Warp-Guided GANs for Single-Photo Facial Animation

JIAHAO GENG, TIANJIA SHAO, and YOUYI ZHENG, State Key Lab of CAD&CG, Zhejiang University
YANLIN WENG and KUN ZHOU, Zhejiang University and ZJU-FaceUnity Joint Lab of Intelligent Graphics, China

Fig. 1. Given single-view portrait photos, our method automatically generates photo-realistic facial animations that closely match the expressions in the driving frames (shown in smaller figures). A realtime demo is shown on the right. From left to right, top to bottom, original photos courtesy of Pedro Haas, Getty Images, Jevgeni Kurnikov, and Universal Studios Licensing LLC.

This paper introduces a novel method for realtime portrait animation in a single photo. Our method requires only a single portrait photo and a set of facial landmarks derived from a driving source (e.g., a photo or a video sequence), and generates an animated image with rich facial details. The core of our method is a warp-guided generative model that instantly fuses various fine facial details (e.g., creases and wrinkles), which are necessary to generate a high-fidelity facial expression, onto a pre-warped image. Our method factorizes out the nonlinear geometric transformations exhibited in facial expressions by lightweight 2D warps and leaves the appearance detail synthesis to conditional generative neural networks for high-fidelity facial animation generation. We show such a factorization of geometric transformation and appearance synthesis largely helps the network better learn the high nonlinearity of the facial expression functions and also facilitates the design of the network architecture. Through extensive experiments on various portrait photos from the Internet, we show the significant efficacy of our method compared with prior arts.

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Neural networks**; *Animation*; *Image processing*;

Additional Key Words and Phrases: Portrait animation, expression transfer, generative adversarial networks

*Corresponding authors: Tianjia Shao (tianjiashao@gmail.com) and Kun Zhou (kunzhou@acm.org).
Authors' addresses: Jiahao Geng; Tianjia Shao; Youyi Zheng, State Key Lab of CAD&CG, Zhejiang University; Yanlin Weng; Kun Zhou, Zhejiang University , ZJU-FaceUnity Joint Lab of Intelligent Graphics, Hangzhou, 310058, China.

ACM Reference Format:
Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. 2018. Warp-Guided GANs for Single-Photo Facial Animation. *ACM Trans. Graph.* 37, 6, Article 231 (November 2018), 12 pages. https://doi.org/10.1145/3272127.3275043

1 INTRODUCTION

> *Self expression is a vital part of understanding life, and enjoying it to the full.* – Oliver Bowden

Facial expression, one of primary nonverbal communication form, plays a vital role in our daily social interactions. As a highly complex process, facial expressions typically involve the movements of various motions and positions of the muscles beneath the skin. For example, a smile could cause a closing of the eye, an opening of the mouth, and folds around the nasion.

Image is one of the most common visual forms that can carry realistic facial expressions, within which, photo-realistic editing results can be achieved [Averbuch-Elor et al. 2017; Cao et al. 2014a; Thies et al. 2016]. A number of research works have been devoted to facial expression editing, expression synthesis, and facial reenactment [Averbuch-Elor et al. 2017; Garrido et al. 2014; Thies et al. 2016; Vlasic et al. 2005]. Previous work on facial manipulation typically requires an input of a driving video or a video of the target portrait so that the contents in those videos could be either used for 3D reconstruction [Breuer et al. 2008] or borrowed for fine-scale detail synthesis (e.g., creases, wrinkles, and hidden teeth, etc.) [Averbuch-Elor et al. 2017; Thies et al. 2016]. Expression editing in a single

image is also possible [Blanz et al. 2003; Blanz and Vetter 1999; Cao et al. 2014a], but often requires manual initialization or fall short in generating photo-realistic effects [Piotraschke and Blanz 2016]. Recent research of [Garrido et al. 2014] and [Averbuch-Elor et al. 2017] showed that through lightweight 2D warps, highly compelling results can be achieved via the extrapolation of structural fiducial points and subsequent fine-grained detail composition using, e.g., ERI [Liu et al. 2001].

In this paper, we are interested in animating the subject in a single portrait photo captured in a frontal pose with neutral expression, to bring it to life and mimic various expressions in a high realism manner. We aim to imitate the movements of the face in portrait photos with different backgrounds as in [Averbuch-Elor et al. 2017]. To this end, we decouple the process into multiple stages. Our key insight is that while the global structural movements of the facial expression can be well captured by the 2D facial landmarks and preserved via 2D warps, the distribution of fine-scale local details and hidden regions could be naturally synthesized by generative models. More importantly, the 2D warps could factorize out the nonlinear geometric transformations exhibited in the facial expressions and better help the network to focus on the appearance synthesis.

We perform global 2D warp on the target portrait photo by a set of control points imposed on facial and non-facial regions of the portrait. The displacements of these control points are transferred from the motion parameters of the driving source (see details in Section 4). We then extract the facial region and interpolate the 2D facial landmarks to generate a per-pixel displacement map which carries the fine movements of the face under the global 2D warp. In a key stage, the displacement map is fed into a generative adversarial neural network together with the warped face image, to generate a final detail-refined facial image. The network, *wg-GAN* which we term, is end-to-end trained with tons of warped facial images and 2D displacement maps derived from publicly available video datasets. We alter the network structure and loss functions to suit our purpose of fine-scale detail synthesis. Since the derived network might not fully unfold hidden regions such as the inner mouth region, we particularly train another generative adversarial neural network [Iizuka et al. 2017] to inpaint such hidden regions.

We show that the factorization of geometric transformations exhibited in facial expressions through 2D warps largely benefits the design of the generative model to allow it to focus on the local detail synthesis and in the meanwhile eases the network training. Using a learning-based generative model also enables us to bypass the need for imposing requirements on the driving source as well as the need for ad-hoc heuristic algorithms to account for fine-scale detail synthesis. Moreover, the utilization of computing power of GPUs in neural nets also enables our pipeline to operate in real time. We demonstrate the efficacy of our method through extensive experiments on various internet portrait photos as well as two user studies. Our results illustrate a significant improvement over the current state-of-the-art approaches and our method is feasible for a variety of applications such as single-photo portrait animation and facial expression editing.

2 RELATED WORK

Literature in facial manipulation stems from the seminal work of [Blanz and Vetter 1999], where a 3D morphable model is fitted to a single image and texture mapped to enable parametric changes in pose and appearance of the face. While having a 3D morphable model could benefit the subsequent manipulation and enable more faithful 3D reconstruction [Breuer et al. 2008; Piotraschke and Blanz 2016], these techniques often fall short in achieving the realism of the manipulated faces at the fine-scale details as these features cannot be fully spanned by the principal components [Averbuch-Elor et al. 2017], not even with multiple images [Vlasic et al. 2005].

Having a video of the target face and a driving video of the source can largely alleviate this problem as the contents such as fine-scale details can be either inherited from the target sequence or borrowed from the source [Mohammed et al. 2009]. This leads to a series of research works which utilize an input video or a performance database of the target face. For example, Vlasic et al. [2005] use a 3D morphable model to drive the facial expression in a video by editing the expression parameters while Dale et al. [2011] use it for face reenactment and later on Garrido et al. [2014] present a method for facial replacement in a video. Li et al. [2012] use a facial performance database of the target face. The work of [Thies et al. 2016] introduces a real-time framework for face reenactment in a video where they also assume the target video carries rich and sufficient data to synthesize the facial details.

There are also a number of works in facial manipulation which have their particular focuses. For example, [Fried et al. 2016] introduce a method to manipulate the camera viewpoint from a single input facial image. The work of [Kuster et al. 2012] and [Ganin et al. 2016] show their interests in manipulating the gaze of a single image. Other works, such as [Garrido et al. 2015], focus on transferring lip motion to an existing target video and [Blanz et al. 2003; Kawai et al. 2013, 2014] focus on the realism of the mouth region. Facial manipulation techniques have also been introduced for purposes of data augmentation [Masi et al. 2016], magnifying (or suppressing) expressions [Yang et al. 2012], removing large-scale motion [Bai et al. 2013], or face frontalization [Hassner et al. 2015], where local edits are commonly performed without significant changes in facial expressions such as those in our cases.

Our method requires face tracking for expression transfer. There is a line of research works in 3D facial performance capture and animation from monocular RGB cameras [Cao et al. 2015, 2014a, 2013, 2016; Shi et al. 2014; Wang et al. 2016], video-audio [Liu et al. 2015], and depth cameras [Bouaziz et al. 2013; Hsieh et al. 2015; Li et al. 2013; Weise et al. 2011]. Through performance capture, one can calculate the rigid head transformation and the non-rigid facial expressions and subsequently use them to transfer the expressions to the target. As their main focus is 3D animation, manual interactions are often required when going back to 2D to generate photo-realistic edits [Cao et al. 2014b].

Recently, deep neural networks have been extensively exploited towards facial manipulation and expression synthesis [Ding et al. 2018; Korshunova et al. 2017; Olszewski et al. 2017; Qiao et al. 2018; Song et al. 2017; Yeh et al. 2016]. Among them, [Yeh et al. 2016] introduce a variational autoencoder to learn expression flow maps

[Yang et al. 2011] for facial manipulation; [Korshunova et al. 2017] introduce a CNN-based framework for face swapping in analogy to image style transfer [Li and Wand 2016]. Other works leverage the generative models (see a pioneer work in [Susskind et al. 2008]) to handle fine-scale details such as wrinkles and teeth [Olszewski et al. 2017], or use expression code to condition on the generative model [Ding et al. 2018]. Recent works of [Qiao et al. 2018; Song et al. 2017] utilize geometric facial landmarks to guide the network to control the facial details synthesis, we show that such geometric cues could significantly boost the network performance when coupled with global and local transformations. Most of the aforementioned methods are applicable within particular cropped facial regions do not handle other regions in the image. Trivial stitching algorithms such as in [Korshunova et al. 2017] will not work if one turns his head and leaves distorted regions between the background and the head. A most recent work of [Averbuch-Elor et al. 2017] addresses this issue by taking confidence-aware lightweight 2D warps in both body and head region followed by procedures of fine-scale details transfer and hidden region hallucination. Unitizing the 2D structural fiducial points helps their method to bypass the need for precise tracking procedure such as the ones presented in [Thies et al. 2015] and [Cao et al. 2015]. Our method fits in by taking advantages of such global 2D transformation and leverages it for the details synthesis and hidden region hallucination with generative adversarial models.

In a concurrent work of [Kim et al. 2018], they introduce a generative neural network to predict photo-realistic video frames from synthetic renderings of a parametric face model. Their method achieves compelling results and is able to generate a full 3D head motion and eye gaze to the target portrait. However, their method requires a target video as input and the adversarial network they propose is target-oriented and needs to be retrained for a new target subject. In contrast, our method requires only a single portrait photo and our generative network is generic which can be applied to any target once trained.

Our method exploits the recent technique of generative adversarial networks (GANs), which was originally designed for generating visually realistic images [Goodfellow et al. 2014]. We found its potential power in learning the semantic fine-grained facial details when incorporated with geometric guidance and unsupervised learning. Through extensive experiments with alternative network structures such as those in [Ding et al. 2018; Qiao et al. 2018; Song et al. 2017], we show the superiority of our fine-grained adversarial model for our particular purpose of facial detail refinement.

3 OVERVIEW

Our method takes as input a single target portrait photo with the face in the neutral-frontal pose. Our aim is to animate the portrait and make the subject express various emotions as in our daily life. There are a few challenges we need to address. First, as facial expression involves complex and nonlinear geometrical transformations, we need a model to capture the global transformations of major facial components such as eyes, mouth, and nose since our humans are highly sensitive to the subtle variations in these regions. Second, to achieve photo-realistic results, we need to seamlessly transfer all fine-scale facial details such as wrinkles, creases, and self-shadows. Third, realistic animation of a portrait normally involves movements in the head and the body. Thus, our algorithm should be able to handle the motion of the head and the upper body adequately well.

To tackle the above mentioned challenges, we formulate the problem into multiple stages: a global stage where we allow the structural facial transformations to be carried by a set of *fiducial points* commonly studied in the literature for various applications (e.g., in [Cao et al. 2015]), and a set of extended feature points to account for head and body movements [Averbuch-Elor et al. 2017]; a local stage where we add back all necessary fine-scaled facial details and remove artifacts brought in by the 2D warps with a generative adversarial neural network, which is trained with cropped facial images and their corresponding 2D displacement maps; and a hidden region hallucination stage where an inpainting generative adversarial neural network is employed to synthesize the inner mouth region (we assume the portrait photo to be taken in neutral expression with the mouth closed [Averbuch-Elor et al. 2017]).

Our paradigm is built upon two key observations: First, the global lightweight 2D warps, although not carrying the full range of facial transformations, can capture well the structural changes of the face expression since essentially the major notable changes involved in facial expressions are in the regions of the mouth, eyes, nose and the silhouettes; Second, those warps partially carry the global and local nonlinearity of the per-pixel transformation functions, which could help the generative networks focus on the details synthesis and avoid the learning of these nonlinear geometric transformations.We consolidate our findings by thoroughly experimenting with alternatives. Below we describe our technical components in details.

4 THE METHOD

Our pipeline is shown in Fig. 2. To enable the global structural manipulation of the portrait, we detect the facial landmarks as well as non-facial feature points in the head and body region as the control points of the portrait similar to [Averbuch-Elor et al. 2017] (Fig. 2(a)). When the control points are modified (e.g., by tracking a source face), our goal is to create novel natural facial expressions.

To transfer the movements of the source landmarks to that of the target, a direct 2D mapping could lead to unnatural expressions if the source and target faces differ dramatically (e.g., a fat face v.s. a thin one). We use the algorithm of DDE [Cao et al. 2014a] to track the face of the source person, which simultaneously detects the 2D facial landmarks and recovers the 3D blendshapes, as well as the corresponding expression and the 3D pose [Cao et al. 2014a]. We also use DDE to recover these initial properties for the target image. We then transfer the expression and 3D pose from the source to the target [Thies et al. 2016]. Subsequently, the transformed target 3D facial landmarks on the face mesh are projected onto the target image to get the displaced 2D landmarks. Then we apply the confidence-aware warping to the target image as in [Averbuch-Elor et al. 2017].

Fig. 5 left shows the coarse face image after warping. As noted by [Averbuch-Elor et al. 2017], coarse 2D warps do not fully carry the richness of facial expressions. For example, the wrinkles and creases merited in the facial transformation need to be transferred properly. This is also true for hidden regions such as the inner

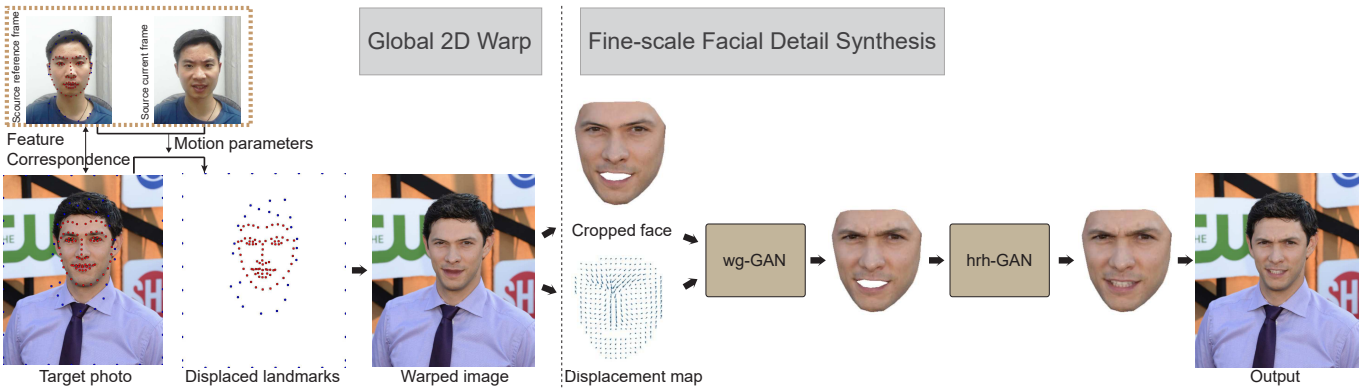

Fig. 2. An overview of our method. Given a single portrait photo and a driving source (either a photo or a video), our method first globally warps the image with tracked feature landmarks. Then the facial region extracted from the warped image, along with a computed displacement map, are fed into a refiner generative adversarial network to generate photo-realistic facial details. The refined face is then fed into another generative adversarial network to hallucinate the hidden regions (i.e., the inner mouth region). Both networks operate in 2D. Finally, the refined face is seamlessly integrated into the warped image to generate the animated result. Original photo courtesy of Pedro Haas.

mouth (Fig. 5 middle). To address these issues, [Averbuch-Elor et al. 2017] proposed algorithms to transfer the wrinkles and creases by expression ratio image (ERI) [Liu et al. 2001] and hallucinate the hidden region directly using the teeth from the source video. These operations could unavoidably introduce undesired artifacts. This is because both the teeth transfer and ERI technique require the source and target faces to merit certain similarities in shape, pose, and fine details, otherwise the transferred details from the source cannot be blended well with the target (e.g., the outlier detection in ERI could fail if the two faces differ too much (Fig. 8, the third row)), leading to undesired artifacts such as unnatural wrinkles, residual shadows, or incompatible teeth (Fig. 8). We handle both of the problems using data-driven approaches to alleviate the artifacts inherited in these heuristics. Since facial detail synthesis and hidden region hallucination are two different tasks, i.e., one is to add details to the face while the other is to fill in the missing parts, learning a unified generative model may bring in additional challenges (see Section 5.1). We train two generative networks for the two tasks separately.

4.1 Refining Warped Face with wg-GAN

Portrait photos are usually taken with a focus on the facial expression but often with diverse background elements such as clothes, environments, light conditions, and hairstyles, which pose potential challenges for a generative model to learn all these convoluted variations. Thus, to avoid the deficiencies in network learning, we focus on the facial regions and leave the 2D global warps [Averbuch-Elor et al. 2017] to delegate the rest transformations.

We exploit a typical conditional GAN framework with a generator network G which we call the *face refinement network* for the purpose of refining a warped face image with fine details and a discriminator network D for the purpose of discerning if a synthesized facial image from G is real or fake. Here, we have a few key issues to consider when designing our network architectures. First, the warped fiducial structure needs to be maintained to mimic the desired expression, i.e., the warped position of the eyes, nose, mouth, etc., since our humans are extremely sensitive to subtle changes in these regions. Second, the generator should be able to generate images that retain high-fidelity of the original face with all necessary details such as wrinkles, creases, and illumination conditions. Finally, the network should be efficient enough to enable real-time performance.

The conditional GAN model consumes a cropped face region which is warped and without the inner mouth region, together with a displacement map which serves as a guidance of the underlying facial global and local transformations (see Fig. 4), and generates a facial image with full synthesized details except the inner mouth region. As for the ground truth images for network training, we crop the real faces and remove the inner mouth region as well. As in [Song et al. 2017], we rectify the face before feeding into the network.

Displacement map generation. The displacement map is an essential element serving in our network design. It carries pixel-wise nonlinear transformations and serves as a condition for the network. We generate our displacement map as follows. For each landmark on the cropped warped image I_w, we compute its offset from the rest pose I. We note that the movements of the major facial components during a facial expression are anisotropic. For example, the displacements at the eyebrow region are typically much smaller than the ones around the mouth. Thus if we use the original values of displacements, the displacement of eyebrows may get ignored by the network. So we normalize the displacements of each semantic parts individually. Specifically, taking the eyebrow for an example, we compute the standard variance of its landmark displacements (the standard variance is computed among all training images) and use it to normalize the displacement vectors of the eyebrow (the normalized values follow a normal distribution). We then interpolate the offsets to the whole image of I_w based on the triangulation of the image using these landmarks (Fig. 4).

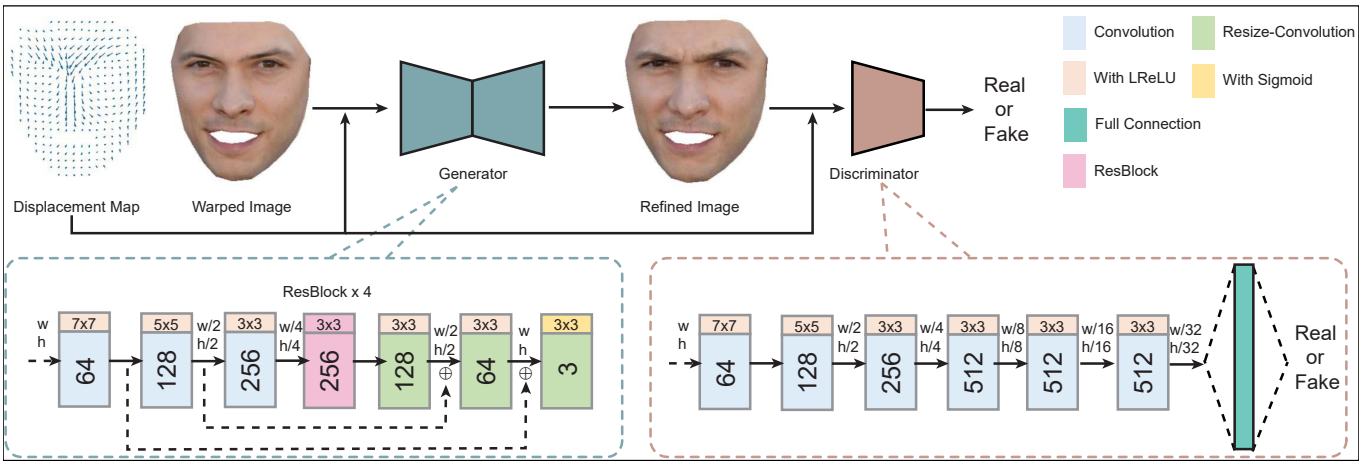

Fig. 3. The architecture of our wg-GAN. It is composed of a facial refinement network (generator) which synthesizes a detail-refined facial image and a discriminator network which discerns whether the refined image is real or fake. Original photo courtesy of Pedro Haas.

The architecture of our wg-GAN network is shown in Fig. 3. We mark the facial region of the input face (determined via landmark points) and discard output pixels that are outside the mask region. The architecture of the *refinement network* follows an encoder-decoder structure. To preserve the network from compressing too much of the information, we only downsample the image to 1/4 of its original resolution (see the first 3 convolutional layers of the *refinement network*). The convoluted images are then sent through four residual blocks [He et al. 2016]. The residual block is commonly used for efficient training of deep structures. Afterward, the output is restored to the original resolution using resize-convolution layers instead of deconvolution layers to avoid the "uneven overlap" problem commonly happening in deconvolution [Gauthier 2014]. We also add skip-connection between the first and last two convolutional layers to preserve the image structure as much as possible [Isola et al. 2017].

The discriminator network takes as input a refined face image or a real face image with the corresponding displacement map and compresses the images through convolutional layers into small feature vectors, which are sent to a fully connected layer to predict a continuous value indicating the confidence of the image being real. It consists of six convolutional layers whose output is flattened to a 32768-dimensional vector (see the bottom of Fig. 3). Similar to the *refinement network*, we use 7×7 and 5×5 kernel sizes in the first two convolution layers respectively while using 3×3 kernels for all the rest layers. All layers in both the *refinement network* and the discriminator network use Leaky Rectified Linear Unit (LReLU) as activation function [Maas et al. 2013] except for the last layer of the *refinement network* where sigmoid function is used to output values in the range of $[0, 1]$, and the last layer of the discriminator network where no activation function is used. We use stride sizes of 2 when downsizing the image and 1 otherwise.

Fig. 4. Displacement map generation. Left: initial displacements of facial landmarks computed from the driving source; middle: adaptively normalized displacements of facial landmarks; right: computed facial displacement map. Original photo courtesy of Pedro Haas.

Loss Function. Let $R(x_w, M)$ denote the *refinement network* in a functional form, with x_w the input warped image and M the displacement map. For the *refinement network*, we use the L_1 loss between the refined face image $R(x_w, M)$ and the ground truth face image x_g as a regularization term that penalizes large changes between the real and refined facial regions:

$$L(R) = \mathbb{E}_{x_w, M, x_g} \|R(x_w, M) - x_g\|_1. \quad (1)$$

For the adversarial loss of discriminative net, we use the loss function of Wasserstein GAN [Arjovsky et al. 2017] for stable training:

$$\min_R \max_D \mathbb{E}_{x_w, M, x_g} [D(x_g, M) - D(R(x_w, M), M)]. \quad (2)$$

The two loss functions are combined together to finally train our wg-GAN:

$$\min_R \max_D \mathbb{E}_{x_w, M, x_g} [\alpha L(R) + D(x_g, M) - D(R(x_w, M), M)], \quad (3)$$

where α is a weighing hyper parameter and is 0.004 in our implementation. To improve the stability of adversarial training, we follow the work of [Shrivastava et al. 2017] to update the discriminator using a history of refined images plus the ones in the current mini-batch.

Training data. Our training data are collected from video sequences. Given a video sequence starting from a rest expression, we detect facial landmarks for every 10th frame and generate the warped image for the frame. Then for each training image, we obtain its ground truth (real frame), warped image, and the displacement

Fig. 5. Left: cropped coarse face image after global 2D warps; middle: face image after the facial refinement using wg-GAN (note the fine details around the eyebrow and the noise); right: result after inner mouth hallucination. Original photo courtesy of Pedro Haas.

map. We gather such training data from public datasets including MMI [Pantic et al. 2005; Valstar and Pantic 2010], MUG [Aifanti et al. 2010], and CFD [Ma et al. 2015]. We find some videos in MMI dataset are captured under undesirable conditions (e.g., side view, changing lighting condition, or rather low resolution), so we select 390 sequences from them as training data, which includes 35 people with 3-20 expressions. For MUG, we gather 329 sequences of 47 people, each with 7 expressions. We also gather 158 people as training data from CFD which typically consists of 4 images for one person with different expressions. Similar to G2GAN [Song et al. 2017], we augment the training data by flipping and random cropping, please refer to [Song et al. 2017] for details.

4.2 Hidden Region Hallucination with hrh-GAN

To fully synthesize a realistic mouth inner region, we take the global-and-local learning-based inpainting approach of [Iizuka et al. 2017]. In their method, a fully convolutional network is designed as generator network to complete images, which is concatenated with two discriminator networks: a global discriminator network to ensure the generated image to be coherent as a whole and a local discriminator network to ensure local patch consistency. We take the same network structure as theirs and employ it for our purpose of inner-mouth synthesis. We call the hidden region hallucination network hrh-GAN for short in our context.

The input of the inpainting network is a cropped face without the inner mouth region, and the network will generate a complete face with teeth and tongue inside the mouth. We use both the training data derived from MMI, MUG, and CFD as well as portrait images collected from the internet. In total, we gather 6211 images for the network training. From those training images, the mask of "hidden" inner mouth region is computed with the detected landmarks. Similar to [Iizuka et al. 2017], the loss function of the network includes a Mean Squared Error (MSE) loss that minimizes per-pixel difference between the generated image and the ground truth image within the mask, a global GAN loss that measures the reality of the whole face and a local GAN loss that measures the reality of the local mouth region (see details in [Iizuka et al. 2017]). We also use flipping and random cropping to augment the training data.

Since our training data size is significantly smaller than that used in [Iizuka et al. 2017], a direct training of 256×256 resolution could lead to unnatural outputs. To alleviate this, we follow the work of [Karras et al. 2017] to train the GAN hierarchically from low resolution to high resolution. Specifically, we first train the hrh-GAN using our training data with the resolution of 128×128. We keep the network architecture the same as [Iizuka et al. 2017] in this step. Then in the second step, we replace the first convolutional layer (128×128×3 → 128×128×64) with three convolutional layers (256×256×3 → 256×256×16 → 256×256×32 → 128×128×64). The output layers are modified similarly while the intermediate layers are kept the same. Finally, the whole network is fine tuned to adapt to the resolution of 256×256. We find this adaption works well in practice. Fig. 5 right shows the results of inner-mouth hallucination.

5 EXPERIMENTAL RESULTS

We implement our algorithm in Python and Cuda on a desktop with an Intel Core i7-4790 CPU (3.6 GHz) and a GeForce 1080Ti GPU (11GB memory). We use the TensorFlow framework [Abadi et al. 2016] for our network implementation.

Transferring the expression from a source frame of 640×480 pixels to a target image of 640×480 pixels takes about 55 milliseconds in total, where the warping takes about 12 milliseconds, the refining net takes about 11 milliseconds and the inpainting net takes about 9 milliseconds. The rest computations are mainly with the face tracking and CPU-GPU communication. We do not require any precomputation, thus our algorithm can be used for real-time facial reenactment. Please refer to supplementary videos for our real-time demo.

We evaluate our algorithm on challenging internet portraits obtained from Flickr and Unsplash covering people with different genders, ages, skin colors and backgrounds (Fig. 13). The source expressions are obtained from the video sequence in the MMI Facial Expression Database, MUG Database, and our own captured data. In Fig. 13 we demonstrate some sample source driving frames and the corresponding transferred expressions. Please refer to our supplementary video for full results. The results show that our method is able to consistently produce natural expressions in analogy to the source frame for all the tested examples.

For better understanding our algorithm pipeline, we also illustrate intermediate results of different stages in our pipeline (see Fig. 14). It can be observed that the lightweight 2D warps are indeed able to provide a good global structural transformation for the deformed face while ensuring consistency with the backgrounds. Our refinement network then effectively reduces the artifacts from warping and adds more natural details. Finally, the inpainting network hallucinates compatible teeth and tongue for the new expression.

5.1 Evaluation

Evaluation on warping and displacement map. To illustrate the effectiveness of our warp-guided network architecture, we experiment with various alternative solutions. Specifically, we examine how the results could degenerate if we do not perform the warp on the facial image or do not use the displacement map. Let W denote warped face image, and M denote displacement map. We alternatively train our network by switching on/off between W and M by preparing different training inputs. For example, $[W-, M+]$ means training with unwarped face image and displacement map while $[W+, M-]$

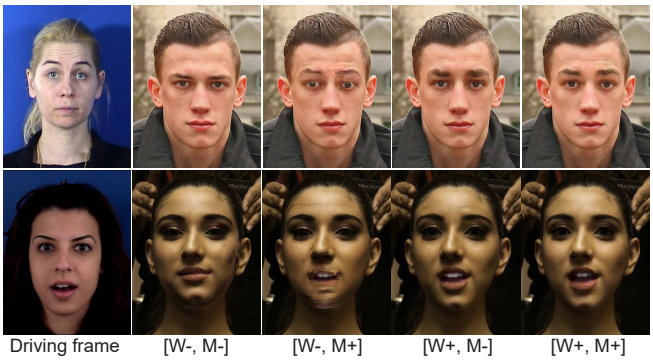

Fig. 6. Experiments with warping and displacement map. The results get degenerated without either warping or the displacement map. From top to bottom, original photos courtesy of Gert Defever and Antoine Lassalle Photography AntoineLphotos.com.

means training with warped face image but without displacement map.

Fig. 6 shows some visual examples. The results show that using the unwarped image without the displacement map, the transferred expression remains almost neutral. This is because without any conditions on the network, the network is not aware of any underlying geometric transformations thus it fails to transfer the expression and synthesize the details. Using the unwarped image with the displacement map, there will be a lot of artifacts on the regions with large motions (e.g., the widely opened eyes and mouth). This is because the network has to hallucinate these regions that do not exist on the neutral face; similarly, using the warped image without displacement map, many detailed wrinkles are not transferred correctly, because the network again loses guidance of the nonlinear displacements for facial pixels. The warped facial image, together with the displacement map, should be seamlessly integrated to guide the network to generate realistic results.

Evaluation on the hierarchical facial refinement. To validate our design of adopting two generative adversary networks to produce natural faces, we have also done an experiment which uses wg-GAN to simultaneously recover the face details and hidden regions (i.e., teeth and tongue) from a warped image. We use the same input of a warped image and a displacement map and train our network using the same training data as before (i.e., MMI + MUG + CFD), and compare the results on the test data. For a fair comparison, we also train our inpainting network using the data from MMI, MUG and CFD only. As shown in Fig. 7, coupling both the task of detail refinement and hidden region hallucination in a single architecture will significantly impose challenges and confusions to the deep neural network. The single GAN generates teeth which have much more artifacts than those obtained from the hrh-GAN.

5.2 Comparison

We compare our results to the state-of-the-art video-to-image reenactment techniques [Averbuch-Elor et al. 2017; Thies et al. 2016] and the method of G2-GAN which exploits geometric guidance for expression synthesis. Since the authors did not release the source code, we implement their algorithms. Our wg-GAN is trained with

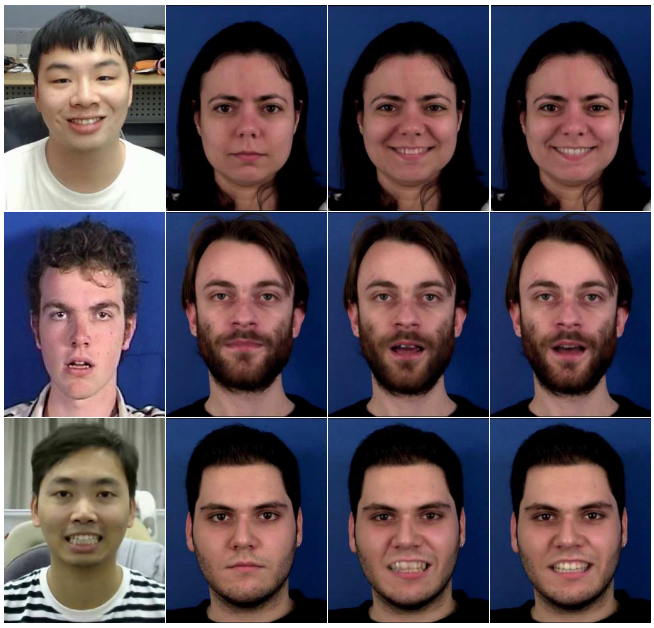

Fig. 7. Evaluation of the hierarchical network refinement. From left to right: driving frames; target frames in neutral expression; results with facial refinement and inner mouth hallucination through a single wg-GAN; and results with our two-GAN framework.

coherent facial expressions extracted from videos thus the temporal coherence is inherently guaranteed when applied to videos. To ensure the temporal coherence of the hidden region hallucination among consecutive frames, we blend two frames in texture space after optic flow alignment similar to [Thies et al. 2016].

Comparison with BP2L. We evaluate our algorithm on the internet images collected from Flickr and Unsplash as target images and compare our results with the ones from [Averbuch-Elor et al. 2017]. The results are shown in Fig. 8. Please see the supplementary video for the full animation. Note for a fair comparison, we use the same landmarks as [Averbuch-Elor et al. 2017]. We can see that our method can effectively reduce the artifacts due to image warping and produce natural wrinkles and teeth. Specifically, in the top row, when the teeth of the source person differs a lot from the target person, transferring the teeth from source to target will bring inevitable artifacts. Since our method does not rely on the teeth from the source video, we can generate more compatible teeth from our trained generative neural network. In the second row, our method effectively reduces the artifacts of closing eyes caused by warping. In the third row, as the beard and nasolabial folds are connected on the source face, the detail transfer algorithm based on ERI in [Averbuch-Elor et al. 2017] failed to clean the beard, resulting in a dark region above the mouth, while our results are natural and clean. In the fourth row, our method keeps the size of eyeball well, while in [Averbuch-Elor et al. 2017], the eyeballs are stretched due to warping. In the bottom row, our generative network is able to synthesize frown lines which do not exist in the source.

Another benefit of our method is that our facial reenactment does not rely on the resolution of the driving frame. As illustrated in Fig. 9, the face sizes of driving frames are about 130×174 pixels

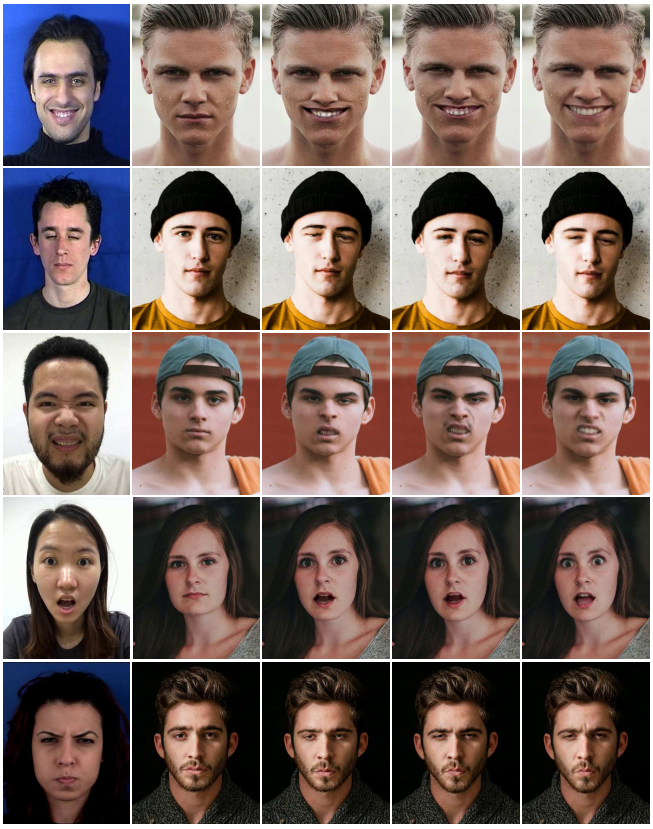

Fig. 8. Comparison of our method with prior work. From left to right, the source, the target, [Averbuch-Elor et al. 2017], [Thies et al. 2016], and our method. From top to bottom, original photos courtesy of Christopher Campbell, Clarisse Meyer, Corey Motta, Alexandre Croussette, and Albert Dera.

and 120×160 pixels separately, and the target face sizes are about 200×280 pixels and 230×310 pixels. If we transfer the source teeth to the target face as [Averbuch-Elor et al. 2017], the smaller teeth are scaled to fit the larger mouth, causing the blurred teeth in the target. In contrast, our generative neural network is able to produce the teeth suitable for the target resolution.

Comparison with Face2Face. We have also compared our algorithm with the state-of-the-art video-to-video reenactment technique [Thies et al. 2016]. We perform comparisons on the portrait images downloaded from the internet. The results are shown in the 3^{rd} column of Fig. 8. Note the differences between the warping results of [Thies et al. 2016] and [Averbuch-Elor et al. 2017] are due to the different landmarks they used. In the work of [Thies et al. 2016], they construct a mouth dataset from the target input video. Since we only have one target image, we follow [Averbuch-Elor et al. 2017; Thies et al. 2016] to extend the work by constructing a mouth database from the source sequence, and then perform mouth retrieval and blend the best mouth to the target. We can see that, because the method of [Thies et al. 2016] is also warping-based, the results unavoidably contain artifacts as mentioned above.

Comparison with G2-GAN. The work of [Song et al. 2017] uses a generative adversarial network to synthesize new facial expressions

Fig. 9. Our method is insensitive to image resolutions. From left to right, the source, the target, [Averbuch-Elor et al. 2017], and our method. The incompatible image resolutions between the source and the target causes the blurred teeth in [Averbuch-Elor et al. 2017]. From top to bottom, original photos courtesy of Gert Defever and Roman Akhmerov.

for a frontal-looking face. The input of the network is the face image with the neutral expression and the facial landmarks of the target expression. For a fair comparison, we use the same face image and target landmarks as in [Song et al. 2017] to generate the warped image and displacement maps for our networks. The authors have kindly given us the source code for the network they used in the paper, which aims to process images of 128×128 pixels, so we resize the images of MMI, MUG, and CFD to 128×128, and use the same training data to train the network of [Song et al. 2017] and ours. Some results on the test data are shown in Fig. 10. Since the method of [Song et al. 2017] does not perform warping on the whole image, we crop the facial regions for comparison. We can see that our results are much cleaner than the results of [Song et al. 2017], especially in the region of eyes and mouth. The results again confirm our observation, that recovering the global structure, local details, and hidden mouth regions together using a single neural network is very challenging. Decomposing the problem to multiple steps makes it much easier for deep neural networks to learn the highly nonlinear geometric transformations. Moreover, only providing the geometric

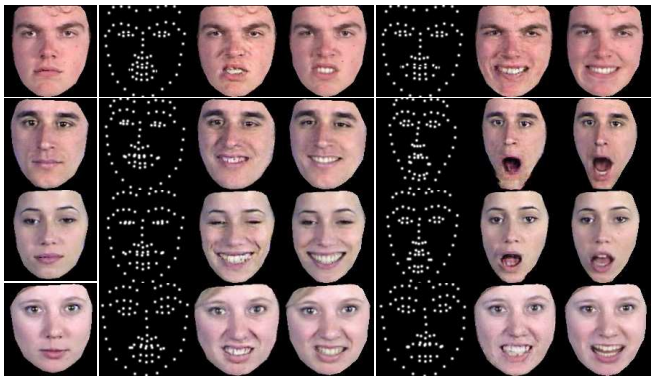

Fig. 10. Comparison of our method (the 4th and 7th columns) with [Song et al. 2017] (the 3rd and 6th columns).

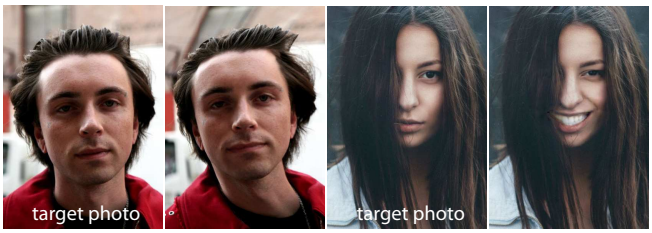

Fig. 11. Limitations of our work. Our method could fail to generate faithful results when large motion in the head and body occurs or a large portion of the facial region gets occluded by hairs. Note the distorted hairs and neck on the left example, and the broken hairs and imperfect teeth in the inner mouth region of the portrait in the right example. From left to right, original photos courtesy of Maria Badasian and Fabian Albert.

position of the landmarks does not factorize out the geometric transformations exhibited in facial expressions and could introduce artifacts.

5.3 User Study

Following [Averbuch-Elor et al. 2017], we conduct two pilot studies to quantitatively evaluate the quality of our results and compare with the method of [Averbuch-Elor et al. 2017]. We randomly selected videos from the MUG Facial Expression Database [Aifanti et al. 2010] containing videos of persons expressing various emotions. Similar to [Averbuch-Elor et al. 2017], 4 subjects were selected which had a complete set of the following four emotions: anger, fear, surprise, happiness. We generated the animated videos by selecting the first video frames to be the target images and driving these target images by one of the 3×4 (3 other subjects and 4 available emotions) driving videos.

We recruited 33 participants (15 females) in the age range of 20-40. The participants were presented with 30 randomly selected videos (6 of them were real, 12 of them were animated videos generated with our method and the rest were generated by the method of [Averbuch-Elor et al. 2017]). The participants were allowed to watch each video only once, to evaluate their first impression. In the first study, the participants were asked to rate them based on how real the animation looks. As in [Averbuch-Elor et al. 2017], we used the same 5-point Likert range of scores: *very likely fake, likely fake, could equally be real or fake, likely real, very likely real.* In the second study, the participants were asked to rate the videos based on how close the animation looks to a particular emotion. We also used the 5-point Likert range of scores for a particular emotion. Take the emotion anger for example, the Likert range of scores are: *very likely not angry, likely not angry, could equally be angry or not, likely angry, very likely angry.* We did not normalize the differences in individual aptitude.

The results of the user studies are illustrated in Table 1. The studies show that 88% of the real videos were identified as such (were rated as either likely real or very likely real). Our animated videos were identified as real 62% of the time while the animated videos generated by the method of [Averbuch-Elor et al. 2017] were identified as real 40% of the time. Similar results are shown for particular emotions (on average: 84%:64%:53%). The "happy" animations were perceived as the most real and expressive (identified as real 66% of the time and identified as happy 76% of the time), while the "surprise" animations were perceived as the least real (identified as real 55% of the time) and the "anger" animations were perceived as the least expressive (identified as angry 45% of the time). Nevertheless, our method consistently outperforms the method of [Averbuch-Elor et al. 2017] in both tests.

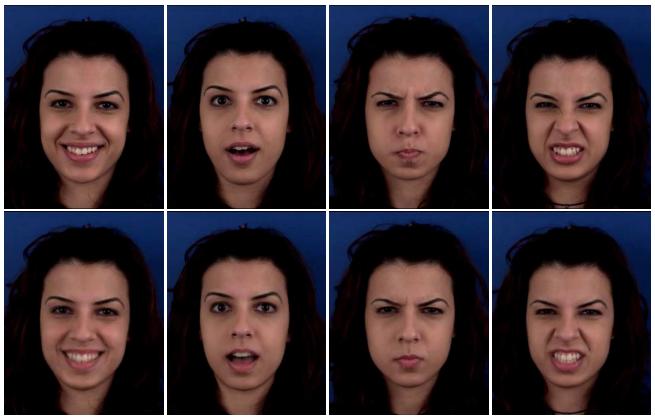

Fig. 12. The results of self-expression recovery. Top row shows the original sequence while the bottom row shows the corresponding recovered results by our method.

5.4 Limitations

Our method has a few limitations. First, we allow the movements of the head and body parts of the portrait, however, the range is limited as in [Averbuch-Elor et al. 2017] due to the 2D warping (see Fig. 11). To allow for full motions of these parts requires more sophisticated approaches to inpaint the missing regions caused by large motions. An option could be the technique used in the concurrent work of [Kim et al. 2018]. Second, as currently our training datasets were taken in a frontal neutral pose, our method requires the portrait photo to be taken in a frontal pose. In the future, it is possible to allow target images with other expressions and non-frontal poses by training the network with more diverse data. Third, since our method is built upon generative models which are trained on real images, it could fail to generate realistic results for unseen data such as cartoon and ancient painting portraits, or when large occlusion occurs in the face region (e.g., Fig. 11 right). In addition, the network synthesis procedure, by nature, is not able to recover exactly the same details as the original face and may cause some deviation from the source expression (see Fig. 12).

6 CONCLUSION

We have introduced a novel method for real-time portrait animation in a single photo. Our method takes as input a single target portrait photo with the face in the neutral-frontal pose and generates photo-realistic animations mimicking a driving source. Our method leverages lightweight 2D warps and generative adversarial networks for high-fidelity facial animation generation. Our generative network is conditioned with geometric transformations merited in the 2D warps, and can instantly fuse fine-scale facial details onto

Table 1. The user study results. The rankings (1-5) signify low (very likely fake/not angry, etc.) to high (very likely real/angry, etc.) scores.

Realism															
	Real Videos					Our Method					[Averbuch-Elor et al. 2017]				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Happy	0.04	0.08	0.02	0.34	0.52	0.08	0.12	0.14	0.38	0.28	0.24	0.30	0.17	0.20	0.10
Fear	0	0.02	0.04	0.33	0.6	0.08	0.12	0.17	0.40	0.24	0.13	0.3	0.19	0.22	0.16
Anger	0	0.12	0.04	0.33	0.51	0.03	0.17	0.16	0.36	0.28	0.07	0.15	0.18	0.36	0.24
Surprise	0.04	0.04	0.02	0.39	0.51	0.07	0.24	0.15	0.36	0.19	0.21	0.30	0.16	0.26	0.07
Average	0.08	0.06	0.03	0.35	0.54	0.06	0.16	0.15	0.37	0.25	0.16	0.26	0.17	0.26	0.14
Expressiveness															
	Real Videos					Our Method					[Averbuch-Elor et al. 2017]				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Happy	0.02	0.02	0.03	0.31	0.62	0.05	0.06	0.14	0.32	0.44	0.04	0.13	0.17	0.48	0.19
Fear	0	0.06	0.06	0.4	0.49	0.06	0.17	0.19	0.40	0.17	0.12	0.28	0.34	0.24	0.02
Anger	0.07	0.11	0.20	0.41	0.20	0.11	0.20	0.25	0.35	0.1	0.07	0.18	0.28	0.38	0.10
Surprise	0.02	0.02	0.02	0.41	0.53	0.03	0.10	0.09	0.42	0.35	0.05	0.08	0.16	0.53	0.18
Average	0.11	0.051	0.079	0.38	0.46	0.06	0.13	0.17	0.37	0.27	0.07	0.17	0.24	0.41	0.12

a warped face image in a high realism manner. Unlike the concurrent work of [Kim et al. 2018], our network is generic and does not rely on contents of either a source driving video or the target photo (video).

Our pipeline achieves significantly better results than the state-of-the-art methods towards fine-scale detail synthesis such as wrinkles, creases, self-shadows, and teeth, etc., thanks to our carefully designed global-and-local paradigm with conditional generative adversarial neural networks. The utilization of lightweight 2D warps and GPUs in neural nets also enables our method to operate in real-time. A number of experiments showed that our approach is suitable for realtime face reenactment in a single photo. In the future, we plan to explore more sophisticated portrait animation techniques through generic inpainting techniques (e.g., towards the background), or leveraging 3D face and hair databases for high-fidelity portrait animation in images.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their constructive comments, Xiaowei Zhou for helpful discussions, MUG, MMI and CFD groups for publishing their database, and the Flickr and Unsplash users (Gert Defever, Kris Krüg, Jevgeni Kurnikov, Antoine Lassalle, Elvis Ripley) for letting us use their work under the Creative Commons License. This work is partially supported by the National Key Research & Development Program of China (2016YFB1001403), NSF China (No. 61772462, No. 61572429, No. 61502306, U1609215), Microsoft Research Asia, and the China Young 1000 Talents Program.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. TensorFlow: A System for Large-Scale Machine Learning.. In *OSDI*, Vol. 16. 265–283.

Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. 2010. The MUG facial expression database. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th international workshop on*. IEEE, 1–4.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* (2017).

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. 2017. Bringing portraits to life. *ACM Trans. Graph.* 36, 6 (2017), 196:1–196:13.

Jiamin Bai, Aseem Agarwala, Maneesh Agrawala, and Ravi Ramamoorthi. 2013. Automatic cinemagraph portraits. *Computer Graphics Forum* 32, 4 (2013), 17–25.

Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. 2003. Reanimating faces in images and video. *Computer Graphics Forum* 22, 3 (2003), 641–650.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. ACM Press/Addison-Wesley Publishing Co., 187–194.

Sofien Bouaziz, Yangang Wang, and Mark Pauly. 2013. Online modeling for realtime facial animation. *ACM Trans. Graph.* 32, 4 (2013), 40:1–40:10.

Pia Breuer, Kwang-In Kim, Wolf Kienzle, Bernhard Scholkopf, and Volker Blanz. 2008. Automatic 3D face reconstruction from single images or video. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*. IEEE, 1–8.

Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Trans. Graph.* 34, 4 (2015), 46:1–46:9.

Chen Cao, Qiming Hou, and Kun Zhou. 2014a. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Trans. Graph.* 33, 4 (2014), 43:1–43:10.

Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 2013. 3D shape regression for real-time facial animation. *ACM Trans. Graph.* 32, 4 (2013), 41.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014b. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2014), 413–425.

Chen Cao, Hongzhi Wu, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2016. Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph.* 35, 4 (2016), 126:1–126:12.

Kevin Dale, Kalyan Sunkavalli, Micah K Johnson, Daniel Vlasic, Wojciech Matusik, and Hanspeter Pfister. 2011. Video face replacement. *ACM Trans. Graph.* 30, 6 (2011), 130:1–130:10.

Hui Ding, Kumar Sricharan, and Rama Chellappa. 2018. ExprGAN: Facial Expression Editing with Controllable Expression Intensity. In *AAAI*.

Ohad Fried, Eli Shechtman, Dan B Goldman, and Adam Finkelstein. 2016. Perspective-aware manipulation of portrait photos. *ACM Trans. Graph.* 35, 4 (2016), 128:1–128:10.

Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. 2016. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *European Conference on Computer Vision (ECCV)*. Springer, 311–326.

Pablo Garrido, Levi Valgaerts, Ole Rehmsen, Thorsten Thormahlen, Patrick Perez, and Christian Theobalt. 2014. Automatic face reenactment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4217–4224.

Pablo Garrido, Levi Valgaerts, Hamid Sarmadi, Ingmar Steiner, Kiran Varanasi, Patrick Perez, and Christian Theobalt. 2015. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. *Computer Graphics Forum* 34, 2 (2015), 193–204.

Jon Gauthier. 2014. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester* 2014, 5 (2014), 2.

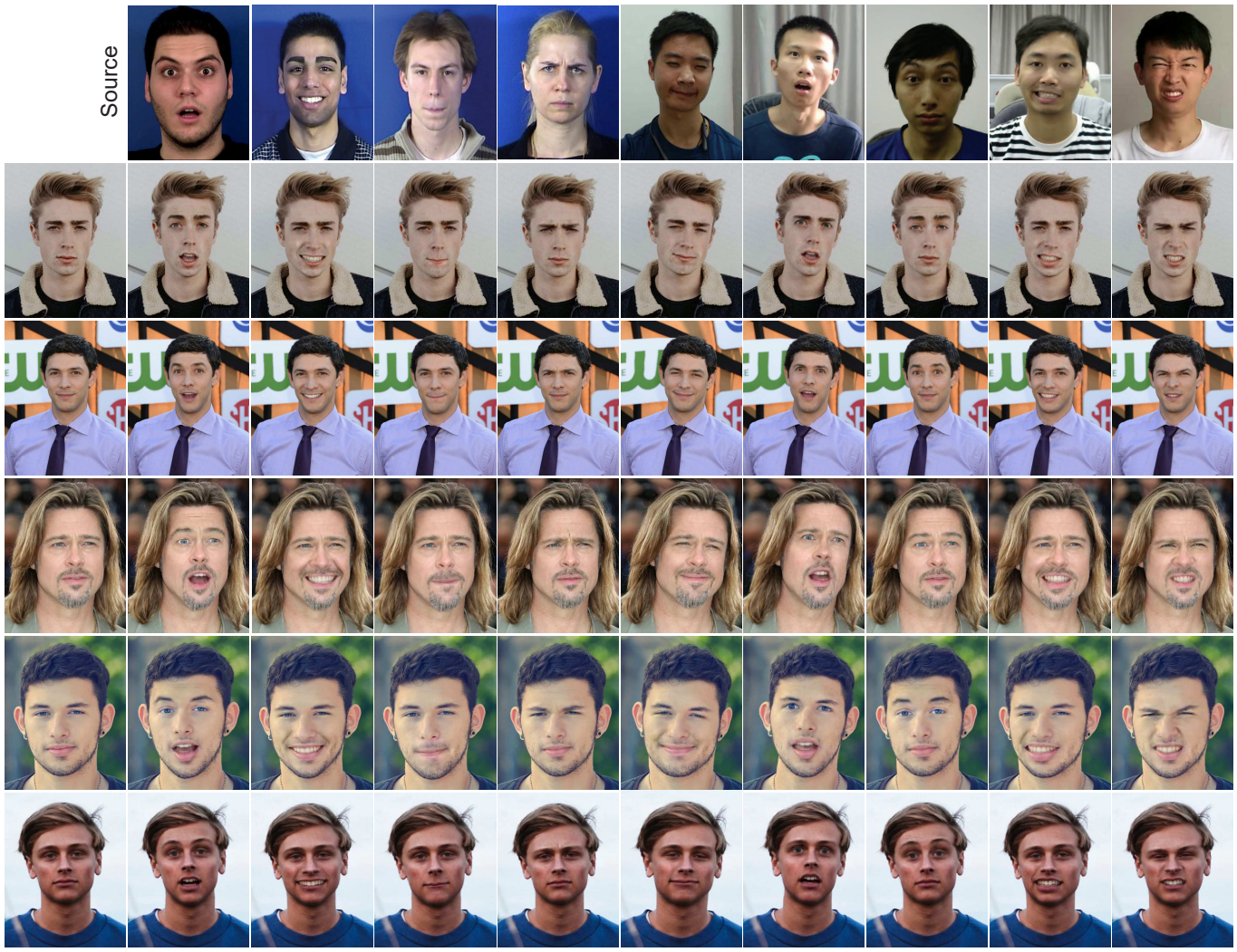

Fig. 13. Portrait animation results by our approach on images randomly collected from Flickr and Unsplash. From top to bottom, original photos courtesy of Parker Whitson, Pedro Haas, Getty Images, Laura Nicola and Oliver Ragfelt.

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

Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. 2015. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4295–4304.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. 2015. Unconstrained realtime facial performance capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1675–1683.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Trans. Graph.* 36, 4 (2017), 107:1–107:14.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017).

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

Masahide Kawai, Tomoyori Iwao, Daisuke Mima, Akinobu Maejima, and Shigeo Morishima. 2013. Photorealistic inner mouth expression in speech animation. In *ACM SIGGRAPH 2013 Posters*. ACM, 9:1–9:1.

Masahide Kawai, Tomoyori Iwao, Daisuke Mima, Akinobu Maejima, and Shigeo Morishima. 2014. Data-driven speech animation synthesis focusing on realistic inside of the mouth. *Journal of information processing* 22, 2 (2014), 401–409.

Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep Video Portraits. *ACM Trans. Graph.* 37, 4 (2018), 163:1–163:14.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. Fast face-swap using convolutional neural networks. In *The IEEE International Conference on Computer Vision*. 3697–3705.

Claudia Kuster, Tiberiu Popa, Jean-Charles Bazin, Craig Gotsman, and Markus Gross. 2012. Gaze correction for home video conferencing. *ACM Trans. Graph.* 31, 6 (2012), 174:1–174:6.

Chuan Li and Michael Wand. 2016. Combining markov random fields and convolutional neural networks for image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2479–2486.

Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. 2013. Realtime facial animation with on-the-fly correctives. *ACM Trans. Graph.* 32, 4 (2013), 42:1–42:10.

Kai Li, Feng Xu, Jue Wang, Qionghai Dai, and Yebin Liu. 2012. A data-driven approach for facial expression synthesis in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 57–64.

Yilong Liu, Feng Xu, Jinxiang Chai, Xin Tong, Lijuan Wang, and Qiang Huo. 2015. Video-audio driven real-time facial animation. *ACM Trans. Graph.* 34, 6 (2015), 182:1–182:10.

Fig. 14. Additional portrait animation results by our approach with intermediate results showing various components of our pipeline. From left to right, original photos courtesy of Gert Defever, Kris Krüg, Elvis Ripley, Riley Kaminer, Genessa Panainte, Pedro Haas, Pedro Haas, Pedro Haas, Oliver Ragfelt, and Arjunsyah.

Zicheng Liu, Ying Shan, and Zhengyou Zhang. 2001. Expressive expression mapping with ratio images. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 271–276.

Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47, 4 (2015), 1122–1135.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, Vol. 30. 3.

Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. 2016. Do we really need to collect millions of faces for effective face recognition?. In *European Conference on Computer Vision*. Springer, 579–596.

Umar Mohammed, Simon JD Prince, and Jan Kautz. 2009. Visio-lization: generating novel facial images. *ACM Trans. Graph.* 28, 3 (2009), 57:1–57:8.

Kyle Olszewski, Zimo Li, Chao Yang, Yi Zhou, Ronald Yu, Zeng Huang, Sitao Xiang, Shunsuke Saito, Pushmeet Kohli, and Hao Li. 2017. Realistic dynamic facial textures from a single image using gans. In *IEEE International Conference on Computer Vision (ICCV)*. 5429–5438.

Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. 2005. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 5–pp.

Marcel Piotraschke and Volker Blanz. 2016. Automated 3d face reconstruction from multiple images using quality measures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3418–3427.

Fengchun Qiao, Naiming Yao, Zirui Jiao, Zhihao Li, Hui Chen, and Hongan Wang. 2018. Geometry-Contrastive Generative Adversarial Network for Facial Expression Synthesis. *arXiv preprint arXiv:1802.01822* (2018).

Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. 2014. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph.* 33, 6 (2014), 222:1–222:13.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russell Webb. 2017. Learning from Simulated and Unsupervised Images through Adversarial Training. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), 2242–2251.

Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. 2017. Geometry Guided Adversarial Facial Expression Synthesis. *arXiv preprint arXiv:1712.03474* (2017).

Joshua M Susskind, Geoffrey E Hinton, Javier R Movellan, and Adam K Anderson. 2008. Generating facial expressions with deep belief nets. In *Affective Computing*. InTech.

Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-time expression transfer for facial reenactment. *ACM Trans. Graph.* 34, 6 (2015), 183:1–183:14.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2387–2395.

Michel Valstar and M Pantic. 2010. Induced disgust, happiness and surprise: An addition to the mmi facial expression database. In *Proc. Int'l Conf. Language Resources and Evaluation, Workshop EMOTION*. 65–70.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face transfer with multilinear models. *ACM Trans. Graph.* 24, 3 (2005), 426–433.

Congyi Wang, Fuhao Shi, Shihong Xia, and Jinxiang Chai. 2016. Realtime 3d eye gaze animation using a single rgb camera. *ACM Trans. Graph.* 35, 4 (2016), 118:1–118:14.

Thibaut Weise, Sofien Bouaziz, Hao Li, and Mark Pauly. 2011. Realtime performance-based facial animation. *ACM Trans. Graph.* 30, 4 (2011), 77:1–77:10.

Fei Yang, Lubomir Bourdev, Eli Shechtman, Jue Wang, and Dimitris Metaxas. 2012. Facial expression editing in video using a temporally-smooth factorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 861–868.

Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. 2011. Expression flow for 3D-aware face component transfer. *ACM Trans. Graph.* 30, 4 (2011), 60:1–60:10.

Raymond Yeh, Ziwei Liu, Dan B Goldman, and Aseem Agarwala. 2016. Semantic facial expression editing using autoencoded flow. *arXiv preprint arXiv:1611.09961* (2016).