Talking Face Generation by Adversarially Disentangled Audio-Visual Representation

Hang Zhou, Yu Liu, Ziwei Liu,* Ping Luo, Xiaogang Wang
The Chinese University of Hong Kong, Hong Kong, China
{zhouhang@link, yuliu@ee, zwliu@ie, xgwang@ee}.cuhk.edu.hk, pluo.lhi@gmail.com

Abstract

Talking face generation aims to synthesize a sequence of face images that correspond to a clip of speech. This is a challenging task because face appearance variation and semantics of speech are coupled together in the subtle movements of the talking face regions. Existing works either construct specific face appearance model on specific subjects or model the transformation between lip motion and speech. In this work, we integrate both aspects and enable arbitrary-subject talking face generation by learning disentangled audio-visual representation. We find that the talking face sequence is actually a composition of both subject-related information and speech-related information. These two spaces are then explicitly disentangled through a novel *associative-and-adversarial* training process. This disentangled representation has an advantage where both audio and video can serve as inputs for generation. Extensive experiments show that the proposed approach generates realistic talking face sequences on arbitrary subjects with much clearer lip motion patterns than previous work. We also demonstrate the learned audio-visual representation is extremely useful for the tasks of automatic lip reading and audio-video retrieval.

1 Introduction

Understanding talking faces visually is of great importance to machine perception and communication. Humans can not only guess the semantic meaning of words by observing lip movement but also imagine the scenario when a specific subject talks (*i.e.* face generation). Recent advances have focused on automatic lip reading, which surpasses human-level performance in certain domains. Here, we explore generating a video of arbitrary-subject speaking, which perfectly syncs with a specific speech where the speech information can be represented by either a clip of audio or video. We refer this problem as arbitrary-subject talking face generation, as shown in Fig. 1.

However, generating identity-preserving talking faces that clearly conveys certain speech information is a challenging task, since the continuous deformation of the face region relates to both intrinsic subject traits (Liu et al. 2015) and extrinsic speech vibrations. Previous efforts in this direction are mainly from computer graphics (Xie and Liu 2007; Wang et al. 2010; Fan et al. 2015; Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Thies et al. 2016). Researchers construct specific 3D face model for a chosen subject and the talking faces are animated by manipulating 3D meshes of the face model. However, these approaches strongly rely on the 3D face model and are hard to scale up to arbitrary identities. More recent attempts (Chung, Jamaludin, and Zisserman 2017) leverage the power of deep generative model and learn to generate talking faces from scratch. Though the resulting models can be applied to an arbitrary subject, the generated face sequences are sometimes blurry and not temporally meaningful. One important reason is that the subject-related and speech-related information are coupled together such that the talking faces are difficult to learn in a purely data-driven manner.

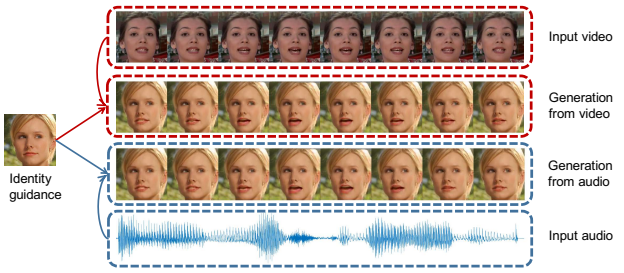

Figure 1: Problem description. Given a single face image of a target person, this work aims to generate the talking video based on the given speech information that is represented by either a clip of video or an audio.

To address the aforementioned problems, we integrate the identity-related and speech-related information by learning disentangled audio-visual representation, as illustrated in Fig. 2. We aim to disentangle a talking face sequence into two complementary representations, one containing identity information while the other containing speech information. However, directly separating these two parts is not a trivial task because the variations of face deformation can be extremely large considering the diversity of potential subjects and speeches.

The key idea here is using audio-visual speech recognition (Chung and Zisserman 2016b; 2017) (*i.e.* recognizing words from talking face sequence and audios, aka lip read-

*Corresponding author.

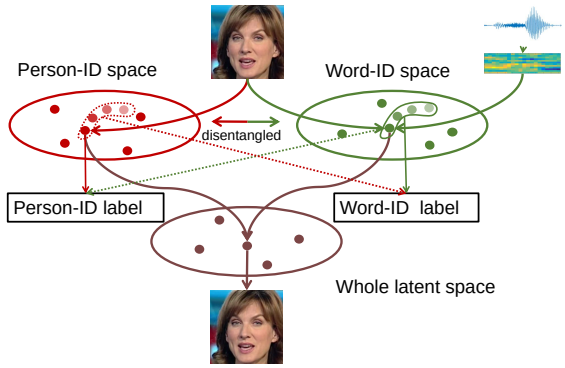

Figure 2: We propose to guide the information flow by using labels to ensure the spaces contain discriminative semantic information dispelling from each other. With the assumption that Word-ID space is shared between visual and audio information, our model can reconstruct faces base on either video or audio.

ing) as a probe task for *associating* audio-visual representations, and then employing *adversarial* learning to disentangle the subject-related and speech-related information inside them. Specifically, we first learn a joint audio-visual space where talking face sequence and its corresponding audio are embedded together. It is achieved by enforcing the lip reading result obtained from talking faces aligns with the speech recognition result obtained from audio. Next, we further utilize lip reading task to disentangle subject-related and speech-related information through adversarial learning (Liu et al. 2018b). Notably, we enforce one of the representations extracted from talking faces to fool the lip reading system, in the sense that it only contains subject-related information, but not speech-related information. Overall, with the aid of *associative-and-adversarial* training, we can jointly embed audio-visual inputs and disentangle subject and speech-related information of talking faces.

The contributions of this work can be summarized as follows. (1) A joint audio-visual representation is learned through audio-visual speech discrimination by associating several supervisions. Experiments show that the joint-embedding improves the baseline of lip reading result on LRW dataset (Chung and Zisserman 2016a). (2) Thanks to the discriminative nature of our joint representation, we disentangle the person-identity and speech information through adversarial learning for better talking face generation. (3) By unifying audio-visual speech recognition and audio-visual synchronizing, we achieve arbitrary-identity talking face generation from either video or audio speech as inputs in an end-to-end framework, which synthesizes high-quality and temporally-accurate talking faces.

2 Related Work

Generating Talking Faces. The work of synthesizing lip motion from either audio (Xie and Liu 2007; Wang et al. 2010; Fan et al. 2015; 2016; Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017; Chung, Jamaludin, and Zisserman 2017) or generating moving faces from videos (Thies et al. 2016; Liu et al. 2017b; Wiles, Koepke, and Zisserman 2018) has long been a task of concern in both the community of computer vision and graphics. However, most synthesis works from audio require a large amount of video footage of the target person for training, modeling, or sampling. They could not transfer the speech information to an arbitrary photo in the wild.

Chung, Jamaludin, and Zisserman (2017) use a setting that is different from the traditional ones. They try to directly generate the whole face image with different lip motions in an image-to-image translation manner based on audios.

But their method base on data-driven training using an autoencoder, which leads to blurry results and lacks continuity. More recently, Song et al. (2018) propose to use conditional RNN adversarial network, and Chen et al. (2018) propose to use correlation loss and three-stream GAN.

(Wiles, Koepke, and Zisserman 2018) use flow to generate high precision arbitrary-identity talking face based on videos and claim to be able to produce videos based on audios, but with no results shown. However, as a common problem, without specific disentangling face and lip motion information, they all cannot generate high-quality results.

Learning Audio-Visual Representation. The task of audio-visual speech recognition is a recognition problem uses either one or both video and audio as inputs. Using visual information only for recognition is also referred to as *Lip Reading*. A review of traditional methods for tackling this task has been made in Zhou et al. (2014) thoroughly. In recent years, this field develop quickly with the usage of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) for end-to-end word-level (Chung and Zisserman 2016a; Stafylakis and Tzimiropoulos 2017), sentence-level (Assael et al. 2016; Chung et al. 2017), and multi-view (Chung and Zisserman 2017) lip reading. In the meantime, the exploration of this topic has been greatly pushed forward by the build-up of large-scale word-level lip reading dataset (Chung and Zisserman 2016a), and the large sentence-level multi-view dataset (Chung and Zisserman 2017).

For the correspondence between human faces and audio clips, a number of works have been proposed to solve the problem of the audio-video synchronization between mouth motion and speech (McAllister et al. 1997; Chung and Zisserman 2016b). Particularly, SyncNet (Chung and Zisserman 2016b; 2017) used two stream CNNs to sync audio *mfcc* with 5 consecutive frames. In Chung and Zisserman (2017), they further fixed the sync image feature as the pre-training for lip reading, but the two tasks are still separate from each other. Recently, works from (Nagrani, Albanie, and Zisserman 2018b; 2018a) also attempt to learn the association between a human face and voice for identity recognition instead of semantic level synchronization.

3 Approach

We propose Disentangled Audio-Visual System (DAVS), an end-to-end trainable network for talking face genera-

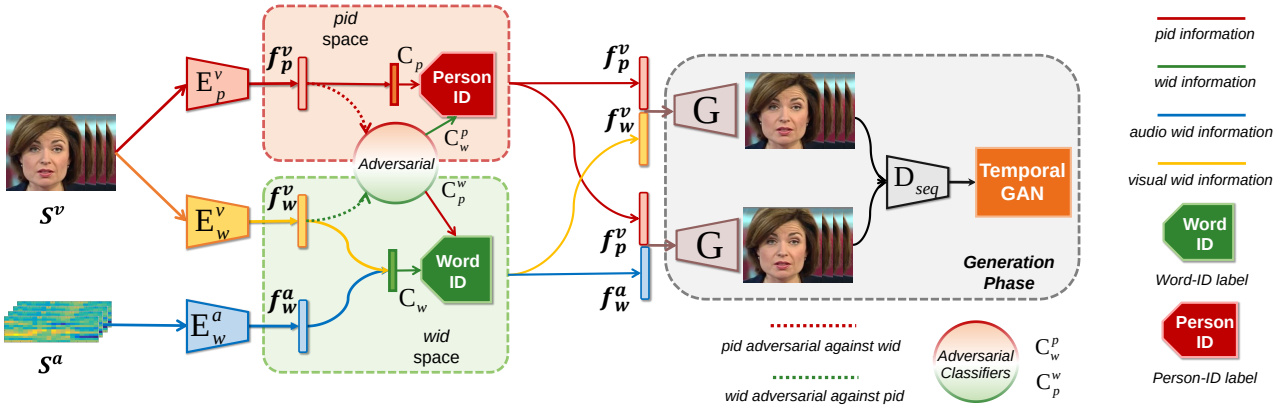

Figure 3: Illustration of our framework. E_p^v is the encoder that encodes Person-ID information from **visual** source to the *pid* space, E_w^v and E_w^a are the Word-ID encoders that extract speech content information to *wid* space from **video** and **audio**. Decoder G takes any combination of features in *pid* and *wid* space to generate faces. D_{seq} is a discriminator used for GAN loss. The adversarial training part contains two extra classifiers C_p^w and C_w^p. The details of embedding the *wid* space and adversarial training are shown in Fig 4 and 5.

tion by learning disentangled audio-visual representations, as shown in Fig. 3.

We leverage both talking video S^v and its corresponding audio S^a as training inputs. For learning the disentangled audio-visual representations between Person-ID space (*pid*) and the Word-ID space (*wid*), there are three encoder networks involved:

- **Video** to **Word-ID** space encoder (E_w^v): E_w^v learns to embed the video frame s^v into a **visual** representation f_w^v which only contains speech-related information. It is achieved by learning a joint embedding space which *associates* video and audio that correspond to the same word.
- **Audio** to **Word-ID** space encoder (E_w^a): E_w^a learns to embed the speech s^a into an **audio** representation f_w^a, which resides in the shared space with f_w^v as introduced above.
- **Video** to **Person-ID** space encoder (E_p^v): E_p^v learns to embed the video frame s^v into a representation f_p^v which only contains subject-related information. It is achieved by the *adversarial* training process, forcing our target representation f_p^v to fool the speech recognition system.

The whole idea of our pipeline is to first learn the discriminative audio-visual joint space *wid*, then disentangle it from the *pid* space. Finally to combine features from the two spaces to get generation results. Specifically, for learning the *wid* space, we employ three supervisions: the supervision of Word-ID labels with shared classifier C_w for associating audio and visual signals with semantic meanings; contrastive loss $\mathcal{L}_C$ for pulling paired video and audio samples closer; and an adversarial training supervision on audio and video features to make them indistinguishable. As for the *pid* space, Person-ID labels from extra labeled face data are used. For disentangling *wid* and *pid* spaces, adversarial training is employed. As for generation, we introduce L_1-norm reconstruction loss $\mathcal{L}_{L_1}$ and temporal GAN loss $\mathcal{L}_{GAN}$ for sharpness and continuity.

3.1 Learning Joint Audio-Visual Representation

We learn a joint audio-visual space that associates representations from both sources. We constrain the extracted audio representation to be close to its corresponding visual representation, forcing the embedded features to share a same distribution and restricting $f_w^a \simeq f_w^v$, so that $G(f_p^v, f_w^v) \simeq G(f_p^v, f_w^a)$ can be achieved. While requiring information of person facial identity flows from the *pid* space, the other space of *wid* would have to be person-ID invariant. The task of audio-visual speech recognition benefits us in achieving the shared latent space assumption and creating a discriminative space through mapping videos and audios to word labels. The implementation of learning the space is shown in Fig 4 (a). Then with the discriminative embedding, we can take the advantage of adversarial training for thoroughly information disentangling as described in Sec. 3.2

Sharing Classifier. After the embedded features are extracted from the *wid* encoders E_w^a, E_w^v to get $F_w^v = [f_{w(1)}^v, \cdots, f_{w(n)}^v]$ and $F_w^a = [f_{w(1)}^a, \cdots, f_{w(n)}^a]$, normally they would be fed into different classifiers for visual and audio speech recognition. Here we share the classifier for both the modalities to enforce them to share their distributions. As a classifier's weight $\mathbf{w}_j$ tend to fall into the center of the clustering of the features belonging to the j'th class, through sharing the weights, the features between both modalities are pulled towards the centroid of the class (Liu et al. 2018a). The supervision is denoted as $\mathcal{L}_w$.

Contrastive Loss. As the problem of mapping audio and visual together is very similar to feature mapping (Chopra, Hadsell, and LeCun 2005), retrieval and particularly the same as lip sync (Chung and Zisserman 2016b), we adopted the contrastive loss which aims at bringing closer paired data while dispelling unpaired as a baseline. During training, for a batch of N audio-video samples, the mth and nth sample are drawn with labels $l_{m=n} = 1$ while the others $l_{m\neq n} = 0$. The distance metric used to measure the dis-

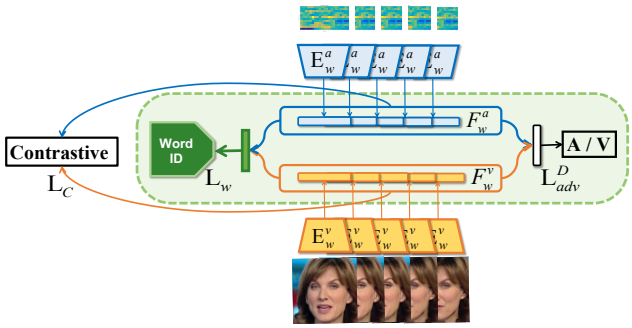

Figure 4: Illustration of embedding the audio-visual shared space wid. The encoded features $F_w^v = [f_{w(1)}^v, \cdots, f_{w(n)}^v]$ and $F_w^a = [f_{w(1)}^a, \cdots, f_{w(n)}^a]$ are constrained by contrastive loss $\mathcal{L}_c$, classification loss $\mathcal{L}_w$ and domain adversarial training $\mathcal{L}_{adv}^D$.

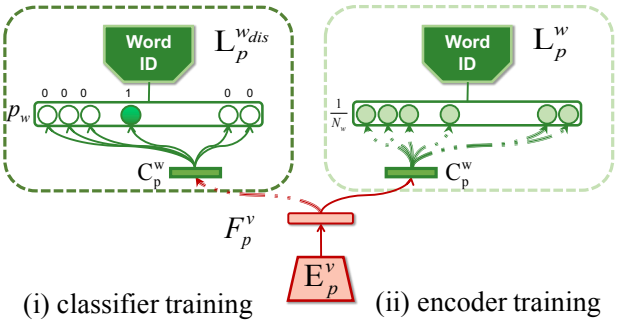

Figure 5: Procedure of adversarial training for dispelling wid information from pid space. The training for classifier C_p^w is illustrated on the left and encoder E_p^v on the right. The weights are updated on solid lines but not on the dashed lines.

tance between $F_{w(m)}^a$ and $F_{w(n)}^v$ here is the euclidean norm $d_{mn} = \|F_{w(m)}^v - F_{w(n)}^a\|_2$. The objective can be written as:

$$\mathcal{L}_C = \sum_{n=1,m=1}^{N,N} (l_{mn} d_{mn} + (1 - l_{mn}) \max(1 - d_{mn}, 0)) \tag{1}$$

During our implementation, all features F_w^v, F_w^a used in this loss are normalized first.

Domain Adversarial Training. To further push the face and audio features to be in the same distribution, we apply a domain adversarial training. An extra two-class domain classifier is appended for distinguishing the source of the feature. The audio and face encoders are then trained to prevent the classifier from success. This is mostly a simple version of the adversarial training described in section 3.2. We refer to the objective of this method as $\mathcal{L}_{adv}^D$.

3.2 Adversarial Training for Latent Space Disentangling

In this section, we describe how we disentangle the subject-related and speech-related information in the joint embedding space using *adversarial* training.

Specifically, we would like the Person-ID feature f_p^v to be free of Word-ID information. The discriminator could be formed to be a classifier C_p^w to map the collection of $F_p^v = [f_{p\,(1)}^v, \cdots, f_{p\,(n)}^v]$ to the N_w Word-ID classes. The objective function for training the classifier is the same as softmax cross-entropy loss. However, the parameter updating is only performed on C_p^w, where ${p_w}^j$ is the one-hot label of the identity classes:

$$\mathcal{L}_p^{w_{dis}}(\mathrm{C}_p^w|\mathrm{E}_p^v) = -\sum_{j=1}^{N_w} {p_w}^j \log(\mathrm{softmax}(\mathrm{C}_p^w(F_p^v))_j). \tag{2}$$

Then we update the encoder while fixing the classifier. The way to ensure that the features have lost all information about speech information is that it produces the same prediction for all classes after being sent into C_p^w. One way to form this limitation is to assign the probabilities of each word-label to be $\frac{1}{N_w}$ in softmax cross-entropy loss. The problem of this loss is that it would still backward gradient for updating parameters even if it reaches the minimum, so we propose to implement the loss using Euclidean distance:

$$\mathcal{L}_p^w(\mathrm{E}_p^v|\mathrm{C}_p^w) = \sum_{j=1}^{N_w} \|\mathrm{softmax}(\mathrm{C}_p^w(F_p^v))_j - \frac{1}{N_w}\|_2^2. \tag{3}$$

The dual feature f_w^v should also be free of pid information accordingly, so the loss for encoding pid information from each f_w^v using classifier C_w^p and loss for wid encoder E_w^v to dispel pid information can be formed as follows:

$$\mathcal{L}_w^{p_{dis}}(\mathrm{C}_w^p|\mathrm{E}_w^v) = -\sum_{j=1}^{N_p} {p_p}^j \log(\mathrm{softmax}(\mathrm{C}_w^p(f_w^v))_j), \tag{4}$$

$$\mathcal{L}_w^p(\mathrm{E}_w^v|\mathrm{C}_w^p) = \sum_{j=1}^{N_p} \|\mathrm{softmax}(\mathrm{C}_w^p(f_w^v))_j - \frac{1}{N_p}\|_2^2. \tag{5}$$

N_p is the number of person identities in the training set for embedding pid space. We summarize the adversarial training procedure for classifier C_p^w and encoder E_p^v as Fig. 5.

3.3 Inference: Arbitrary-Subject Talking Face Generation

In this section, we describe how we generate arbitrary-subject talking faces using the disentangled representations learned above. Combining pid feature f_p^v with either of the **video** wid feature f_w^v or **audio** wid feature f_w^a, our system can generate a frame using the decoder G. The newly generated frame can be expressed as $\mathrm{G}(f_p^v, f_w^v)$, $\mathrm{G}(f_p^v, f_w^a)$.

Here we take synthesizing talking faces from audio wid information as example. The generation results can be expressed as $\mathrm{G}(f_{p\,(k)}^v, F_w^a) = \{\mathrm{G}(f_{p\,(k)}^v, f_{w(1)}^a), \cdots, \mathrm{G}(f_{p\,(k)}^v, f_{w(n)}^a)\}$, where $f_{p\,(k)}^v$ is the pid feature of the random kth frame, which acts

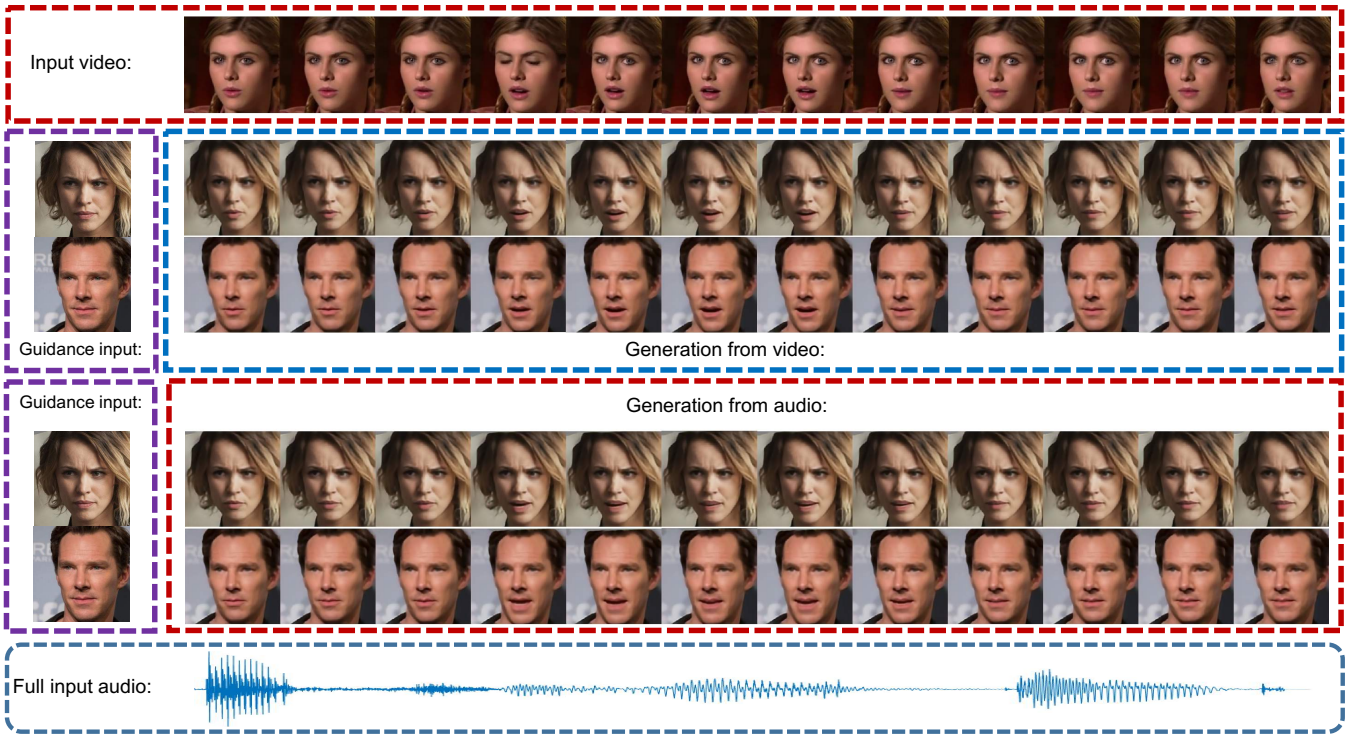

Figure 6: Qualitative results. The guidance input image is on the left . The upper half is the generation from video and lower half is the generation from audio information.

as identity guidance. Our overall loss function consists of a L_1 reconstruction loss and a temporal GAN loss, where a discriminator D_{seq} takes the generated sequence $\mathrm{G}(f^v_{p\,(k)}, F^a_w)$ as input. These two terms can be formulated as follows:

$$\mathcal{L}_{L_1} = \|S^v - \mathrm{G}(f^v_{p\,(k)}, F^a_w)\|_1, \tag{6}$$

$$\begin{aligned}\mathcal{L}_{GAN} = \mathbb{E}_{S^v}[\log \mathrm{D}_{seq}(S^v)] + \\ \mathbb{E}_{F^v_p, F^a_w}[\log(1 - \mathrm{D}_{seq}(\mathrm{G}(f^v_{p\,(k)}, F^a_w)]\end{aligned} \tag{7}$$

The overall reconstruction loss can be written as $\mathcal{L}_{Re}$, α is a hyper-parameter that leverages the two losses.

$$\mathcal{L}_{Re} = \mathcal{L}_{GAN} + \alpha \mathcal{L}_{L_1}. \tag{8}$$

The same procedure can be applied to generation from video information by substituting F^a_w with F^v_w. As the reconstruction from audio and video can perform at the same time during training, we use $\mathcal{L}_{Re}$ to denote the overall reconstruction loss function.

4 Experiments

Datasets. Our model is trained and evaluated on the LRW dataset (Chung and Zisserman 2016a), which is currently the largest word-level lip reading dataset with 1-of-500 diverse word labels. For each class, there are more than 800 training samples and 50 validation/test samples. Each sample is a one-second video with the target word spoken. Besides, the identity-preserving module of the network is trained on a subset of the MS-Celeb-1M dataset (Guo et al. 2016). All the talking faces in the videos are detected and aligned using RSA algorithm (Liu et al. 2017a), and then resized to 256×256. For the audio stream, we follow the implementation in (Chung and Zisserman 2016b) to extract the $mfcc$ features at the sampling rate of 100Hz. Then we match each image with a $mfcc$ audio input with the size of $12 * 20$.

Network Architecture. We adopted a modified VGG-M (Chatfield et al. 2014) as the backbone for encoder E^v_p, and for encoder E^v_w, we modified a simple version of FAN (Bulat and Tzimiropoulos 2017). The encoder E^a_w has a similar structure as that used in Chung and Zisserman (2016b). Meanwhile, our decoder contains 10 convolution layers with 6 bilinear upsampling layers to obtain a full-resolution output image. All the latent representations are set to be 256-dimensional.

Implementation Details. We implemented DAVS using Pytorch. The batch size is set to be 18 with 1e-4 learning rate and trained on 6 Titan X GPUs. It takes about 4 epochs for the audio-visual speech recognition and person-identity recognition to converge and another 5 epochs for further tuning the generator. The whole training process takes about a week. Due to the alignment of the training set, the directly generated results may suffer from a scale changing problem, so we apply the subspace video stabilization (Liu et al. 2011) for smoothness.

Table 1: PSNR and SSIM scores for generation from audio and video wid information with and without GAN loss.

Approach \ Score	PSNR	SSIM
Audio ($\mathcal{L}_{L_1}$)	25.4	0.859
Video ($\mathcal{L}_{L_1}$)	25.7	0.865
Audio ($\mathcal{L}_{Re}$)	**26.7**	**0.883**
Video ($\mathcal{L}_{Re}$)	**26.8**	**0.884**

Table 2: User study of our generation results and reproduced baseline. The results are averaged over person and time.

Method \ Rate	Realistic	Lip-Audio Sync
Reproduced Baseline	44.1%	58.0%
Ours (Generation from Audio)	51.5%	72.3%
Ours (Generation from Video)	**87.8%**	**88.4%**

4.1 Results of Arbitrary-Subject Talking Face Generation

At test time, the input identity guidance s_p^v to E_p^v is any person's face image and only one of the source for speech information S_w^v, S_w^a is needed to generate a sequence of images.

Quantitative Results. To verify the effectiveness of our GAN loss for improving image quality, we evaluate the PSNR and SSIM (Wang et al. 2004) score on the test set of LRW based on reconstruction. We compare the results with and without the GAN loss in Table 1. We can see that both the scores are improved by changing $\mathcal{L}_{L_1}$ to $\mathcal{L}_{Re}$.

Qualitative Results. Video results can be found on our project page[1]. Here we show image results in Fig 6. The input guidance photos are celebrities chosen randomly from the Internet. Our model is capable of generating talking faces based on both audios or videos. The focus of our work is to improve audio guided generation results by using joint audio-visual embedding, so we compare our work with Chung, Jamaludin, and Zisserman (2017) at Fig 7. It can be clearly seen that our results outperform theirs from both the perspective of identity preserving and image quality.

User Study. We also conduct user study to investigate the visual quality of our generated results comparing with a fair reproduction of Chung, Jamaludin, and Zisserman (2017) with our network structure. They are evaluated *w.r.t* two different criteria: whether participants could regard the generated talking faces as realistic (true or false), and how much percent of the time steps the generated talking faces temporally sync with the corresponding audio. We generate videos with the identity guidance to be 10 different celebrity photos. As for speech content information, we use clips from the test set of LRW dataset and selections from the Voxceleb dataset (Nagrani, Chung, and Zisserman 2017), which is not used for training. There are overall 10 participants involved, and the results are average over persons and video time steps. The ground-truth is not included in the user study.

[1]https://liuziwei7.github.io/projects/TalkingFace

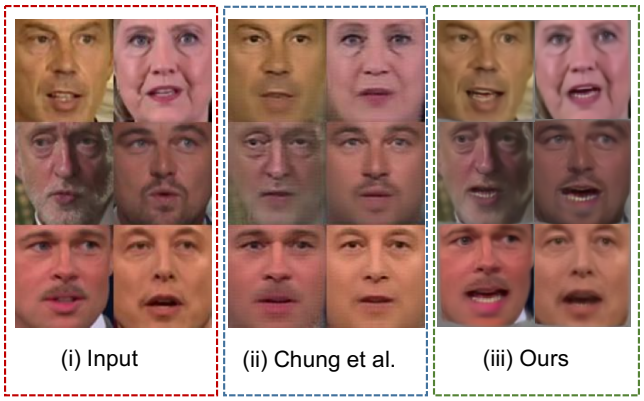

Figure 7: Qualitative results comparing with Chung et al. The mouth shapes are arbitrary.

Different subjects may behave different lip motion given the same audio clip and it is not desirable for the ground-truth to interfere with the participants' perception. When conducting the user study for lip sync evaluation, we asked the participants to only focus on whether the lip motion and given audio are temporally synchronized. Their ratings indicate that our generation results outperform the baseline by synchronizing rate and the extent of realistic, according to Table 2.

4.2 Effectiveness of Audio-Visual Representation

In order to inspect the quality of our embedded audio-visual representation, we evaluate the discriminative power and the closeness of our co-embedded features.

Word-level Audio-Visual Speech Recognition. We report audio-visual speech recognition accuracy on the test set of LRW dataset. Containing the task of visual recognition (lip reading) and audio recognition (speech recognition).

Our model structure for lip reading is similar to the Multiple-Towers method which reaches the highest lip reading results in Chung and Zisserman (2016a), so we consider it as a baseline. The difference is that the concatenation of features is performed at the spacial size of 1×1 in our setting. This would not be a reasonable choice for this task alone for the spatial information in images would be lost across time. However, as shown in Table 3, our results adding the contrastive loss alone outperforms the baseline. With the help of sharing classifier and domain adversarial training, the results improve a large margin.

Audio-Video Retrieval. To evaluate the closeness between the audio and face features, we borrow protocols used in the retrieval community. The retrieval experiments are conducted on the test set of LRW with 25000 samples, which means that given a test target video (audio), we try to find the closest audio (video) based on the distance of wid features F_w^v, F_w^a among all the test samples. Here we report the $R@1$, $R@10$ and $Med\ R$ measurements which is the same as Faghri et al. (2017). As we can see in Table 3, with all supervisions, the highest results can be achieved.

Qualitative Results. Figure 8 shows the sequence generation quality from audio with different supervisions provided

Table 3: Audio-Visual Speech Recognition and 1:25000 audio-video retrieval results with different supervisions. The first column is the supervisions, we use $\mathcal{L}_C$ to represent contrastive loss, SC for sharing classifier, $\mathcal{L}_{adv}^D$ for the adversarial training.

Approach	Audio-Visual Speech Recognition			Video to Audio Retrieval			Audio to Video Retrieval		
	Visual acc.	Audio acc.	Combine acc.	$R@1$	$R@10$	$Med\ R$	$R@1$	$R@10$	$Med\ R$
(Chung and Zisserman 2016a)	61.1%	-	-	-	-	-	-	-	-
Ours ($\mathcal{L}_C$)	61.8%	81.7	90.8%	29.3	56.3	6.0	29.8	56.3	6.0
Ours ($\mathcal{L}_C$ + SC)	65.6%	91.6%	94.9%	38.8	66.4	3.0	44.5	70.9	2.0
Ours ($\mathcal{L}_C$ + $\mathcal{L}_{adv}^D$)	63.5%	88.1%	93.7%	39.3	67.9	3.0	42.2	69.2	2.0
Ours ($\mathcal{L}_C$ + SC + $\mathcal{L}_{adv}^D$)	**67.5%**	**91.8%**	**95.2%**	**64.2**	**84.7**	**1.0**	**67.7**	**85.8**	**1.0**

Table 4: Ablation study on disentangle mechanism.

Experiment	Audio Generation to Source			Video Generation to Source		
	Retrieval $R@1$	Landmark L_2	ID Squared L_2	Retrieval $R@1$	Landmark L_2	ID Squared L_2
Direct Replication	2.5	4.27	-	2.5	4.31	-
Without Disentanglement	53.8	3.94	0.212	90.8	3.60	0.194
With Disentanglement	**60.5**	**3.48**	**0.188**	**95.3**	**2.85**	**0.174**

above. We can observe from the figure that given the same clip of audio, the duration of the mouth opening and to what extent it is opened is affected by different supervisions. Sharing the classifier apparently lengthens the time and strength of the mouth opening to make the image closer to the ground truth. Combining with the adversarial training makes the image quality improves. Note that it is not a one-to-one mapping between audio and lip motion; different subjects may behave different lip motion given the same audio clip so the final results may not perform the same as the ground truth.

4.3 Identity-Speech Disentanglement

To validate our adversarial training is able to disentangle speech information from person-ID branch, we use person-ID encoder on every frame of a video and concatenate them to get $F_p^v = \{f_{p(1)}^v, ..., f_{p(n}^v)\}$. Then we train an SVM to map training samples to their wid labels and test the results, which implies that we attempt to find the wid information left in the pid encoder. The whole procedure is repeated before and after the feature disentanglement. Before the disentanglement, 27.8% of the test set can be assigned to the right class, but only 9.7% left after, indicating that considerable speech content information within the encoder E_p^v is gone.

We then highlight the merits of adversarial disentanglement from two aspects, **identity preserving** and **lip sync quality**. For identity preserving, we use OpenFace's squared L2 similarity score as an indicator and compare the identity distance between the generated faces and the original ones (lower indicates more similar). For lip sync quality, we detect 20 landmarks using dlib library (King 2009) around the lips to characterize its deviation from ground truth, measured by the averaged L2-norm (lower is better). Then we conduct retrieval experiments between all generated results and source videos based on extracted F_{wid}^v features. Experiments are also conducted on a direct replication of every video clip, to prove that the retrieval results are affected by lip motion rather than appearance features. From Table 4, we

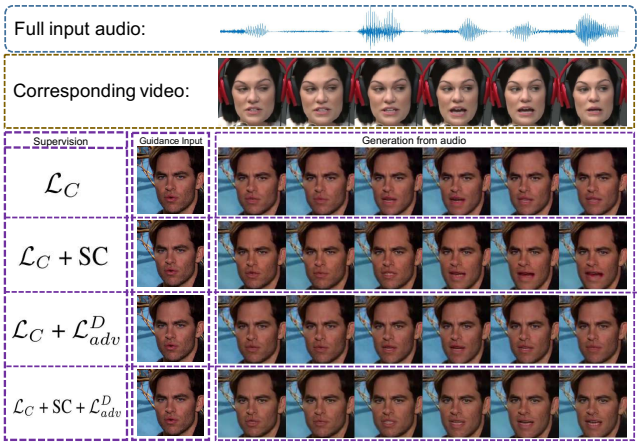

Figure 8: Qualitative results for different types of supervisions. The left indicates different supervisions. All the generations are audio-based.

can observe that adversarial disentanglement indeed helps improves lip sync quality.

5 Conclusion

In this paper, we propose a novel framework called Disentangled Audio-Visual System (DAVS), which generates high quality talking face videos using disentangled audio-visual representation. Specifically, we first learn a joint audio-visual embedding space wid with discriminative speech information by leveraging the word-ID labels. Then we disentangled the wid space from the person-ID pid space through adversarial learning. Compared to prior works, DAVS has several appealing properties: (1) A joint audio-visual representation is learned through audio-visual speech discrimination by associating several supervisions. The disentangled audio-visual representation significantly improves lip

reading performance; (2) Audio-visual speech recognition and audio-visual synchronizing are unified in an end-to-end framework; (3) Most importantly, arbitrary-subject talking face generation with high-quality and temporal accuracy can be achieved by our framework; both audio and video speech information can be employed as input guidance.

Acknowledgements

We thank Yu Xiong for helpful discussions and his assistance with our video. This work is supported by SenseTime Group Limited, the General Research Fund sponsored by the Research Grants Council of Hong Kong and the Hong Kong Innovation and Technology Support Program (No.ITS/121/15FX).

References

Assael, Y. M.; Shillingford, B.; Whiteson, S.; and de Freitas, N. 2016. Lipnet: Sentence-level lipreading. *arXiv preprint arXiv:1611.01599*.

Bulat, A., and Tzimiropoulos, G. 2017. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *ICCV*.

Chatfield, K.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*.

Chen, L.; Li, Z.; Maddox, R. K.; Duan, Z.; and Xu, C. 2018. Lip movements generation at a glance. In *ECCV*.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*.

Chung, J. S., and Zisserman, A. 2016a. Lip reading in the wild. In *ACCV*.

Chung, J. S., and Zisserman, A. 2016b. Out of time: automated lip sync in the wild. In *ACCV*.

Chung, J. S., and Zisserman, A. 2017. Lip reading in profile. In *BMVC*.

Chung, J.; Senior, A.; Vinyals, O.; and Zisserman, A. 2017. Lip reading sentences in the wild. In *CVPR*.

Chung, J. S.; Jamaludin, A.; and Zisserman, A. 2017. You said that? In *BMVC*.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*.

Fan, B.; Wang, L.; Soong, F. K.; and Xie, L. 2015. Photo-real talking head with deep bidirectional lstm. In *ICASSP*.

Fan, B.; Xie, L.; Yang, S.; Wang, L.; and Soong, F. K. 2016. A deep bidirectional lstm approach for video-realistic talking head. *Multimedia Tools and Applications*.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *JMLR*.

Liu, F.; Gleicher, M.; Wang, J.; Jin, H.; and Agarwala, A. 2011. Subspace video stabilization. *TOG*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Liu, Y.; Li, H.; Yan, J.; Wei, F.; Wang, X.; and Tang, X. 2017a. Recurrent scale approximation for object detection in cnn. In *ICCV*.

Liu, Z.; Yeh, R.; Tang, X.; Liu, Y.; and Agarwala, A. 2017b. Video frame synthesis using deep voxel flow. In *ICCV*, volume 2.

Liu, Y.; Song, G.; Shao, J.; Jin, X.; and Wang, X. 2018a. Transductive centroid projection for semi-supervised large-scale recognition. In *ECCV*.

Liu, Y.; Wei, F.; Shao, J.; Sheng, L.; Yan, J.; and Wang, X. 2018b. Exploring disentangled feature representation beyond face identification. *CVPR*.

McAllister, D. F.; Rodman, R. D.; Bitzer, D. L.; and Freeman, A. S. 1997. Lip synchronization of speech. In *Audio-Visual Speech Processing: Computational & Cognitive Science Approaches*.

Nagrani, A.; Albanie, S.; and Zisserman, A. 2018a. Learnable pins: Cross-modal embeddings for person identity. In *ECCV*.

Nagrani, A.; Albanie, S.; and Zisserman, A. 2018b. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*.

Song, Y.; Zhu, J.; Wang, X.; and Qi, H. 2018. Talking face generation by conditional recurrent adversarial network. *arXiv preprint arXiv:1804.04786*.

Stafylakis, T., and Tzimiropoulos, G. 2017. Combining residual networks with lstms for lipreading. *INTERSPEECH*.

Suwajanakorn, S.; Seitz, S. M.; and Kemelmacher-Shlizerman, I. 2017. Synthesizing obama: learning lip sync from audio. *TOG*.

Thies, J.; Zollhöfer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*.

Wang, L.; Qian, X.; Han, W.; and Soong, F. K. 2010. Synthesizing photo-real talking head via trajectory-guided sample selection. In *Eleventh Annual Conference of the International Speech Communication Association*.

Wiles, O.; Koepke, A.; and Zisserman, A. 2018. X2face: A network for controlling face generation by using images, audio, and pose codes. In *ECCV*.

Xie, L., and Liu, Z.-Q. 2007. Realistic mouth-synching for speech-driven talking face using articulatory modelling. *IEEE Transactions on Multimedia*.

Zhou, Z.; Zhao, G.; Hong, X.; and Pietikäinen, M. 2014. A review of recent advances in visual speech decoding. *Image and Vision Computing*.