Animating Portrait Line Drawings from a Single Face Photo and a Speech Signal

Ran Yi
Shanghai Jiao Tong University
China
ranyi@sjtu.edu.cn

Zipeng Ye
Ruoyu Fan
yezp17@mails.tsinghua.edu.cn
fry21@mails.tsinghua.edu.cn
Tsinghua University
China

Yezhi Shu
Tsinghua University
China
shuyz19@mails.tsinghua.edu.cn

Yong-Jin Liu*
BNRist, CS Department, MOE-Key Laboratory of Pervasive Computing, Tsinghua University
China
liuyongjin@tsinghua.edu.cn

Yu-Kun Lai
Cardiff University
UK
LaiY4@cardiff.ac.uk

Paul L. Rosin
Cardiff University
UK
RosinPL@cardiff.ac.uk

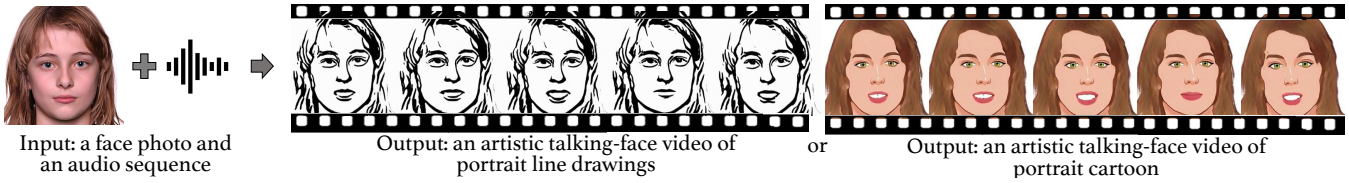

Figure 1: Our method generates an artistic animation of a face portrait from a single face photo and a speech signal. In addition to line drawings, our method can be easily applied to portrait cartoon. The input image by SKV Florbal (Public Domain).

ABSTRACT

Animating a single face photo is an important research topic which receives considerable attention in computer vision and graphics. Yet line drawings for face portraits, which is a longstanding and popular art form, have not been explored much in this area. Simply concatenating a realistic talking face video generation model with a photo-to-drawing style transfer module suffers from severe inter-frame discontinuity issues. To address this new challenge, we propose a novel framework to generate artistic talking portrait-line-drawing video, given a single face photo and a speech signal. After predicting facial landmark movements from the input speech signal, we propose a novel GAN model to simultaneously handle domain transfer (from photo to drawing) and facial geometry change (according to the predicted facial landmarks). To address the inter-frame discontinuity issues, we propose two novel temporal coherence losses: one based on warping and the other based on a temporal coherence discriminator. Experiments show that our model produces high quality artistic talking portrait-line-drawing videos and outperforms baseline methods. We also show our method can be easily extended to other artistic styles and generate good results. The source code is available at https://github.com/AnimatePortrait/AnimatePortrait.

CCS CONCEPTS

• **Computing methodologies** → **Non-photorealistic rendering**.

KEYWORDS

Face Animation, Line Drawing, Stylization, Video Synthesis, Audio-driven Generation

ACM Reference Format:
Ran Yi, Zipeng Ye, Ruoyu Fan, Yezhi Shu, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. 2022. Animating Portrait Line Drawings from a Single Face Photo and a Speech Signal. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3528233.3530720

*Corresponding author.

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

ACM ISBN 978-1-4503-9337-9/22/08.
https://doi.org/10.1145/3528233.3530720

1 INTRODUCTION

Animating a single face photo is an important problem in computer vision and graphics, which finds many applications in film production, virtual avatars and social media (e.g., [Kemelmacher-Shlizerman et al. 2014]). Compared to real faces, artistic line drawing facial animation can evoke different human experiences (e.g., [Luo et al. 2006]), leading to stronger visual effect by removing

the distraction of irrelevant details, enabling new interaction and entertainment applications. However, audio-driven artistic talking drawing video generation from a normal face photo has not been explored much so far.

In our study, we pay attention to face portrait line drawings which are a highly abstract and expressive style. Compared to cluttered styles (such as oil painting and cartoon), portrait drawings only use a sparse set of line stroke elements. We explore the problem of animating portrait drawings from a single face photo and a speech signal (Fig. 1). This problem is challenging since it involves two tasks: (1) audio-driven prediction of changes in facial expression, and (2) domain transfer from photos to artistic drawings.

Directly concatenating realistic talking face generation methods and domain transfer methods (applied independently frame by frame) will cause severe inter-frame discontinuity problems. For the target artistic portrait line drawing style, this problem is even more disastrous because the appearance and disappearance of the sparse line elements are more visible than discontinuous changes of colors (e.g., in other art styles).

Existing talking face generation methods have explored the idea of generating facial landmarks from speech [Chen et al. 2019; Zhou et al. 2020]. Given an audio-synchronized landmark sequence, [Zhou et al. 2020] used warping to animate a single portrait image. However, since every frame is warped from the input portrait image, the temporal coherence is overly strong, which makes the generated video look unnatural and fake under large head movements. As we will later show, such unnaturalness becomes more obvious when applying a similar approach to animating line drawings.

In this paper, we propose a novel method to generate a talking portrait line drawing video from a single face photo and a speech signal. First, we use an audio translation network to predict facial landmark movements from the input speech signal. Then we propose a novel generative adversarial network (GAN) model that simultaneously achieves domain transfer (from photo to drawing) and facial geometry change (according to facial landmarks). The inputs to the proposed generator G include a face photo p, p's facial landmarks l_p and the target facial landmarks l_t. The output is a portrait line drawing $G(p, l_p, l_t)$ with the same identity as p and the same landmarks as l_t. To alleviate the burden of collecting real artistic portrait drawing videos for training, we design a novel training scheme for generators in our model such that only static portrait drawing data is needed. To address the inter-frame discontinuity problem, we propose two novel temporal coherence losses.

The main contributions of our work are three-fold: 1) We propose a novel feature warping framework to generate *artistic* and *expressive* talking portrait line drawing video. 2) We conduct both geometry change and artistic style transfer using a single GAN model learning from static portrait drawing data, and ensure temporally coherent animation via two novel coherence loss terms. 3) We propose a scheme for separate processing of foreground and background to avoid inappropriate background warping. Experiments demonstrate that our method is better than baseline methods and can generate high quality artistic talking drawing videos. We extend our method to other artistic styles and generate good results.

2 RELATED WORK

2.1 Photo to portrait line drawing transfer

Transforming a face photo into a portrait line drawing is a challenging problem in artistic style transfer, because the face has high semantic constraints and the target style is highly abstract. Neural style transfer methods [Gatys et al. 2016; Li et al. 2019b] often fail in this task because the Gram matrix is not suitable for the target style which has little texture, and example-guided transfer cannot capture well the characteristics of the line drawings [Yi et al. 2020].

GANs [Goodfellow et al. 2014] have achieved success in many image-to-image translation problems and researchers began to design GAN-based models for photo to portrait line drawing translation. APDrawingGAN [Yi et al. 2019, 2021] learned portrait line drawing generation from paired data of face photos and corresponding line drawings. This method applies a composite GAN model consisting of global and local GANs, with a new distance transform loss and a line continuity loss to improve the line generation quality. To avoid the requirement of paired data, [Yi et al. 2020, 2022] proposed an asymmetric cycle structure in a novel GAN model, which is also capable of generating multi-style portrait line drawings. However, these methods only deal with static portrait line drawing generation. When synthesizing portrait drawings frame-by-frame in real video, serious discontinuities can be clearly observed.

2.2 Speech-driven talking face generation

Speech-driven talking face generation is a challenging cross-modality learning problem, which has attracted increasing attention in recent years. Many methods have been proposed, which take a speech signal and some visual contents of a target person as inputs, and generate a talking face video of the target person. These methods can be classified into two categories: (1) person-specific methods that require plenty of video information of the target person for training [Suwajanakorn et al. 2017; Thies et al. 2020], and (2) general methods for arbitrary persons [Chen et al. 2019; Chung et al. 2017; Vougioukas et al. 2018, 2020; Zhou et al. 2019, 2020].

In the first category, [Suwajanakorn et al. 2017] synthesized talking face video of Obama by learning from hours of Obama's videos of speeches using a recurrent neural network to predict lip motion from audio features. [Thies et al. 2020] proposed a deep neural network based on a latent 3D face model to generate talking face videos from speech, which requires a 2-3 minutes target person video as training data. [Richard et al. 2021] synthesized 3D face animation from a speech signal based on cross-modality disentanglement.

In the second category, the methods learn a general talking face generation model using 2D image-based techniques, which are applicable to an arbitrary target person. Facial landmarks are usually used to control face generation [Ha et al. 2020; Siarohin et al. 2019; Wang et al. 2019; Zakharov et al. 2019], and some methods use landmarks to guide the talking face generation: [Chen et al. 2019] translated an audio sequence to facial landmarks and generated videos conditioned on the landmarks; [Zhou et al. 2020] also used facial landmarks as an internal representation, but applied a morphing-based method to animate a cartoon image. [Chen et al. 2020] modeled head motion and facial expression separately and generated talking face videos with rhythmic head motion predicted from a reference short video. [Zhou et al. 2021] modularized the

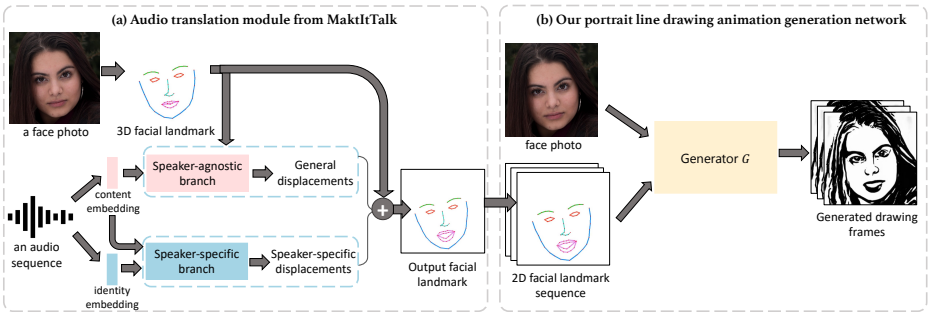

Figure 2: The framework of our method. (a) We use the speech to facial landmark translation module from MakeItTalk [Zhou et al. 2020], which uses two branches (a speaker-agnostic branch and a speaker-specific branch). (b) We propose a GAN model to generate a *temporally coherent* talking line drawing video from the input photo and target 2D facial landmark sequence, with natural and consistent head and hair movement. The face photo is courtesy of Merenda Mattia (Public Domain).

representations into speech content, head pose and identity spaces and generated talking faces with pose controlled by a pose source.

All the above methods only generate face animation based on the input speech, but do not consider domain/style transfer for expressive art forms. In contrast, our method achieves both talking face animation and artistic style transfer in a single GAN model.

3 METHOD

3.1 Overview

Given a single face photo and a speech sequence as inputs, our proposed method generates an expressive *artistic* talking portrait line drawing video that is synchronized with the speech. The problem not only requires facial changes according to speech, but also needs style transfer from normal photos to artistic line drawings.

To solve the problem, we use facial landmarks to bridge the gap between audio and the talking portrait line drawings. The framework of our method consists of two stages (Fig. 2).

Stage 1. We first use the speech to landmark translation module from MakeItTalk [Zhou et al. 2020] to predict facial landmarks from the input speech sequence. The inputs are a speech signal and the 3D facial landmark positions detected from a face photo. The outputs are displacements for the facial landmarks. The module extracts deep representations of content and speaker identity from the audio, and uses a speaker-agnostic branch and a speaker-specific branch to predict landmark displacements. By taking the x, y coordinates of the 3D landmark positions, we can easily get the 2D landmarks to provide input for the next stage. Details are presented in Sec. 3.2.

Stage 2. We design a GAN model to generate artistic talking portrait line drawings from the input face photo and target 2D facial landmarks. The generator needs to achieve both facial geometry changes (according to landmark positions) and style transfer (from photo to artistic drawing). In addition to content loss, identity preserving loss, adversarial loss and geometry loss that constrains facial changes, we propose two novel temporal coherence losses to improve the temporal coherence of the output video. To alleviate the burden of collecting real artistic portrait drawing video, we train the generator using only unpaired photos and static portrait drawing data. Details are presented in Sec. 3.3.

3.2 Speech to facial landmark translation

To predict facial landmarks from a given speech, we use the speech to facial landmark translation module from MakeItTalk [Zhou et al. 2020], which predicts 3D facial landmark displacements from an input audio using disentangled content embedding and speaker embedding of the audio. First, 3D facial landmarks $q \in \mathbb{R}^{68\times 3}$ are detected from a face photo using a 3D landmark detector [Bulat and Tzimiropoulos 2017]. Then the module uses two branches to predict the landmark changes, one speaker-agnostic branch, and one speaker-specific branch (Fig. 2(a)). It extracts speaker-agnostic content representation and speaker identity embedding from the input speech signal. The speaker-agnostic content animation branch takes the content embedding as inputs, and estimates general facial landmark displacements. The speaker-specific animation branch takes the content and identity embedding as inputs, and predicts speaker-specific facial landmark displacements. The final landmark displacements are the sum of predicted displacements from the two branches. Using this module [Zhou et al. 2020], the final landmark displacements contain both general speaker-agnostic displacements predicted from the audio content embedding only, and speaker-specific displacements that also consider speaker identity.

3.3 Talking portrait drawing video generation

To simultaneously achieve style transfer and facial geometry changes, we design a GAN model (Fig. 3) to transform a face photo and a target 2D landmark into a portrait line drawing, which has the same facial geometry as the target landmarks and the same identity as the input face photo. The generation model needs to solve two challenges: (1) simultaneous facial geometry deformation and artistic style transfer; (2) generating temporally coherent artistic portrait drawing video by learning from static portrait drawing data.

Denote the face photo domain as $\mathcal{P}$, the 2D landmark domain as $\mathcal{M}$, and the artistic portrait line drawing domain as $\mathcal{D}$. We learn a mapping function Φ that maps from $(\mathcal{P}, \mathcal{M})$ to $\mathcal{D}$, using a set of real face photos $S(p) = \{p_i | i = 1, 2, \cdots, n_p\} \subset \mathcal{P}$ and a separate set of artistic drawings $S(d) = \{d_i | i = 1, 2, \cdots, n_d\} \subset \mathcal{D}$ as training data. Our model consists of 1) a generator G (transforming a photo to a deformed drawing conditioned on target landmarks), 2) a global discriminator D_g and 3 local discriminators D_{ln}, D_{le}, D_{ll} (discriminating the regions of the generated drawings from real

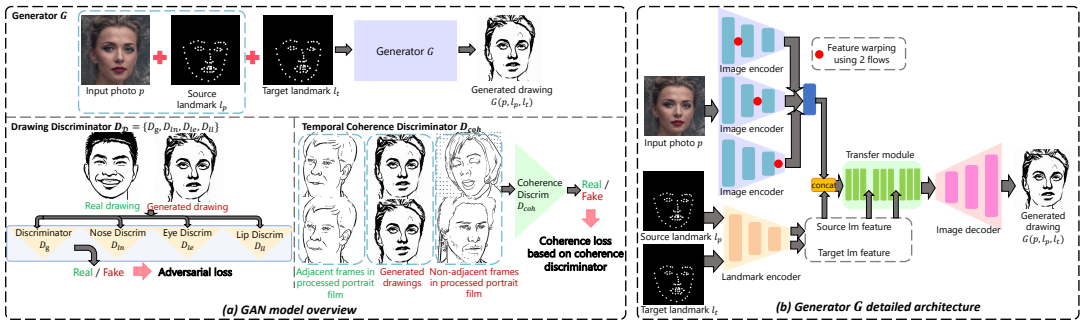

Figure 3: We propose a GAN model to generate temporally coherent and expressive talking portrait drawings from a single photo and a 2D landmark sequence. (a) The model consists of a photo-to-drawing generator G, a global discriminator D_g, local discriminators D_{ln}, D_{le}, D_{ll}, and a temporal coherence discriminator D_{coh}. (b) The detailed architecture of generator G, which performs both facial geometry changes and style transfer. The face photo is courtesy of Sergey Kochkarev (Public Domain).

artist drawings), and 3) a temporal coherence discriminator D_{coh} (discriminating whether two drawings are temporally coherent).

3.3.1 Architecture. Generator G takes a face photo p, the facial landmarks l_p in p (encoded as a binary image where the background is black and the landmarks are drawn as white dots), and target 2D facial landmarks l_t as input, and outputs a portrait line drawing $G(p, l_p, l_t)$ whose facial geometry is consistent with l_t while the identity is similar to p. G is designed with a *feature space warping* strategy to deform the facial geometry and consists of five parts:

(1) An image encoder consisting of one flat convolution and two down convolution blocks, which encodes face photos and extracts necessary image features.
(2) A landmark encoder consisting of one flat convolution and two down convolution blocks (with fewer channels than the image encoder), which encodes landmarks and extracts landmark features.
(3) A feature warping module which warps the image feature according to flow map. We warp with two flow maps (first warping with each separately, and then concatenating the warped features): given two 2D landmarks detected by [Chen 2021], a) flow map A is predicted by a flow regression network [Li et al. 2019a], which takes two landmarks as input and outputs a flow map, and the training data is synthesized by fitting a 3D face model to a photo pair and projecting them to 2D to compute the ground-truth flow map, and b) flow map B is calculated by interpolating the position difference between two sparse landmarks. Flow map A is predicted by a flow regression network trained using a 3D face model, which is more accurate, but is unconstrained for regions outside the face. Flow map B complements it to better cope with regions outside the face. Each time, the feature warp is applied once after one convolution layer of the image encoder, and we repeat the process for each convolution layer, as shown in Fig. 3(b). This essentially applies feature warping at different scales, which better aligns face image features to the target pose. The outputs of the 3 image encoders are merged using a convolution layer. See the appendix for an ablation study to validate our design.
(4) A transfer module consisting of nine residual blocks [He et al. 2016], which combines warped image features and landmark features and transfers the feature from the source to the target domain. the landmark features are concatenated with the image features 3 times, before the 1st, 4th and 7th residual blocks, respectively.
(5) An image decoder consisting of two up convolution blocks and a final convolution, which reconstructs the final output drawing. See appendix for full details.

Drawing discriminators $D_g, D_{ln}, D_{le}, D_{ll}$. The drawing discriminators discriminate generated drawings from real ones. We design a global drawing discriminator D_g to extract full image features and decide whether a full drawing is real. Following [Yi et al. 2020], we further design three local discriminators D_{ln}, D_{le}, D_{ll} to discriminate the nose, eye and lip parts of the drawings respectively, in order to enhance the generation quality of these important features. $D_g, D_{ln}, D_{le}, D_{ll}$ adopt the PatchGAN [Isola et al. 2017] structure, with three down convolution blocks and two flat convolution blocks. The discriminators output probability maps of real/fake for patches.

Temporal coherence discriminator D_{coh}. The temporal coherence discriminator discriminates whether two drawings are temporally coherent. The input to D_{coh} is the concatenation of two drawings. D_{coh} also adopts the PatchGAN structure with three down convolution blocks and two flat convolution blocks. It outputs the probability maps of temporal coherency for two corresponding patches in the input drawings. In order to train the temporal coherence discriminator D_{coh}, we make use of portraits from an artistically stylised film, which is transferred into line drawings by a line extractor [HAT 2018][1] (Sec. 4.1), and feed 3 types of inputs, i.e., 1 kind of real samples and 2 kinds of fake samples: 1) line drawings in adjacent frames of the portrait film as real samples, 2) two generated drawings conditioned on two slightly changed 2D target landmarks as fake samples, 3) line drawings in non-adjacent frames of the film as fake samples.

3.3.2 Foreground extraction and blending strategy. Since the geometry change is guided by facial landmarks, some undesirable background warping effects are likely to occur following the current design. However, the background should remain static before and after the warping. To avoid unwanted background warping, we process the foreground part and the background part separately.

[1]The processed film portraits are only an approximation to the target portrait line drawings, and are only used for training the temporal coherence discriminator, but not for any generator.

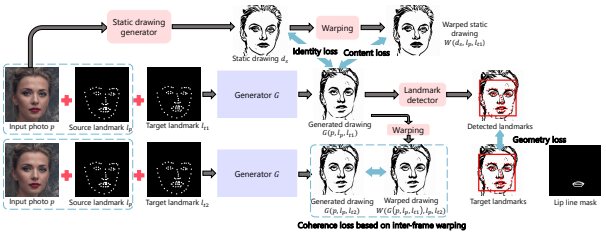

Figure 4: Visualization of the loss calculation process. The face photo is courtesy of Sergey Kochkarev (Public Domain).

Given an input face photo p, we first extract foreground mask m_p using MODNet [Ke et al. 2020]. We then feed the foreground image of the face photo p_f into the generator G, and predict the line drawing corresponding to the target landmarks, i.e. $d_o = G(p_f, l_p, l_t)$. For the background image p_b, we apply the method [Yi et al. 2020] to generate a *static* portrait line drawing d_b.

The final output drawing is the blended result of d_o and d_b. Since the foreground mask changes after the facial geometry changes, we predict the foreground of d_o by warping the foreground mask m_p with the flow map predicted in the feature warping module in Sec. 3.3.1. Denoting the warped foreground mask as m'_p, the blended result is calculated as:

$$d = d_o \cdot m'_p + d_b \cdot (1 - m'_p) \tag{1}$$

3.3.3 Loss functions. Our loss function consists of six loss terms.

Adversarial loss. The adversarial loss measures the discrimination ability of drawing discriminators. Denote the set of global and local discriminators as $D_{\mathcal{D}}$. The adversarial loss is formulated as

$$\begin{aligned} L_{adv}(G, D_{\mathcal{D}}) = &\sum_{D \in D_{\mathcal{D}}} \mathbb{E}_{d \in S(d)}[\log D(d)] \\ &+ \sum_{D \in D_{\mathcal{D}}} \mathbb{E}_{p \in S(p)}[\log(1 - D(G(p, l_p, l_t))] \end{aligned} \tag{2}$$

where l_p contains the facial landmarks of p, l_t includes target landmarks randomly selected from artistic drawing landmarks. G and $D_{\mathcal{D}}$ are trained in an adversarial manner (i.e., G minimizes this loss and $D_{\mathcal{D}}$ maximizes this loss) to direct the generator towards generating drawings closer to real drawings.

Warping-based content loss. To constrain the content in generated drawings, we use a synthetic drawing to approximate the ground truth. We first apply the method [Yi et al. 2020] to generate a *static* portrait line drawing d_s for the input photo p. Then we apply the differential image warping module W [2] based on sparse landmarks to warp d_s into the target landmarks' geometry. The warped static drawing $W(d_s, l_p, l_t)$ has the same geometry as l_t and the same identity as p, and thus could serve as an approximate ground truth. The content loss is formulated as

$$L_{content}(G) = \mathbb{E}_{p \in S(p)}[||W(d_s, l_p, l_t) - G(p, l_p, l_t)||_1] \tag{3}$$

Note that the warped drawing may contain artifacts since the warping flow is estimated based on sparse landmarks (see warped drawings in Fig. 4). Thus this loss alone is not sufficient for good results.

Geometry loss. We further use landmarks to constrain the geometry of generated drawings. We use a landmark detector [Chen 2021](which uses MTCNN [Zhang et al. 2016] to detect face and then MobileFaceNet [Chen et al. 2018b] to detect landmark) to extract the 2D coordinates of the 68-point landmarks of generated drawings. Denote the 2D coordinates of the target landmarks l_t as $x(l_t)$, and the landmark detector as R_{land}. This loss term is

$$L_{geom}(G) = \mathbb{E}_{p \in S(p)}[||x(l_t) - R_{land}(G(p, l_p, l_t))||_2] \tag{4}$$

where both coordinates are normalized into the range $[0, 1]$.

To generate better lines for the lip region, we define a lip line mask M_{lip_line} (see the example in Fig. 4) from the lip landmarks and expect the pixels in the lip line mask region of the generated drawing to be close to black (so that the lip contour is drawn). The geometry loss for lip is formulated as

$$L_{geom\_lip}(G) = \mathbb{E}_{p \in S(p)}[||G(p, l_p, l_t) \cdot M_{lip\_line}||_1] \tag{5}$$

Identity preservation loss. We constrain the output drawing to be the same identity as the input photo p by minimizing the distance between identity features of the generated drawing and the static portrait line drawing of p. We use a pre-trained SphereFaceNet [Liu et al. 2017] R_{iden} to extract identity features. The loss term is

$$L_{iden}(G) = \mathbb{E}_{p \in S(p)}[||R_{iden}(d_s) - R_{iden}(G(p, l_p, l_t))||_1] \tag{6}$$

Note that we extract identity features on the static drawing of p instead of directly on p to avoid the influence of different domains.

Temporal coherence loss based on inter-frame warping. For each input photo p, we generate two drawings conditioned on two target landmarks l_{t1} and l_{t2} with slight changes. We use the differential warping module W to warp the first generated drawing $G(p, l_p, l_{t1})$ to the second set of target landmarks l_{t2}, and require that the warped drawing should be similar to the second generated drawing $G(p, l_p, l_{t2})$. This loss term is formulated as

$$L_{coh1}(G) = \mathbb{E}_{p \in S(p)}[||W(G(p, l_p, l_{t1}), l_{t1}, l_{t2}) - G(p, l_p, l_{t2})||_1] \tag{7}$$

Temporal coherence loss based on coherence discriminator. We observe that using the loss term (7) alone is not sufficient to ensure the temporal coherence in generated drawing animation. So we propose one more temporal coherence loss, which is the adversarial loss for the temporal coherence discriminator G_{coh}. To learn from real portrait films, we denote the set of adjacent frame pairs in the processed portrait film as $S(adj)$ and the set of non-adjacent frame pairs as $S(nadj)$. This loss term is formulated as

$$\begin{aligned} L_{coh2}(G, D_{coh}) = &\mathbb{E}_{(d_1, d_2) \in S(adj)}[\log D_{coh}(d_1, d_2)] \\ &+ \mathbb{E}_{p \in S(p)}[\log(1 - D_{coh}(G(p, l_p, l_{t1}), G(p, l_p, l_{t2}))] \\ &+ \mathbb{E}_{(d_3, d_4) \in S(nadj)}[\log(1 - D_{coh}(d_3, d_4))] \end{aligned} \tag{8}$$

For the non-adjacent frame pairs set $S(adj)$ it is hard to determine the temporal extent of multiple similar adjacent frames. Therefore we choose two frames from two different processed film clips, so they are different enough, as shown in Fig. 3.

Finally the overall loss function $L(G, D_{\mathcal{D}}, D_{coh})$ is a combination of the six loss terms and the mapping ϕ is learned by solving:

$$\begin{aligned} &\min_G \max_{D_{\mathcal{D}}, D_{coh}} L(G, D_{\mathcal{D}}, D_{coh}) = L_{adv}(G, D_{\mathcal{D}}) + \lambda_1 L_{content}(G) \\ &+ (\lambda_2 L_{geom}(G) + \lambda_3 L_{geom\_lip}(G)) + \lambda_4 L_{iden}(G) \\ &+ \lambda_5 L_{coh1}(G) + \lambda_6 L_{coh2}(G, D_{coh}) \end{aligned} \tag{9}$$

[2] We use the differential warping module W from [Cole et al. 2017] to warp the image.

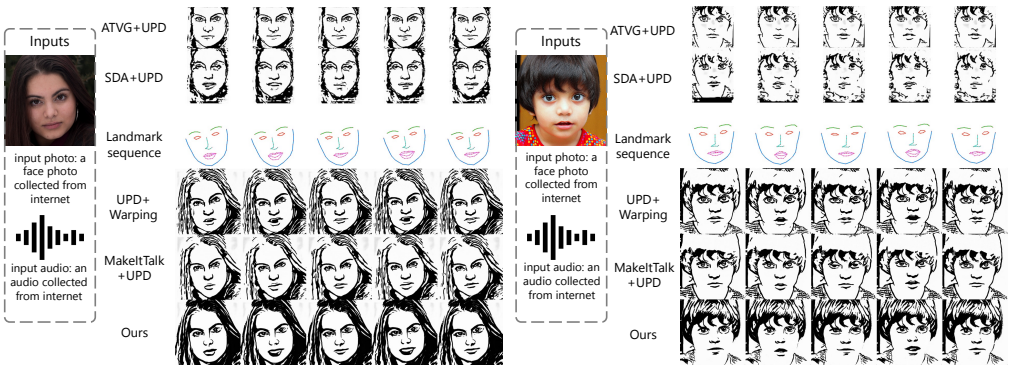

Figure 5: Comparison with baselines. Rows 1-6: ATVG+UPD, SDA+UPD, landmark sequence, UPD+Warping, MakeItTalk+UPD, and Ours. The input face photos are courtesy of Merenda Mattia (Public Domain) and waqas anees (Public Domain).

Table 1: Quantitative evaluation (FID, SSIM, and LMD).

Methods	FID ↓	SSIM ↑	LMD ↓
ATVG+UPD	162.1	0.923	2.74
SDA+UPD	185.2	0.913	2.76
MakeItTalk+UPD	143.0	0.951	**2.64**
UPD+Warp	151.9	**0.975**	**2.64**
Ours	**135.2**	0.974	**2.64**

Table 2: User study results. Each column shows the percentages of four methods selected as the best for each criterion.

Methods	Naturalness	ID preserve	Lip sync	Take-all
ATVG+UPD	1.0%	0.6%	22.7%	1.3%
SDA+UPD	1.5%	0.8%	2.1%	1.1%
UPD+Warp	14.6%	23.7%	12.6%	14.2%
Ours	**82.9%**	**74.8%**	**62.6%**	**83.3%**

4 EXPERIMENTS

We implemented our method in PyTorch. The experiments are performed on a computer with a GeForce RTX 2080 Ti GPU. The parameters in Eq.(9) are empirically set as: $\lambda_1 = 5$, $\lambda_2 = 50$, $\lambda_3 = 50$, $\lambda_4 = 3$, $\lambda_5 = 10$, and $\lambda_6 = 0.5$. The output resolution is 256×256.

4.1 Experiment Setup

Training setting: For the talking line drawing generation network, we collect face photos from both the internet and a facial expression database [Yang et al. 2020], and construct a photo set of 1,603 images. The collected face photos cover different facial expressions, head poses, races, and ages. We also collect portrait line drawings by professional illustrators from the internet. To provide more portrait drawings with intense expressions, we generate some synthetic drawings using the *unpaired portrait drawing* (UPD) method [Yi et al. 2020] and construct a set of 1,451 line drawings (see the second style in UPD and examples shown in the appendix). We further use the portrait film "A Scanner Darkly" to provide samples for the temporal coherence discriminator. The frames in the film are color portraits drawn by artists. We use a line extractor [HAT 2018] to convert them into black and white line drawings to approximate our target style. Some examples of line drawings and processed portrait film frames are presented in Fig. 3, and more are presented in the appendix. We use Adam optimizer with learning rate 0.00005, $\beta_1 = 0.5$, $\beta_2 = 0.999$, and batch size is 1.

Test setting: (1) First, we set the first frame of a real video as the input face photo, and set the audio signal extracted from that video as the input speech signal. We collect 118 video clips (of different identities) from the VoxCeleb2 dataset [Chung et al. 2018] and 18 real videos (each of $7 \sim 35$ seconds length) from the internet. (2) Second, we take an arbitrary face photo and an arbitrary speech signal collected from the internet as the inputs.

4.2 Comparisons

To the best knowledge of the authors, we are the first to generate an *artistic* and expressive talking drawing video from a real face photo and a speech signal. For comparison, we construct three baselines by concatenating existing realistic talking face generation and static portrait line drawing generation methods.

Baseline 1 combines two steps: (1) generating a realistic talking face video from a single face photo and a speech signal, and (2) transferring each frame in the video into a portrait line drawing. We select two state-of-the-art open source methods (ATVG [Chen et al. 2019] and SDA [Vougioukas et al. 2018]) for the first step, and select the UPD method for the second step. The resulting baselines are denoted as ATVG+UPD and SDA+UPD.

Baseline 2 first transfers the input face photo into a portrait line drawing using UPD, and then applies facial reenactment to this drawing by using the speech-to-landmark network and image warping based on landmarks. We denote this baseline as UPD+Warping.

Baseline 3 first generates realistic animation by MakeItTalk, and then transfers it into line drawing style by UPD.

We compare our method and baselines: ATVG+UPD, SDA+UPD, UPD+Warping, MakeItTalk+UPD. Some qualitative results are shown in Fig. 5 and animated results are shown in the accompanying demo video. Inter-frame discontinuities are clearly observed in baseline1 methods (ATVG+UPD, SDA+UPD). This is due to the fact that each video frame is generated separately without considering inter-frame relations. The results of baseline2 (UPD+Warping) often show unpleasant distortion. The overly strong temporal coherence and the distortion effects degrade the naturalness, as there are not even minor variations normally expected with facial dynamics. The baseline 3 (MakeItTalk+UPD) results are less temporally coherent, since UPD stylizes photos without guaranteeing coherence. In comparison, our method generates naturally artistic results with inter-frame continuity, good lip synchronization and identity preservation.

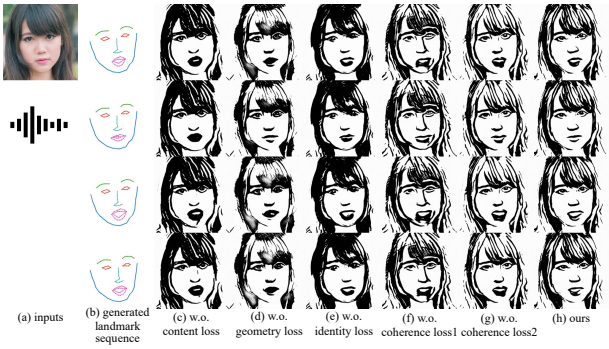

Figure 6: Ablation study: (a) inputs, (b) landmark sequence, (c-g) results w/o content loss (c), geometry loss (d), identity loss (e), temporal coherence loss1 (f) and loss2 (g), and ours (h). The input photo by GRAPHISLIMITED (Public Domain).

4.3 Quantitative evaluation

Metric evaluation. We conduct quantitative evaluation using three metrics: (1) FID [Heusel et al. 2017] for individual frame quality and similarity between the generated and real distributions; (2) inter-frame SSIM for inter-frame coherency evaluation; (3) lip landmark distance (LMD) [Chen et al. 2018a] for lip synchronization. We evaluate ours and the comparison methods on 118 video clips from the VoxCeleb2 dataset [Chung et al. 2018], and calculate the average score of the metrics. Quantitative results are summarized in Table 1. The results show that ours, MakeItTalk+UPD and UPD+Warping are the best on LMD, ours and UPD+Warping are the best on SSIM, and ours is the best on FID. See appendix for full details.

User study. Due to the subjective nature of the portrait line drawings, we conducted a user study to compare our method with three baseline methods, i.e., ATVG+UPD, SDA+UPD, UPD+Warping. We randomly chose 18 real videos from the VoxCeleb2 dataset and the internet, and used their first frame and audio tracks to generate 18 talking drawing videos for each method. Participants were shown the input face photos and four generated drawing videos side by side (using a randomized order to avoid bias), and they were required to answer four questions: 1) which result is the best in terms of head movement and naturalness; 2) which result is the best in terms of line drawing quality and identity preservation; 3) which result is the best in terms of audio synchronization; 4) which result is the best by considering all the factors in the previous three questions. 57 participants finished the user study and 4,104 responses were collected in total. Results of each method being selected as the best under the criterion in each question are summarized in Table 2. Our method achieves the best performance in all evaluation criteria.

4.4 Ablation study

We perform an ablation study on the key factors in the proposed method: (1) content loss, (2) geometry loss, (3) identity preservation loss, (4) temporal coherence loss1 based on inter-frame warping, (5) temporal coherence loss2 based on the temporal coherence discriminator, and (6) foreground-background separation scheme.

As shown in Fig. 6, without the content loss, the semantic information in local facial regions is quite different from the input photo. Without the geometry loss, the lip movements are small and lip synchronization is poor. Without the identity loss, the generated faces look less similar to the input face. Without the two coherence losses, the generated results have worse inter-frame continuity than ours. As shown in Fig. 7, without the foreground-background separation scheme, the background is warped and generates worse results: as shown in Fig. 7(c), the text on the whiteboard is warped in the results without the foreground-background separation, while in our results (Fig. 7(d)) the background text remains static.

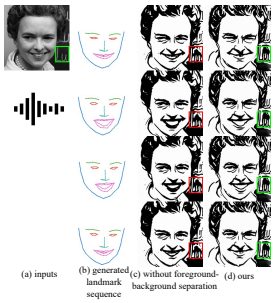

Figure 7: Ablation study: (a) inputs, (b) landmark sequence, (c) results w/o foreground-background separation (background text moves), and (d) Ours (background text static). The photo by Smithsonian Institution (Public Domain).

4.5 Extension to other artistic styles

Our method can be easily extended to artistic portrait animation of different styles. To demonstrate this, we further extend our method to a cartoon style, by training on cartoon images collected from the internet. An example is shown in Fig. 1. More results, details and more styles are in the appendix and demo video.

5 CONCLUSION AND LIMITATION

In this paper, we propose a novel method for generating an expressive artistic talking line drawing video from a single face photo and a speech signal. Our method (1) achieves both facial geometry change and artistic style transfer in a single GAN model, (2) ensures temporally coherent generation via two novel coherence loss terms, and (3) proposes a scheme for separate processing of foreground and background to avoid unnatural background warping. Experiments demonstrate that our method is better than three baseline methods and can generate high quality artistic talking line drawing videos. Our proposed method can be easily extended to other styles, e.g. cartoon style, and generate high quality results. There are still some limitations: our method does not perform well when the input is blurry or with exaggerated expression, or the foreground separation is inaccurate, which we will improve in future work.

ACKNOWLEDGMENTS

This work was supported by National Key Research and Development Program of China (2019YFC1521104), the Natural Science Foundation of China (61725204, 72192821, 61972157), Tsinghua University Initiative Scientific Research Program, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Science and Technology Commission of Shanghai Municipality (21511101200), Shanghai Sailing Program (22YF1420300), and Art major project of National Social Science Fund (I8ZD22).

REFERENCES

Adrian Bulat and Georgios Tzimiropoulos. 2017. How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230, 000 3D Facial Landmarks). In *IEEE International Conference on Computer Vision (ICCV)*. 1021–1030.

Cunjian Chen. 2021. PyTorch Face Landmark: A Fast and Accurate Facial Landmark Detector. https://github.com/cunjian/pytorch_face_landmark

Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. 2020. Talking-Head Generation with Rhythmic Head Motion. In *16th European Conference, Glasgow (ECCV) (Lecture Notes in Computer Science, Vol. 12354)*. Springer, 35–51.

Lele Chen, Zhiheng Li, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. 2018a. Lip Movements Generation at a Glance. In *15th European Conference, Munich (ECCV)*, Vol. 11211. 538–553.

Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. 2019. Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7832–7841.

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018b. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In *Biometric Recognition - 13th Chinese Conference (CCBR) (Lecture Notes in Computer Science, Vol. 10996)*. 428–438.

Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. 2017. You said that?. In *British Machine Vision Conference (BMVC 2017)*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 1086–1090.

Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T. Freeman. 2017. Synthesizing Normalized Faces from Facial Identity Features. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 3386–3395.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2016. Image Style Transfer Using Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2414–2423.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems (NeurIPS)*. 2672–2680.

Sungjoo Ha, Martin Kersner, Beomsu Kim, Seokjun Seo, and Dongyoung Kim. 2020. MarioNETte: Few-Shot Face Reenactment Preserving Identity of Unseen Targets. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*. 10893–10900.

Project HAT. 2018. Line Distiller. https://github.com/hepesu/LineDistiller Accessed: 2021-12-14.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems, (NeurIPS)*. 6629–6640.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976.

Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson W.H. Lau. 2020. Is a Green Screen Really Necessary for Real-Time Portrait Matting? *CoRR* abs/2011.11961 (2020), 1–11.

Ira Kemelmacher-Shlizerman, Eli Shechtman, Rahul Garg, and Steven M. Seitz. 2014. Moving portraits. *Commun. ACM* 57, 9 (2014), 93–99.

Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. 2019b. Learning Linear Transformations for Fast Image and Video Style Transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3809–3817.

Yining Li, Chen Huang, and Chen Change Loy. 2019a. Dense Intrinsic Appearance Flow for Human Pose Transfer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3693–3702.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. SphereFace: Deep Hypersphere Embedding for Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6738–6746.

Yuan Luo, M.L. Gavrilova, and M.C. Sousa. 2006. NPAR by Example: Line Drawing Facial Animation from Photographs. In *International Conference on Computer Graphics, Imaging and Visualisation*. 1–8.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 1153–1162.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems (NeurIPS)*. 7135–7145.

Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: learning lip sync from audio. *ACM Trans. Graph.* 36, 4 (2017), 95:1–95:13.

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural Voice Puppetry: Audio-Driven Facial Reenactment. In *16th European Conference, Glasgow (ECCV)*. 716–731.

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2018. End-to-End Speech-Driven Facial Animation with Temporal GANs. In *British Machine Vision Conference (BMVC)*. 133.

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2020. Realistic Speech-Driven Facial Animation with GANs. *Int. J. Comput. Vis.* 128, 5 (2020), 1398–1413.

Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Bryan Catanzaro, and Jan Kautz. 2019. Few-shot Video-to-Video Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5014–5025.

Tao Yang, Zeyun Yang, Guangzheng Xu, Duoling Gao, Ziheng Zhang, Hui Wang, Shiyu Liu, Linfeng Han, Zhixin Zhu, Yang Tian, Yuqi Huang, Lei Zhao, Kui Zhong, Bolin Shi, Juan Li, Shimin Fu, Peipeng Liang, Michael J. Banissy, and Pei Sun. 2020. Tsinghua facial expression database – A database of facial expressions in Chinese young and older women and men: Development and validation. *PLoS ONE* 15, 4 (2020), e0231304.

Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. 2019. APDrawingGAN: Generating Artistic Portrait Drawings From Face Photos With Hierarchical GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10743–10752.

Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. 2020. Unpaired Portrait Drawing Generation via Asymmetric Cycle Mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8214–8222.

Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. 2022. Quality Metric Guided Portrait Line Drawing Generation from Unpaired Training Data. *IEEE Trans. Pattern Anal. Mach. Intell.* (2022). https://doi.org/10.1109/TPAMI.2022.3147570

Ran Yi, Mengfei Xia, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. 2021. Line Drawings for Face Portraits from Photos using Global and Local Structure based GANs. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 10 (2021), 3462–3475.

Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 9458–9467.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* 23, 10 (2016), 1499–1503.

Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. 2019. Talking Face Generation by Adversarially Disentangled Audio-Visual Representation. In *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI 2019)*. 9299–9306.

Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4176–4186.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. 2020. MakeItTalk: Speaker-Aware Talking Head Animation. *ACM Trans. Graph.* 39, 6, Article 221 (2020), 15 pages.