View	#Intensity	Resolution	#Clips
SAVESS [23]	7 min 21 s	4	7	1	1	1280 × 1024	480
RAVDESS [32]	3 min 42 s	24	8	1	2	1920 × 1080	7,356
GRID [9]	18 min 54 s	34	-	1	-	720 × 576	34,000
Lombard [1]	4 min 1 s	54	-	2	-	854 × 480	10,800
CREMA-D [4]	N/A	**91**	6	1	3(only 1/12)	1280 × 720	7442
Ours	**38 min 57 s**	60	**8**	**7**	**3**	**1920 × 1080**	**281,400**

A fundamental design choice for emotional talking-face corpus is the choice between (i) the in-the-wild approach [7,39], i.e., annotating videos gathered from sources such as the Internet and (ii) the controlled approach [29,33], i.e., recording coordinated performers in a constant, controlled environment. It is easy to scale up with the in-the-wild approach, but the data suffer from inconsistency in both the quality of the audio/video and the annotation of emotions [35]. The controlled approach, on the other hand, ensures the quality of the data but takes considerably higher costs to build. *MEAD* is an effort to build a dataset that is at the same time abundant in quantity and consistently good in quality. As far as we know, our dataset is the largest controlled dataset in terms of the number of video clips. In order to ensure the naturalness of the performed emotions, our data acquisition pipeline is carefully designed from the selection of actors/actresses to the performance feedback and correction. A team led by a professional actor guide the participants to speak in natural and accurate emotional status. To ensure the quality of the audio, we carefully select emotionally consistent speech texts that cover different phonemes.

Together with the dataset, we propose an emotional talking face generation baseline that enables the manipulation of the emotion and its intensity. A two-branch architecture is designed to process the audio and emotional conditions separately. Specifically, one of the branches is responsible for mapping audio to lip movements and the other branch is responsible for synthesizing the desired emotion on the target face. Finally, the intermediate representations are fused in a refinement network to render the emotional talking-face video.

In summary, our contributions are twofold:

- We build a large-scale, high-quality emotional audio-visual dataset *MEAD*, which is the largest emotional audio-visual corpus in terms of the number of video clips and viewpoints.
- Together with the dataset, we propose an emotional talking face generation baseline that enables the manipulation of the emotion and its intensity. Extensive experiments measure video generation and emotion manipulation performance for future reference.

2 Related Work

Talking-Face Generation. Talking-face generation is a long-standing problem [14,31,36] which is gaining attention [15,49,51]. Researchers' current focus is mainly on improving the realisticness of the generated videos. Chung *et al.*[8] proposes the Speech2Vid model to animate static target faces. Zhou *et al.* [58] adopted a representation disentanglement framework to drive different identities to utter the same speech contents. Chen *et al.* [6] used a cascade GAN approach to improve the temporal continuity of the generation. Song *et al.* [48] propose an audio-driven talking face generation method to solve head pose and identity challenges by utilizing 3D face model. However, how to manipulate the *emotion* in generated talking-face is still an open question.

Emotion Conditioned Generation. Emotion conditioned image generation has been advancing under the inspiration of recent progress in unsupervised image translation [22,30,43,55,60]. These frameworks are able to transfer expression according to specified emotion categories. However, obvious artifacts are frequently observed in dynamic changing areas in the results. Pumarola *et al.* [42] propose the GANimation model, which is based on an unsupervised framework to describe expressions in a continuous rather than discrete way, representing facial expressions as action units activations. Ding *et al.* [12] designed a novel controller module in an encoder-decoder network to adjust expression intensity continuously, however, the method is not explicit enough for more fine-grained control. Several works have also studied the generation of emotional talking sequences as well. Najmeh *et al.* [46] introduces a conditional sequential GAN model to learn the relationship between emotion and speech content, and generate expressive and naturalistic lip movements. Konstantinos [52] uses three discriminators to enhance details, synchronization, and realistic expressions like blinks and eyebrow movements. Both of the methods could basically capture the facial movements related to emotion categories, however, the emotion manipulation is completely determined by the speech audio, and cannot be more delicate to achieve manipulation in different intensities. Although some works [5,13,25,59] have proposed thought-provoking solutions towards this problem from a 3D facial animation perspective, the lack of suitable emotional audio-visual dataset still hinders further progress.

Emotional Audio-visual Dataset. There are some high-quality in-the-lab audio-visual datasets [1,3,9], but none of these take emotion information into consideration in design. The SAVEE [23] dataset is one of the datasets that considers emotion in speech. But only 4 actors are featured to read the designed TIMIT corpus [16]. Some datasets annotated not just emotion categories but also the intensities. AffectNet [38] and SEWA [28] included continuous intensity annotations based on dimensional model valence and arousal circumflex [45]. There are also datasets with discrete emotion intensity annotations. CREMA-D dataset [4] contains affective data with three intensity levels. Actors are required to express each emotion in two intensities when collecting data for RAVESS

dataset [32]. However, the limited number of recorded sentences makes it hard for networks trained on it to generalize to real-life applications.

3 MEAD

In order to ensure the naturalness of the performed emotions, our data acquisition pipeline is carefully designed from the selection of actor/actresses to the performance feedback and correction. A team led by a professional actor guide the participants to speak in natural and accurate emotional status. To ensure the quality of the audio, we carefully select emotionally consistent speech texts that cover different phonemes.

3.1 Design Criteria

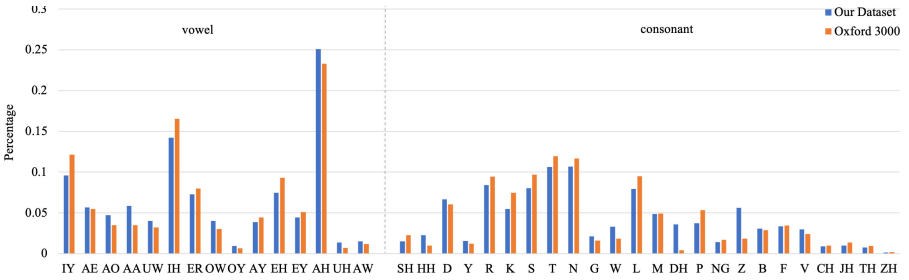

Fig. 2. Vowel and consonant distribution. Although we design different corpora for different emotions, for simplicity, we demonstrate the distribution for all emotion categories. The distribution of our corpus is basically consistent with that of frequently used 3000 words in [10].

Definition of Emotion Categories and Intensities. We use eight emotion categories following [32] and three levels of emotion intensity which is intuitive to human understanding. More intensity levels would be hard to distinguish and cause confusion and inconsistency in data acquisition. The first level is defined as *weak*, which describes delicate but detectable facial movements. The second level *medium* is the normal state of the emotion, which stands for the typical expression of the emotion. The third level is defined as *strong*, which describes the most exaggerated expressions of this emotion, requiring intense movements in related facial areas.

Design of the Speech Corpus. For audio speech content, we follow the phonetically-diverse speech corpus TIMIT [16], which is originally designed in automatic speech recognition [19,40,50]. We carefully select the sentences covering all phonemes in each emotion category, and the sentences in each emotion category is divided into three parts: 3 common sentences, 7 emotion-related sentences, and 20 generic sentences. We provide more details of the speech corpus in supplementary materials.

3.2 Data Acquisition

In the data acquisition process we mainly consider the following two aspects. First, the captured emotion should be natural and stable during talking. Second, the three levels of emotion intensities should be distinguishable. A guidance team led by a professional actor oversees the process.

Speaker Selection. We recruit fluent English speakers aged from 20 to 35 with previous acting experience. To evaluate the skills of the actors, video samples of each emotion in different intensities performed by the professional actor is recorded and the candidates are required to imitate the expressions in the videos. The guidance team evaluated the performance of each speaker according to the imitation result to ensure the main features of emotions are expressed correctly and naturally.

Recording. Before the recording begins, training courses are offered to help the speaker get into desired emotional status. We require the speakers to express different emotions spontaneously in their mother tongue to help them release tension. Then, an emotion-arousal session helps the speakers raise their emotional status so that the speakers can manage extreme expressions in level 3. During the recording, the guidance team arranges recording order of the different intensities to suit each speaker. Most of the speakers are recorded in the order of *weak*, *strong* and *medium*, as it will be easier to master medium intensity when the speaker is aware of the two extremes of one emotion.

Supervision and Feedback. During the recording, the guidance team would provide supervision from both emotion and speaking perspectives. In terms of emotion, the expression must cover all the features of the corresponding category, and the expressiveness should be consistent with the given intensity. Meanwhile, the speaker is requested to read the whole sentence with no pause and no pronunciation error. The guidance team would make the final judgment of whether the clip is qualified or not. In general, it would take the speaker two or three times to finish a qualified clip when first switching to another emotion category.

3.3 Analysis and Comparison

In this section, we show statistics of MEAD and compare with related datasets.

Analysis of the Speech Corpus. We keep track of the number of occurrences of different phonemes, including 15 vowels and 24 consonants based on the ARPAbet symbol set [27]. The distributions of vowels and consonants are shown as Fig. 2. Our corpus fully covers all vowels and consonants, and their occurrence frequency intuitively accords with the frequency of daily usage [10]. In RAVDESS [32], GRID [9], Lombard [1] and CREMA-D [4], the speech corpora have been greatly simplified, *e.g.* RAVDESS [32] uses only two sentences, GRID [9] and Lombard [1] use fixed sentence patterns, and CREMA-D [4] provides only 12 sentences for each emotion. These corpora are much less diverse than the TIMIT [16] corpus used by SAVEE [23] and our dataset. In our dataset,

30 sentences are used for each of the 7 basic emotion categories and 40 for the neutral category. Please refer to the supplementary materials for more details.

Analysis of the Audio-Visual Dataset. We demonstrate the quantitative comparison in Table 1. To enable the manipulation of emotion and its intensity in a more fine-grained way, our dataset is designed to contain neutral and 7 basic emotions including 3 intensity levels. The emotion categories are designed following [32], but our dataset provides 3 intensity levels for each emotion additionally. Thus, compared to recent datasets [23,32], our dataset provides richer emotion information. Another feature of MEAD is the inclusion of multi-view data. We place 7 cameras at different viewpoints to capture our portrait videos simultaneously, the detailed set-up is shown in supplementary. The viewpoint number is the largest among recent audio-visual datasets. The AV Digits [41] database is recorded from three angles (front, 45°, and profile), and the Ouluvs2 [2] dataset extends the view setting of the former to five for more fine-grained coverage. The Lombard [1] dataset and TCD-TIMIT [17] dataset only provide the front and side views. [9,23,32] provide data captured from the front view only. In terms of resolution, our dataset provides videos in the resolution of 1920×1080, which can be used in high fidelity portrait video generation. Similar to recent datasets [1,9,23,32], our audio sample rate is 48 kHz and video frame rate is 30. Note that SAVEE provides video data of fps 60, which is higher than that of us. However, many video feature extraction networks [53] usually downsample video frames of 30 fps. Thus, we adopt 30 fps which is sufficient for many video tasks like emotion recognition and portrait video generation.

3.4 Evaluation

We design a user study to evaluate the quality of our dataset, specifically, to examine if (i) the emotion performed by actors can be correctly and accurately recognized and (ii) the three levels of emotion intensity can be correctly distinguished. A hundred volunteers are gathered from universities for this experiment. The age range of the participants is from 18 to 40. We randomly selected six actors' data from MEAD for the user study, the testing data includes four males and two females.

Two types of experiments are conducted, namely emotion classification and intensity classification. For each type of experiment, two kinds of evaluations are performed – one on normal videos and the other on silenced videos. For emotion classification, we prepare test videos with varying emotion intensities in random order and the user needs to select one of 8 emotion categories. This evaluation is conducted 144 times for each user. The results of the silent video experiment are shown in Table 2, where the "Generated" stands for the user study results of the videos generated by our proposed baseline. It demonstrates that most of the testing video convey correct emotion to users. In emotions such as angry, happy and sad, we get an accuracy rate of over 0.90, while the results in neutral is much less satisfying, as the neutral expression is much easier to be in confusion with delicate emotional expressions in level 1. Considering that the intensity of

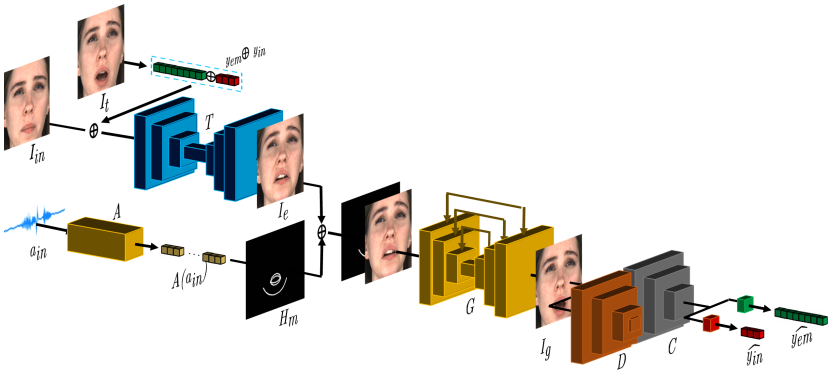

Fig. 3. Architecture of our baseline method. The overview of our emotional controllable talking-face approach. Our method includes three modules to drive the input neutral portrait image, audio clip and controllable emotion condition vector to obtain the output speech video.

emotion could affect the perception of the category of the emotion, we design the intensity classification experiment as follows: three videos of different intensities of one emotion is displayed in random order, and the participants are required to give the correct order from weak to strong. This experiment is conducted 42 times for each participant. From the results shown in the upper part of Table 3, we can observe that progressive emotion intensities are well distinguishable. Level 1 and level 2 are sometimes confused on some difficult emotions like disgust and contempt. More details about the user study can be found in the supplementary materials.

4 Emotional Talking-Face Baseline

Based on MEAD, we propose an emotional talking-face generation baseline that is able to manipulate emotion category and emotion intensity in talking-faces. The overview of our generation approach is depicted in Fig. 3. A two-branch architecture is designed to process the audio and emotional conditions separately. Specifically, the audio-to-landmarks module maps the audio to lip movements and neutral-to-emotion transformer synthesizes the desired emotion on the upper face of the target. Finally, the intermediate representations are fused in a refinement network to render the emotional talking-face video. The training phase requires three inputs, namely the input audio feature a_{in}, the identity-specifying image $\mathbf{I}_{in}$ and a target emotional image $\mathbf{I}_t$. Note that only the first two inputs are required in the testing phase.

Audio-to-Landmarks Module. We extract the D-dim Mel-Frequency Cepstral Coefficients (MFCC) feature from the input audio. A one-second temporal sliding window is used to pair audio features with video frames. The sample rate for the audio feature is set as 30, same as the video frame rate.

The audio feature is fed into the audio-to-landmarks module, which is composed of a long short term memory (LSTM) [21] module followed by a fully connected layer. A heatmap of the lower face $\mathbf{H}_m$ is formed by the output landmarks coordinates as shown in Fig. 3. We formulate the regression loss of the audio-to-landmark module as:

$$L_{reg} = ||A(a_{in}) - l_p||_2, \tag{1}$$

where $A(a_{in})$ and l_p are the predicted and ground truth mouth landmark coordinates respectively. Both $A(a_{in})$ and l_p are dimension-reduced representation by applying PCA.

Neutral-to-Emotion Transformer. Emotional image $\mathbf{I}_t$ is expected to be obtained based on the input neutral image $\mathbf{I}_{in}$ and emotion status vector y, specifically, $y = y_{em} \oplus y_{in}$ is the concatenation of two one-hot vectors: emotion category y_{em} and emotion intensity y_{in}. The module is composed of an encoder-decoder architecture where the encoder and decoder are constructed by symmetric 6-layer residual blocks [18] and 4 convolutional layers. We expand the emotion status vector y to the width and height of neutral face image $\mathbf{I}_{in}$ and concatenate them along the color channel as the input of our neutral-to-emotion transformer [42].

We supervise the neutral-to-emotion transformer with the reconstruction loss L_{rec} and perceptual loss [47] L_{con1} as follows:

$$\begin{aligned} L_{rec} &= \lambda_{rec}||\mathbf{I}_t - T(\mathbf{I}_{in}|y)||_1 \\ L_{con1} &= ||VGG_i(T(\mathbf{I}_{in}|y)) - VGG_i(\mathbf{I}_t)||_1, \end{aligned} \tag{2}$$

where the $T(\mathbf{I}_{in}|y)$ is the transformed emotional image, VGG_i is the activation layer for specific layer i on pre-trained VGG_16 [47] model. We can transform the neutral face into a face with input emotion category and intensity to achieve the emotion manipulation on the upper face.

Refinement Network. A refinement network is used to produce the final high-resolution face image $\mathbf{I}_g$ conditioned on the input lower face heatmap $\mathbf{H}_m$ and the generated emotional upper face image $T(\mathbf{I}_{in}|y)$. A U-Net [44] structure is adopted as the generator in this module. To generate realistic talking sequences with natural emotion, we first reconstruct the mouth region with the supervision of the target emotional image $\mathbf{I}_t$, and then constrain a content loss between the output image and $\mathbf{I}_t$ on the whole face. The mouth reconstruction loss and the content loss are defined as:

$$\begin{aligned} L_{mou} &= \lambda_{rec}||M(\mathbf{I}_t) - M(G(T(\mathbf{I}_{in}|y), \mathbf{H}_m))||_1 \\ L_{con2} &= ||VGG_i(G(T(\mathbf{I}_{in}|y), \mathbf{H}_m)) - VGG_i(\mathbf{I}_t)||_1, \end{aligned} \tag{3}$$

where M stands for a mouth area crop function and G is the generation network. We adopt LSGAN [34] scheme to train this module with the following adversarial loss:

$$L_{adv} = \frac{1}{2}E(D(\mathbf{I}_t)^2) + \frac{1}{2}E((1 - D(G(T(\mathbf{I}_{in}|y), \mathbf{H}_m)))^2). \tag{4}$$

Besides, we also use the total variation loss [24] to reduce the spiky artifacts and make the output image smooth:

$$L_{TV} = \lambda_{TV} \sum_{i,j}^{H,W} ||(\mathbf{I}_{g(i+1,j)} - \mathbf{I}_{g(i,j)})^2 + (\mathbf{I}_{g(i,j+1)} - \mathbf{I}_{g(i,j)})^2||, \tag{5}$$

where $\mathbf{I}_{g(i,j)}$ means the (i, j) pixel of $\mathbf{I}_g$. To further improve the generation quality, two pre-trained classifier models, both trained by cross-entropy loss from given labels, are used for emotion and intensity monitoring. We add two classification losses $L_{c_{em}}$ and $L_{c_{in}}$ to improve the performance of our generation network. Therefore, the final loss function should be formulated as:

$$L_{total} = L_{reg} + L_{rec} + L_{mou} + \lambda_{con} L_{con} + L_{adv} + L_{TV} + L_{c_{em}} + L_{c_{in}}, \tag{6}$$

where L_{con} is the summation of L_{con1} and L_{con2}. We empirically set all the coefficients of loss terms as 1, except 1e-5 for λ_{TV}.

5 Experiments and Results

5.1 Experiment Setup

Pre-processing. We evaluate our baseline method on our proposed dataset and set aside 20% of all collected data as the test set. For the videos, we crop and align the face in each frame using facial landmarks detected with an open source face alignment tool [56]. For the audios, we extract the 28-dim MFCC features. The size of the audio features of a one-second clip is set as 30×28 in accordance with the frame rate of the video.

Implementation Details. We train our network with Adam [26] optimizer using a learning rate of 0.001, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. We determine the weight coefficients of different loss functions through empirical validation. It takes nearly 4 hours to train the audio-to-landmark module, 24 hours to train the emotion transformer module and another 36 hours to continue training the result refinement module. Note that the transformer module can be trained together with the result refinement as well, but separate training is more stable according to experiment results. The training and testing phases are conducted on a single Nvidia GTX1080Ti GPU with a batch size of 1.

5.2 Baseline Comparison

We compare our emotion controllable talking-face generation approach with the following three methods.

CycleGAN [60] is an unsupervised image translation framework designed by Zhu *et al.*

ADAVR [58] is an talking-face generation method based on adversarially disentangled audio-visual representation proposed by Zhou *et al.*

Table 2. User study on emotion category discrimination. The accuracy rate of generated videos is nearly 10% lower compared to that of the captured videos, however, the accuracy distributions of generated and captured videos are basically consistent.

Emotion	angry	disgust	contempt	fear	happy	sad	surprise	neutral	mean
Captured	0.91	0.86	0.82	0.87	0.90	0.93	0.86	0.65	0.85
Generated	0.75	0.76	0.67	0.79	0.81	0.85	0.78	0.52	0.74

Table 3. User study on emotion intensity discrimination. The result is not satisfying enough, illustrating the intensity manipulation strategy still needs further improvement.

Groups	wrong types	angry	disgust	contempt	fear	happy	sad	surprised	mean
Captured	weak medium	0.13	0.17	0.19	0.11	0.12	0.13	0.08	0.13
	medium strong	0.07	0.05	0.09	0.13	0.07	0.10	0.07	0.08
	weak strong	0.03	0.03	0.04	0.04	0.02	0.04	0.03	0.03
	all wrong	0.02	0.01	0.02	0.01	0.01	0.02	0.00	0.01
Generated	weak medium	0.40	0.32	0.37	0.34	0.41	0.32	0.29	0.35
	medium strong	0.31	0.36	0.28	0.31	0.32	0.41	0.28	0.32
	weak strong	0.10	0.08	0.09	0.12	0.07	0.09	0.09	0.09
	all wrong	0.04	0.03	0.06	0.07	0.06	0.06	0.03	0.05

ATVGnet [6] is a hierarchical cross-modal method for talking-face generation.

Emotion Category Manipulation. This experiment attempts to generate talking-face videos with desired emotion category. Note that all the three baselines mentioned above are not capable of directly controlling facial expressions through conditional manipulation as our proposed method is. Therefore, we train several emotion-specific models for the audio-driven methods ATVGnet [6] and ADAVR [58], i.e. they are a set of models, each trained to generate emotional talking-face videos of only one kind of emotion. Similarly, the CycleGANs are each trained to translate a neutral face to a face of one specific emotion category. In Fig. 4, our method is able to generate diversified expressions from the input emotion categories while the audio-driven methods ATVGnet and ADAVR do not produce convincing emotional expressions. The results produced by CycleGAN contain obvious artifacts around areas where intense modification is needed, such as teeth, lips and eyes regions. The unpaired training style of CycleGAN also inherently impairs the temporal continuity of the produced talking-face videos.

We also conduct an compound emotion generation experiment by first generating an intermediate image with one kind of emotion and then modifying the mouth area with another kind of emotion in accordance with the emotion of the input audio. Please see the supplementary materials for the interesting results.

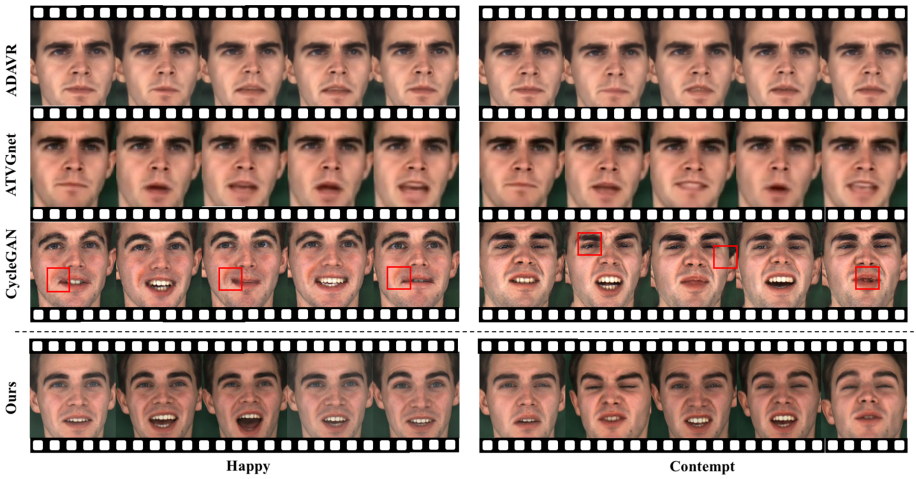

Fig. 4. Emotion category manipulation. Comparative results of emotion category manipulation on our dataset: three baseline methods against our method. The face regions in the red rectangles contain obvious artifacts caused by rapid dynamic variation.

Emotion Intensity Manipulation. This experiment attempts to generate talking-face videos with desired emotion category and intensity. Similar to the Emotion Catagory Manipulation experiment, ATVGnet [6], ADAVR [58] and CycleGAN [60] are trained as emotion-intensity-specific models. Figure 5 shows the intensity manipulation results. Specifically, the results from audio-driven methods ATVGnet and ADAVR do not exhibit observable intensity distinction, and the results of the CycleGAN exhibit slight differences between intensity levels while being prone to generating artifacts.

5.3 Evaluation Results for Our Baseline

Generation Quality. To quantitatively evaluate the quality of our generated portrait videos, we generate 48 portrait video clips of 8 different emotions (6 clips each emotion) of each actor for evaluation. We first adopt an emotion-video classification network [37] which achieves state-of-the-art performance on CK+ [33] to evaluate the emotion categories generation accuracy. Note that, the emotion categories in our dataset are consistent with CK+, and we retrained the network with the video data in our dataset and got the best result of accuracy 86.29%. Then we use this model to test on our generated videos of different emotions and got an accuracy of 86.26%, which reflects the genuineness and correctness of the emotion videos we generated. Then, we compare FID score [20] between baseline methods based on videos generated by same audios or videos. As shown in Tab. 4, neutral and surprise videos get the best quality and similarity compared to the training data. Since these baseline methods include no specified module to generate or manipulate emotions, the generated videos either are always recognized

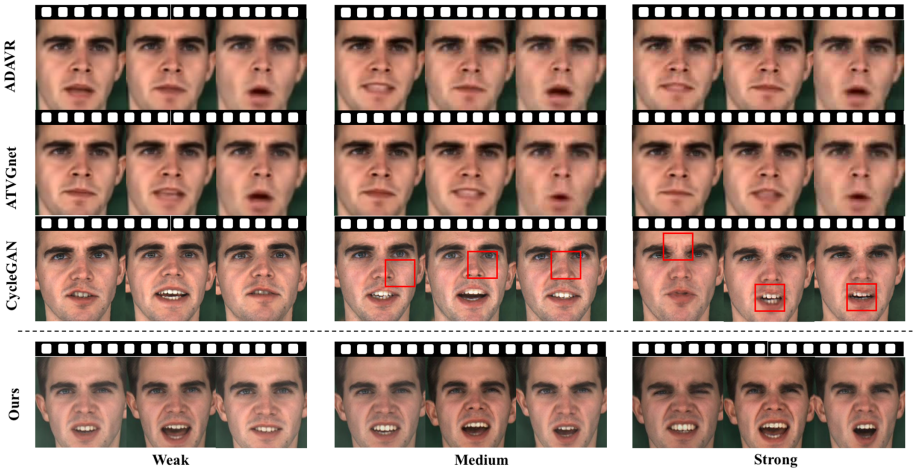

Fig. 5. Emotion intensity manipulation. Comparative results of emotion intensity manipulation on disgust: three baseline methods against our method. The face regions in the red rectangles contain unnatural skin texture and blurred artifacts

Table 4. Evaluation on generation quality. FID scores of different emotion categories. The results indicates that the generated videos can clearly express discrepancy between categories.

emotion	angry	disgust	contempt	fear	happy	sad	surprise	neutral	mean
FID	36.14	36.99	43.02	35.06	32.81	32.64	25.97	28.06	33.84

as neutral or contain strong artifacts which influence the evaluation seriously, resulting in the final FID scores over 100. We also provide quantitative and qualitative results of detailed ablation study in our supplementary materials, when different loss term in Eq. 6 is removed.

Emotion Generation Accuracy. Consistent with the dataset evaluation in Sect. 3.4, we also conduct a user study with 100 participants on our generated videos. The study includes two aspects: the accuracy of the generated facial emotion category and the accuracy of the generated emotion intensity. The emotion category accuracy is shown in Table 2, and the user study results on the ground truth video clips can be used as a reference. In general, the user study accuracy of generated facial emotion has decreased by nearly 10% on average, however, the accuracy distribution still mostly remains the same with the reference. Since some emotions are hard to distinguish, *e.g. fear* and *surprise*, *disgust* and *contempt*, even nearly 15% of the ground truth video clips are not rated to contain the correct emotion. We get the overall accuracy of 0.74 in all categories, indicating that our method performs well in manipulating facial emotion in portrait videos. As shown in Table 4, neutral and surprise videos get the best quality and similarity with the training data.

Similar as the intensity experiments of the dataset, the participants are given a sample with 3 video clips concatenated that respectively contain 3 different intensities of the same emotion. Then, the participants are asked to rank the emotion intensities and the accuracy is shown in Table 3. In Table 3, we note that the different levels of "sad" is best distinguished while the levels of "fear" and "surprised" are worst distinguished. Only about 20% of participants agree with the comparative emotional intensities generated by our method, which proves that our method cannot provide intensity manipulation results that are convincing enough in all categories.

6 Limitations and Future Work

Although our approach shows the ability of manipulating emotion for talking-face generation, there still exists some limitations. First, our method cannot disentangle emotion from input audio signal. The emotion in the lip is totally up to the input audio, we have not achieved emotion manipulation in the lips. The next step is to edit the whole emotional talking-face driven by neutral audio features. Second, according to the user study results, the generation results still need improvement in image quality and discrimination in emotion intensities. How to measure discrimination in emotion intensities is a challenging problem. One direction towards it is to depend on more explicit auxiliary annotations like FACS, which will be further implemented in our dataset. Furthermore, based on our three levels of intensity, we are getting close to achieving more accurate emotion intensity manipulation on talking-face task. However, the inadequacy of intensity levels in the recent datasets still forms the biggest obstacle for more fine-grained emotion generation. Collecting more fine-grained emotion data may requires smarter design and more complex procedure.

7 Conclusion

The generation of emotion in talking-face generation task is often neglected in previous works due to the absence of suitable emotional audio-visual dataset. We contribute a large-scale high-quality emotional audio-visual dataset, MEAD, providing rich and accurate affective visual and audio information with great detail. The emotional talking head generation baseline trained on MEAD achieves the manipulation of emotion and its intensity with favorable performance compared with current methods. We believe MEAD would benefit the community of talking-face generation and other research fields such as conditional generation, cross-modal understanding and expression recognition.

Acknowledgement. This work is supported by the SenseTime-NTU Collaboration Project, Singapore MOE AcRF Tier 1 (2018-T1-002-056), NTU SUG, and NTU NAP.

References

1. Alghamdi, N., Maddock, S., Marxer, R., Barker, J., Brown, G.J.: A corpus of audio-visual Lombard speech with frontal and profile views. J. Acoust. Soc. Am. **143**(6), EL523 (2018)
2. Anina, I., Zhou, Z., Zhao, G., Pietikäinen, M.: Ouluvs2: a multi-view audiovisual database for non-rigid mouth motion analysis. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, pp. 1–5. IEEE (2015)
3. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
4. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. **5**(4), 377–390 (2014)
5. Cao, Y., Tien, W.C., Faloutsos, P., Pighin, F.: Expressive speech-driven facial animation. ACM Trans. Graph. (TOG) **24**(4), 1283–1302 (2005)
6. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7832–7841 (2019)
7. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: deep speaker recognition. In: INTERSPEECH (2018)
8. Chung, J.S., Jamaludin, A., Zisserman, A.: You said that? In: British Machine Vision Conference (2017)
9. Cooke, M., Cunningham, S., Shao, X.: An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**, 2421 (2006)
10. Cowie, A.P., Gimson, A.: Oxford Advanced Learner's Dictionary of Current English. Oxford University Press, Oxford (1992)
11. Cowie, R., et al.: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag. **18**(1), 32–80 (2001)
12. Ding, H., Sricharan, K., Chellappa, R.: Exprgan: facial expression editing with controllable expression intensity. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
13. Edwards, P., Landreth, C., Fiume, E., Singh, K.: Jali: an animator-centric viseme model for expressive lip synchronization. ACM Trans. Graph. (TOG) **35**(4), 127 (2016)
14. Ezzat, T., Geiger, G., Poggio, T.: Trainable videorealistic speech animation, vol. 21. ACM (2002)
15. Fried, O., et al.: Text-based editing of talking-head video. ACM Trans. Graph. (TOG) **38**, 1–14 (2019)
16. Garofolo, J.S.: Timit acoustic phonetic continuous speech corpus. Linguistic Data Consortium (1993)
17. Harte, N., Gillen, E.: Tcd-timit: an audio-visual corpus of continuous speech. IEEE Trans. Multimedia **17**(5), 603–615 (2015)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)

19. Healy, E.W., Yoho, S.E., Wang, Y., Wang, D.: An algorithm to improve speech recognition in noise for hearing-impaired listeners. J. Acoust. Soc. Am. **134**(4), 3029–3038 (2013)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, pp. 6626–6637 (2017)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
22. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 172–189 (2018)
23. Jackson, P., Haq, S.: Surrey audio-visual expressed emotion (savee) database. http://kahlan.eps.surrey.ac.uk
24. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
25. Karras, T., Aila, T., Laine, S., Herva, A., Lehtinen, J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Trans. Graph. (TOG) **36**(4), 94 (2017)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
27. Klautau, A.: Arpabet and the timit alphabet (2001)
28. Kossaifi, J., et al.: Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. arXiv preprint arXiv:1901.02839 (2019)
29. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the radboud faces database. Cogn. Emot. **24**(8), 1377–1388 (2010)
30. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 35–51 (2018)
31. Lewis, J.: Automated lip-sync: background and techniques. J. Vis. Comput. Animation **2**(4), 118–122 (1991)
32. Livingstone, S.T., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): a dynamic, multimodal set of facial and vocal expressions in north American English. PLoS One **13**, e0196391 (2018)
33. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94–101. IEEE (2010)
34. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802 (2017)
35. Mariooryad, S., Busso, C.: Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, pp. 85–90. IEEE (2013)
36. Mattheyses, W., Verhelst, W.: Audiovisual speech synthesis: an overview of the state-of-the-art. Speech Commun. **66**, 182–217 (2015)
37. Meng, D., Peng, X., Wang, K., Qiao, Y.: frame attention networks for facial expression recognition in videos. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3866–3870. IEEE (2019). https://github.com/Open-Debin/Emotion-FAN

38. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affect. Comput. **10**(1), 18–31 (2017)
39. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. In: INTERSPEECH (2017)
40. Narayanan, A., Wang, D.: Ideal ratio mask estimation using deep neural networks for robust speech recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7092–7096. IEEE (2013)
41. Petridis, S., Shen, J., Cetin, D., Pantic, M.: Visual-only recognition of normal, whispered and silent speech. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6219–6223. IEEE (2018)
42. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: anatomically-aware facial animation from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 818–833 (2018)
43. Qian, S., et al.: Make a face: towards arbitrary high fidelity face manipulation. In: ICCV (2019)
44. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
45. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161 (1980)
46. Sadoughi, N., Busso, C.: Speech-driven expressive talking lips with conditional sequential generative adversarial networks. IEEE Trans. Affect. Comput. (2019)
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
48. Song, L., Wu, W., Qian, C., Qian, C., Loy, C.C.: Everybody's talkin': let me talk as you want. arXiv preprint arXiv:2001.05201 (2020)
49. Song, Y., Zhu, J., Wang, X., Qi, H.: Talking face generation by conditional recurrent adversarial network. arXiv preprint arXiv:1804.04786 (2018)
50. Srinivasan, S., Roman, N., Wang, D.: Binary and ratio time-frequency masks for robust speech recognition. Speech Commun. **48**(11), 1486–1501 (2006)
51. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. ACM Trans. Graph. (TOG) **36**(4), 95 (2017)
52. Vougioukas, K., Petridis, S., Pantic, M.: Realistic speech-driven facial animation with gans. Int. J. Comput. Vis., 1–16 (2019)
53. Wang, T.C., et al.: Video-to-video synthesis. In: NeurIPS (2018)
54. Williams, C.E., Stevens, K.N.: Emotions and speech: some acoustical correlates. J. Acoust. Soc. Am. **52**(4B), 1238–1250 (1972)
55. Wu, W., Cao, K., Li, C., Qian, C., Loy, C.C.: Transgaga: geometry-aware unsupervised image-to-image translation. In: CVPR (2019)
56. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: a boundary-aware face alignment algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2129–2138 (2018)
57. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9459–9468 (2019)
58. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: The Association for the Advancement of Artificial Intelligence Conference (2019)

59. Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., Singh, K.: Visemenet: audio-driven animator-centric speech animation. ACM Trans. Graph. (TOG) **37**(4), 161 (2018)
60. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)