SIP (2020), vol. 9, e22, page 1 of 13 © The Author(s), 2020. Published by Cambridge University Press.. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/ATSIP.2020.20

ORIGINAL PAPER

PortraitGAN for flexible portrait manipulation

JIALI DUAN,[1] XIAOYUAN GUO[2] AND C.-C. JAY KUO[1]

Previous methods have dealt with discrete manipulation of facial attributes such as smile, sad, angry, surprise, etc., out of canonical expressions and they are not flexible, operating in single modality. In this paper, we propose a novel framework that supports continuous edits and multi-modality portrait manipulation using adversarial learning. Specifically, we adapt cycle-consistency into the conditional setting by leveraging additional facial landmarks information. This has two effects: first cycle mapping induces bidirectional manipulation and identity preserving; second pairing samples from different modalities can thus be utilized. To ensure high-quality synthesis, we adopt texture-loss that enforces texture consistency and multi-level adversarial supervision that facilitates gradient flow. Quantitative and qualitative experiments show the effectiveness of our framework in performing flexible and multi-modality portrait manipulation with photo-realistic effects. Code will be made public: shorturl.at/chopD.

Keywords: Generative adversarial learning, Photo-realistic

Received 8 June 2020; Revised 2 September 2020

I. INTRODUCTION

Our digital age has witnessed a soaring demand for flexible, high-quality portrait manipulation, not only from smart-phone apps but also from photography industry, e-commerce, and movie production, etc. Portrait manipulation is a widely studied topic [1–6] in computer vision and computer graphics. From another perspective, many computer vision problems can be seen as translating image from one domain (modality) to another, such as colorization [7], style transfer [8–11], image inpainting [12], and visual attribute transfer [13], etc. This cross-modality image-to-image translation has received significant attention [14, 15] in the community. In this paper, we define different styles as modalities and try to address multi-modality transfer using a single model. In terms of practical concern, transfer between each pair of modalities as opposed to SPADE [16] or GANPaint [17] whose manipulation domain is fixed.

Recently, generative adversarial networks have demonstrated compelling effects in synthesis and image translation [14, 15, 18–21], among which [15, 22] proposed cycle-consistency for unpaired image translation. In this paper, we extend this idea into a conditional setting by leveraging additional facial landmarks information, which is capable of capturing intricate expression changes. Benefits that arise with this simple yet effective modifications include: First, cycle mapping can effectively prevent many-to-one mapping [15, 23] also known as mode-collapse. In the context of face/pose manipulation, cycle-consistency also induces identity preserving and bidirectional manipulation, whereas previous method [1] assumes neutral face to begin with or is unidirectional [24, 25], manipulating in the same domain. Second, face images of different textures or styles are considered different modalities and current landmark detector will not work on those stylized images. With our design, we can pair samples from multiple domains and translate between each pair of them, thus enabling landmark extraction indirectly on stylized portraits. Our framework can also be extended to makeups/de-makeups, aging manipulation, etc., once corresponding data are collected. In this work, we leverage [10] to generate pseudo-targets, i.e. stylized faces to learn simultaneous expression and modality manipulations, but it can be replaced with any desired target domains.

However, there remain two main challenges to achieve high-quality portrait manipulation. We propose to learn a single generator G as in [26]. But StarGAN [26] deals with discrete manipulation and fails on high-resolution images with irremovable artifacts. To synthesize images of photo-realistic quality (512×512), we propose multi-level adversarial supervision inspired by [27, 28] where synthesized images at different resolution are propagated and combined before being fed into multi-level discriminators. Second, to avoid texture inconsistency and artifacts during translation between different domains, we integrate Gram matrix [8] as a measure of texture distance into our model as it is differentiable and can be trained end-to-end using back propagation. Figure 1 shows the result of our model.

[1]University of Southern California, 3740 McClintock Avenue, Los Angeles, USA
[2]Emory University, 201 Dowman Dr, Atlanta, USA

Corresponding author:
Jiali Duan
Email: jialidua@usc.edu

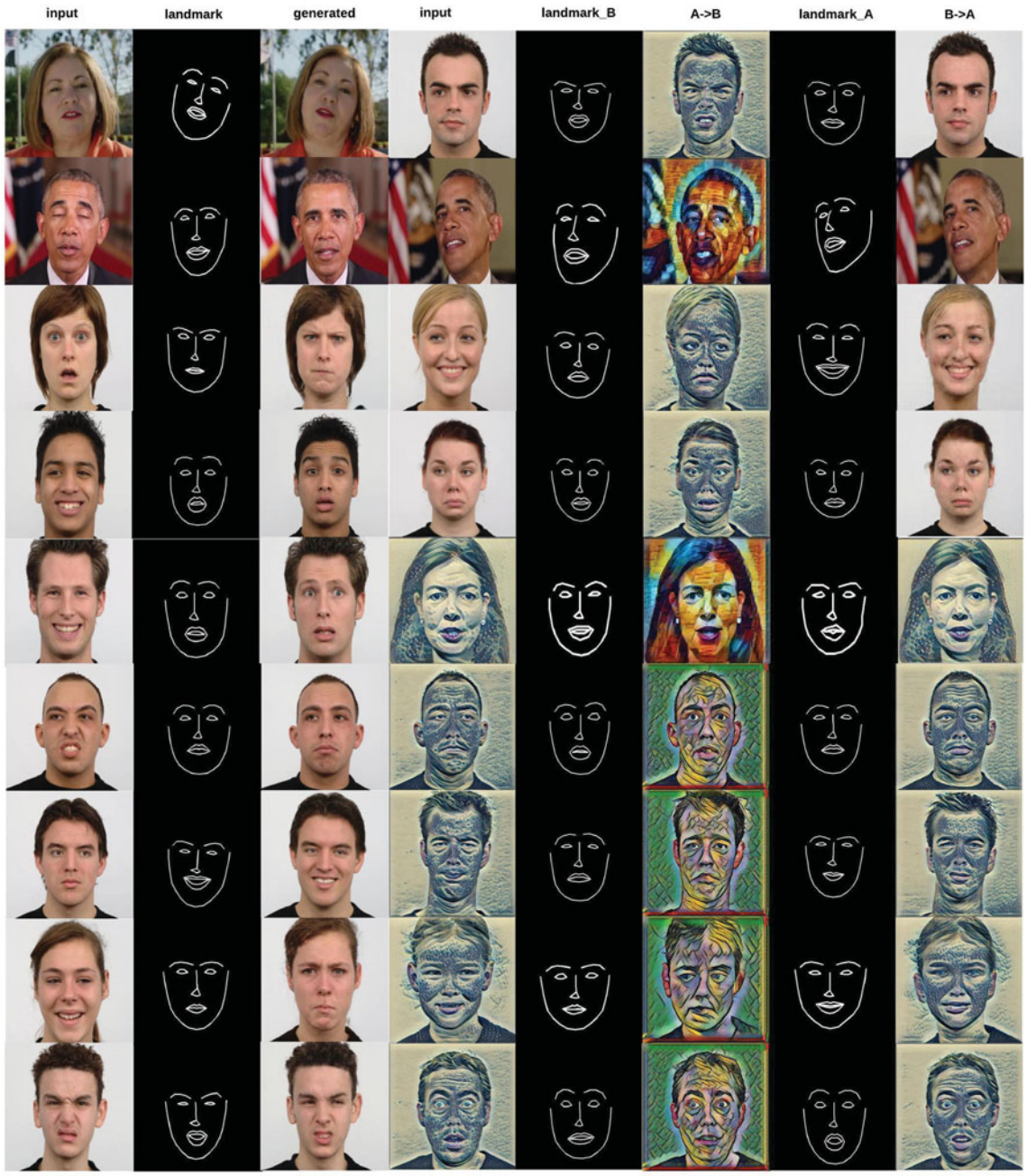

Fig. 1. More results for continuous shape edits and simultaneous shape and modality manipulation results by PortraitGAN.

Extensive evaluations have shown both quantitatively and qualitatively that our method is comparable or superior to state-of-the-art generative models in performing high-quality portrait manipulation. Our model is bidirectional, which circumvents the need to start from a neutral face or a fixed domain. This feature also ensures stable training, identity preservation, and is easily scalable to other desired domain manipulations. In the following section, we review related works to ours and point out the differences. Details of PortraitGAN are elaborated in Section III. We evaluate our approach in Section IV and conclude the paper in Section V.

II. RELATED WORK

Face editing

Face editing or manipulation is a widely studied area in the field of computer vision and graphics, including face morphing [29], expression edits [30, 31], age progression [32], facial reenactment [1, 6, 33]. However, these models are designed for a particular task and rely heavily on domain knowledge and certain assumptions. For example, [1] assumes neutral and frontal faces to begin with while [6] employs 3D model and assumes the availability of target videos with variation in both poses and expressions. Our

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

model differs from them as it is a data-driven approach that does not require domain knowledge, designed to handle general face manipulations.

Image translation

Our work can be categorized into image translation with generative adversarial networks [14, 18, 22, 27, 34, 35], whose goal is to learn a mapping $G : \mathcal{X} \rightarrow \widehat{\mathcal{Y}}$ that induces an indistinguishable distribution to target domain $\mathcal{Y}$, through adversarial training. For example, Isola *et al.* [14] take image as a condition for general image-to-image translation trained on paired samples. Later, Zhu *et al.* [15] build upon [14] by introducing cycle-consistency loss to obviate the need of matched training pairs. In addition, it alleviates many-to-one mapping during training generative adversarial networks also known as mode collapse. Inspired by this, we integrate this loss into our model for identity preservation between different domains.

Another seminal work that inspired our design is StarGAN [26], where target facial attributes are encoded into a one-hot vector. In StarGAN, each attribute is treated as a different domain and an auxiliary classifier used to distinguish these attributes is essential for supervising the training process. Different from StarGAN, our goal is to perform continuous edits in the pixel space that cannot be enumerated with discrete labels. This implicitly implies a smooth and continuous latent space where each point in this space encodes meaningful axis of variation in the data. We treat different style modalities as domains in this paper and use two words interchangeably. In this sense, applications like beautification/de-beautification, aging/younger, with beard/without beard can also be included into our general framework. We compare our approach against CycleGAN [15] and StarGAN [26] during experiments and illustrate in more details about our design in the next section.

Landmark guided generation

In [36], an offline interpolation process is adopted for generating face boundary map, to be used for GMM clustering and as conditional prior. There are two key differences: (1) the number of new expressions depends on clustering, possibly not continuous; (2) boundary heat map is estimated offline. In [37], facial landmarks are represented as VAE encoding for GAN. In contrast, the major goal of our framework is to support online, flexible, even interactive in user experience, which is why we process and leverage landmarks in a different way, as a channel map.

There are also works that use pose landmarks as condition for person image generation [25, 38–40]. For example, [24] concatenates one-hot pose feature maps in a channel-wise fashion to control pose generation. Different from our approach, each landmark constitutes one channel. In [41], keypoints and segmentation mask of birds are used to manipulate locations and poses of birds. To synthesize more plausible human poses, Siarohin et.al [39] develop deformable skip connections and compute a set of affine transformations to approximate joint deformations. These works share some similarity with ours as both facial landmark and human skeleton can be seen as a form of pose representation. However, the above works deal with manipulation in the original domain and does not preserve identity.

Style transfer

Exemplar-guided neural style transfer was first proposed by Gatys *et al.* [8]. The idea is to preserve content from the original image and mimic "style" from a reference image. We adopt Gram matrix in our model to enforce pattern consistency. We apply a fast neural-style transfer algorithm [10] to generate pseudo targets for multi-modality manipulations. Another branch of work [16, 42] try to model style distribution in another domain which is in favor of one-to-many mapping in the target domain, or collection style transfer [43].

III. PROPOSED METHOD

Problem formulation

Given domains $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3, \ldots, \mathcal{X}_n$ of different modalities, our goal is to learn a *single* general mapping function

$$G : \mathcal{X}_i \rightarrow \mathcal{X}_j, \quad \forall i, j \in \{1, 2, 3, \ldots, n\} \tag{1}$$

that transforms $\mathcal{I}_A$ from domain A to $\mathcal{I}_B$ from domain B in a continuous manner. Equation 1 implicitly implies that G is bidirectional given desired conditions. We use facial landmark $\mathcal{L}_j \in R^{1\times H\times W}$ to denote facial expression in domain j. Facial expressions are represented as a vector of 2D keypoints with $N = 68$, where each point $u_i = (x_i, y_i)$ is the ith pixel location in $\mathcal{L}_j$. We use attribute vector $\bar{c} = [c_1, c_2, c_3, \ldots, c_n]$ to represent the target domain. Formally, our input/output are tuples of the form $(\mathcal{I}_A, \mathcal{L}_B, c_B)/(\mathcal{I}_B, \mathcal{L}_A, c_A) \in R^{(3+1+n)\times H\times W}$.

Model architecture

The overall pipeline of our approach is straightforward, shown in Fig. 2 consisting of three main components: (1) A generator $G(\mathcal{I}, \mathcal{L}, \bar{c})$, which renders an input face in domain $\overline{c_1}$ to the same person in another domain $\overline{c_2}$ given conditional facial landmarks. G is bidirectional and reused in both forward as well as backward cycle. First mapping $\mathcal{I}_A \rightarrow \widehat{\mathcal{I}_B} \rightarrow \widehat{\mathcal{I}_A}$ and then mapping back $\mathcal{I}_B \rightarrow \widehat{\mathcal{I}_A} \rightarrow \widehat{\mathcal{I}_B}$ given conditional pair $(\mathcal{L}_B, \overline{c_B})/(\mathcal{L}_A, \overline{c_A})$. (2) A set of discriminators D_i at different levels of resolution that distinguish generated samples from real ones. Instead of mapping $\mathcal{I}$ to a single scalar which signifies "real" or "fake", we adopt PatchGAN [9] which uses a fully convnet that outputs a matrix where each element $M_{i,j}$ represents the probability of overlapping patch ij to be real. If we trace back to the original image, each output has a 70×70 receptive field. (3) Our loss function takes into account identity preservation and texture consistency between different domains. In the following sections, we elaborate on each module individually and then combine them together to construct PortraitGAN.

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

A) Base model

To begin with, we consider manipulation of emotions in the same domain, i.e. $\mathcal{I}_A$ and $\mathcal{I}_B$ are of same texture and style, but with different face shapes denoted by facial landmarks $\mathcal{L}_A$ and $\mathcal{L}_B$. Under this scenario, it is sufficient to incorporate only forward cycle and conditional modality vector is not needed. The adversarial loss conditioned on facial landmarks follows equation 2.

$$\mathcal{L}_{\text{GAN}}(G,D) = E_{\mathcal{I}_B}[\log(D(\mathcal{I}_B)] + E_{(\mathcal{I}_A,\mathcal{L}_B)}[\log(1 - D(G(\mathcal{I}_A,\mathcal{L}_B)))] \quad (2)$$

A face verification loss is desired to preserve identity between $\mathcal{I}_B$ and $\widehat{\mathcal{I}_B} = G(\mathcal{I}_A,\mathcal{L}_B)$. However, in our experiments, we find ℓ_1 loss to be enough and it is better than ℓ_2 loss as it alleviates blurry output and acts as an additional regularization [14].

$$\mathcal{L}_{id}(G) = E_{(\mathcal{I}_A,\mathcal{L}_B,\mathcal{I}_B)}||\mathcal{I}_B - G(\mathcal{I}_A,\mathcal{L}_B)||_1 \quad (3)$$

The overall loss is a combination of adversarial loss and ℓ_1 loss, weighted by λ.

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{GAN}}(G,D) + \lambda\mathcal{L}_{\text{id}}(G) \quad (4)$$

B) Multi-level adversarial supervision

Manipulation at a landmark level requires high-resolution synthesis, which is challenging [44], because it is harder to optimize.

Here we use two strategies for improving generation quality and training stability. First our conditional facial landmark acts as an additional constraint for generation. Second, we adopt a multi-level feature matching loss [8, 45] to explicitly require G to match statistics of real data that D finds most discriminative at feature level as follows.

$$\mathcal{L}_{FM}(G,D_k) = \mathbb{E}_{(I_A,I_B)} \sum_{i=1}^{T} \frac{1}{N_i} \|D_k^i(\mathcal{I}_\mathbf{B}) - D_k^i(G(\mathcal{I}_\mathbf{A},\mathcal{L}_\mathbf{B}))\|_1 \quad (5)$$

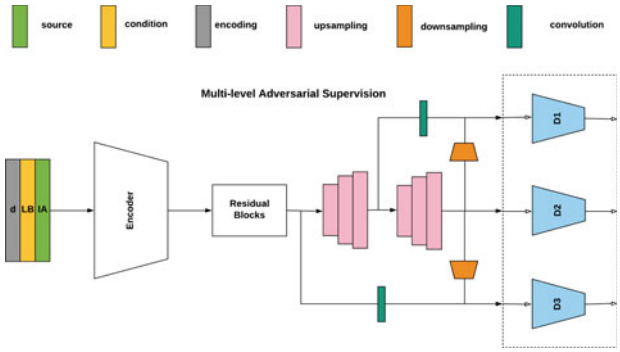

Fig. 3. Multi-level adversarial supervision.

we denote the ith-layer feature extractor of discriminator D_k as D_k^i, where T is the total number of layers and N_i denotes the number of elements in each layer.

Third, we provide fine-grained guidance by propagating multi-level features for adversarial supervision (Fig. 3). Cascaded upsampling layers in G are connected with auxiliary convolutional branches to provide images at different scales $(\widehat{\mathcal{I}_{B1}}, \widehat{\mathcal{I}_{B2}}, \widehat{\mathcal{I}_{B3}} \ldots \widehat{\mathcal{I}_{Bm}})$, where m is the number of upsampling blocks. These images are fed into discriminators at different scales D_k. Applying it to equation 4 we get,

$$\mathcal{L}_{multi} = \sum_k [\mathcal{L}_{\text{GAN}}(G,D_k) + \beta\mathcal{L}_{FM}(G,D_k)] + \gamma\mathcal{L}_{\text{id}}(G) \quad (6)$$

Compared to [28], our proposed discriminators responsible for different levels are optimized as a whole rather than individually for each level. The increased discriminative ability from D_k in turn provides further guidance when training G (equation 6).

C) Texture consistency

When translating between different modalities in high-resolution, texture differences become easy to observe. Inspired by [8], we let $\psi^k_{\mathcal{I},L}$ be the vectorized kth extracted

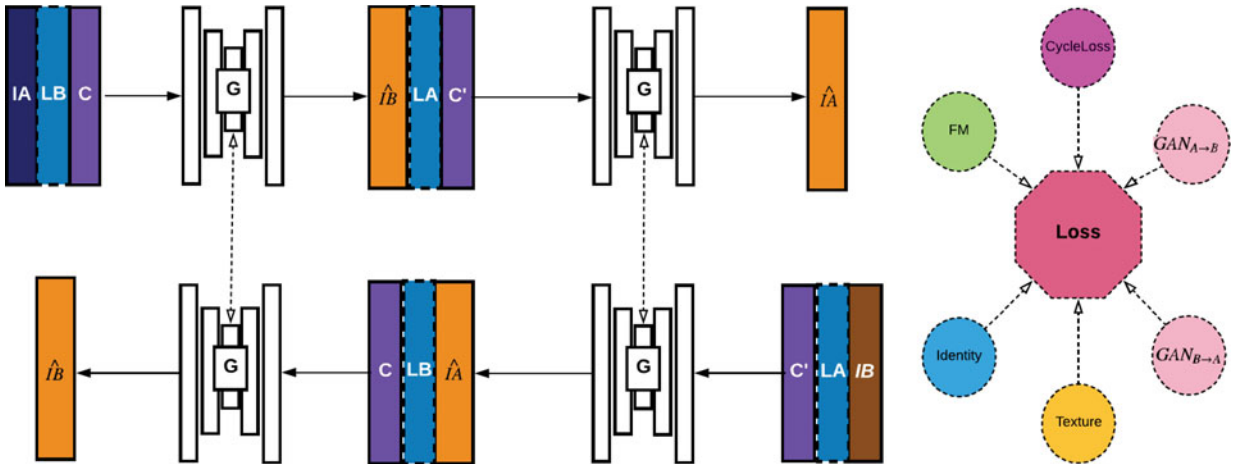

Fig. 2. Overview of training pipeline: In the forward cycle, original image IA is first translated to $\widehat{IB}$ given target emotion LB and modality C and then mapped back to $\widehat{IA}$ given condition pair (LA,C') encoding the original image. The backward cycle follows similar manner starting from IB but with opposite condition encodings using the same generator G. Identity preservation and modality constraints are explicitly modeled in our loss design.

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

feature map of image $\mathcal{I}$ from neural network ψ at layer L. $\mathcal{G}_{\mathcal{I},L} \in R^{\kappa \times \kappa}$ is defined as,

$$\mathcal{G}_{\mathcal{I},L}(k,l) =< \psi^k_{\mathcal{I},L}, \psi^l_{\mathcal{I},L} >= \sum_i \psi^k_{\mathcal{I},L}(i) \cdot \psi^l_{\mathcal{I},L}(i) \quad (7)$$

where κ is the number of feature maps at layer L and $\psi^k_{\mathcal{I},L}(i)$ is ith element in the feature vector. Equation 7 also known as Gram matrix can be seen as a measure of the correlation between feature maps k and l, which only depends on the number of feature maps, not the size of $\mathcal{I}$. For image $\mathcal{I}_A$ and $\mathcal{I}_B$, the texture loss at layer L is,

$$\mathcal{L}^L_{\text{texture}}(\widehat{\mathcal{I}_B}, \mathcal{I}_B) = ||\mathcal{G}_{\widehat{\mathcal{I}_B},L} - \mathcal{G}_{\mathcal{I}_B,L}||^2 \quad (8)$$

where $\widehat{\mathcal{I}_B} = G(\mathcal{I}_A, \mathcal{L}_B)$. We obtain obvious improvement in quality of texture in cross-modality generation and we use pretrained VGG19 for texture feature extraction in our experiments with its parameters frozen during optimization.

D) Bidirectional portrait manipulation

To transfer to a target domain $\mathcal{X}$, an additional one-hot encoding vector $\bar{c} \in R^n$ is conditioned as input. Specifically, each element is first replicated spatially into size $H \times W$ and then concatenated with image and landmark along the channel axis. The only change to previous equations is that instead of taking $(\mathcal{I}_A, \mathcal{L}_B)$ as input, the generator G now takes $(\mathcal{I}_A, \mathcal{L}_B, \bar{c})$, where $\bar{c}$ indicates the domain where $\mathcal{I}_B$ belongs to.

To encourage bijection between mappings in different modality manifold and to prevent mode collapse, we adopt cycle-consistency structure similar to [15], which consists of a forward and a backward cycle, for both generating directions.

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G) = {} & E_{(\mathcal{I}_A,\mathcal{L}_B,\bar{c},\bar{c}')}[||G(G(\mathcal{I}_A, \mathcal{L}_B, \bar{c}), \mathcal{L}_A, \bar{c}') - \mathcal{I}_A||]_1 \\ & + E_{(\mathcal{I}_B,\mathcal{L}_A,\bar{c},\bar{c}')}[||G(G(\mathcal{I}_B, \mathcal{L}_A, \bar{c}'), \mathcal{L}_B, \bar{c}) - \mathcal{I}_B||]_1 \end{aligned} \quad (9)$$

where $\bar{c}$ and $\bar{c}'$ encodes modality for $\mathcal{I}_B$ and $\mathcal{I}_A$ respectively. Note that only one set of generator/discriminator is used for bidirectional manipulation. Our final optimization objective for PortraitGAN is as follows,

$$\begin{aligned}\mathcal{L}_{\text{PortraitGAN}} = {} & \mathcal{L}_{\text{multi}_{A\to B}} + \mathcal{L}_{\text{multi}_{B\to A}} \\ & + \alpha * \mathcal{L}_{\text{cyc}} + \eta * \mathcal{L}_{\text{texture}} \end{aligned} \quad (10)$$

where α,η controls the weight for cycle-consistency loss and texture loss respectively.

IV. EXPERIMENTAL EVALUATION

Our goal in this section is to test our model's capability in (1) continuous shape editing; (2) simultaneous modality transfer. We also created testbed for comparing our model against two closely related SOTA methods [15, 26], though they do not support either continuous shape editing and multi-modality transfer directly. The aim is to provide quantitative and qualitative analysis in terms of perceptual quality. Additionally, we also conducted ablation studies for our components.

Implementation details

Each training step takes as input a tuple of four images ($\mathcal{I}_A$, $\mathcal{I}_B$, $\mathcal{L}_A$, $\mathcal{L}_B$) randomly chosen from possible modalities of the same identity. Attribute conditional vector, represented as a one-hot vector, is replicated spatially before channel-wise concatenation with corresponding image and facial landmarks. Our generator uses 4 stride-2 convolution layers, followed by nine residual blocks and 4 stride-2 transpose convolutions while auxiliary branch uses one-channel convolution for fusion of channels. We use two three-layer PatchGAN [9] discriminators for multi-level adversarial supervision and Least Square loss [46] for stable training. Layer conv1_1-conv5_1 of VGG19 [47] are used for computing texture loss. We set α, β, γ, η as 2, 10, 5, 10 to ensure that loss components are at the same scale. There are four styles used in our experiment, for training a unified deep model for shape and modality manipulation. The training time for PortraitGAN takes around 50 h on a single Nvidia 1080 GPU.

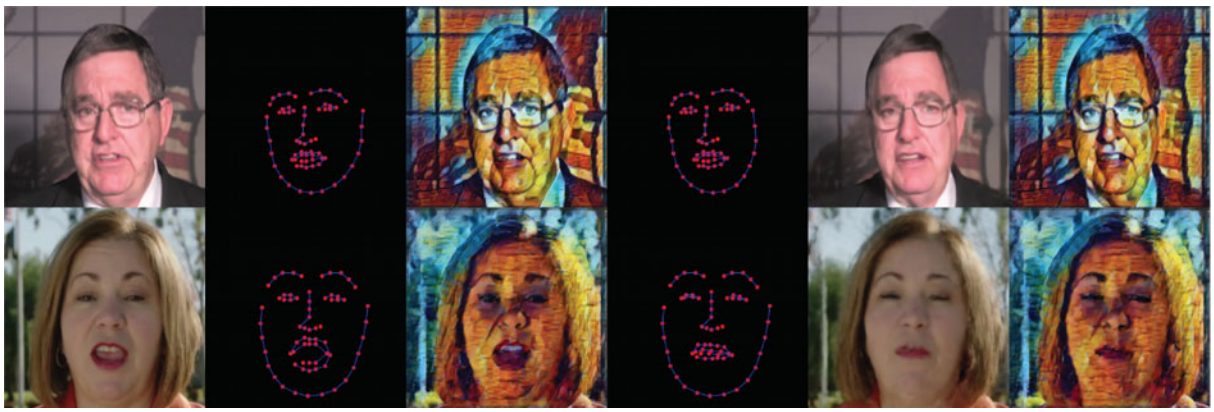

Fig. 4. Interactive manipulation without constraints. Column 1st–2nd: Original image and auto-detected facial landmarks; 3rd: generated image from 1st-2nd; 4th: manipulated target landmark; 5th: inverse modality generation from 3rd–4th; 6th: photo to style generation with landmarks of 5th.

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

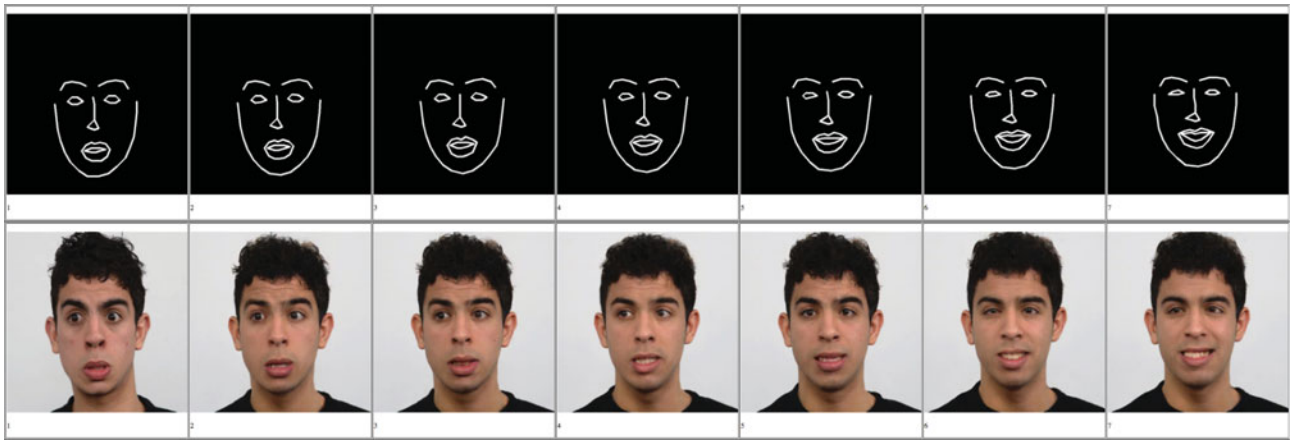

Fig. 5. Given leftmost and rightmost face, we first interpolate the middle one (e.g. the 4th one), then we can interpolate 2nd (with 1st and 4th) and 5th (with 4th and 7th). Lastly, we interpolate 3rd (with 2nd and 4th) and 6th (with 5th and 7th).

Dataset

We collected and combined the following three emotion dataset for experiments and performed a 7/3 split based on identity for training and testing. (1) The Radboud Faces Database [48] contains 4,824 images with 67 participants, each performing eight canonical emotional expressions: anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral. (2) iCV Multi-Emotion Facial Expression Dataset [49] is designed for micro-emotion recognition, which includes 31,250 facial expressions from 125 subjects performing 50 different emotions. (3) We also collected 20 videos of high-resolution from Youtube (abbreviated as HRY Dataset) containing 10 people giving speech or talk. For the above dataset, we use dlib [50] for facial landmark extraction and [10] for generating portraits of multiple styles. Extracted landmarks and stylized images correspond to groundtruth $\mathcal{L}_B$ and $\mathcal{I}_B$ respectively for equation 5.

Comparison protocol

CycleGAN [15] is considered state-of-the-art in image translation and is closely related to our work in terms of consistency-loss design. StarGAN [26] is also related because it supports multiple attribute transfer using a single generator. However, direct comparison is not possible since none of the two approaches support continuous shape edits. Therefore, to compare with CycleGAN, we use the following pipeline: Given image pair $\mathcal{I}_A$,$\mathcal{I}_B$, which are from domain A and B, CycleGAN translates $\mathcal{I}_A$ to $\widehat{\mathcal{I}_B}$, which has content from $\mathcal{I}_A$ and modality from $\mathcal{I}_B$. This can be achieved with our approach with landmark $\mathcal{L}_A$ unchanged. To compare with StarGAN, we train StarGAN on discrete canonical expressions and compare it with our approach which is conditioned on facial landmarks.

A) Portrait manipulation

Our model is sensitive for edits in eyebrows, eyes, and mouth but less so for nose. The reason is because there is little change in nose shape in our collected database. Nevertheless, our model is able to handle continuous edits because

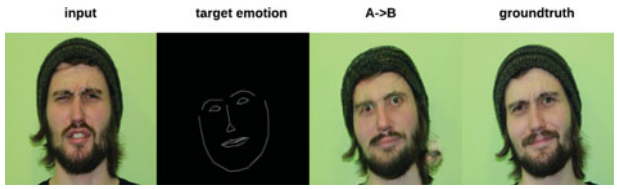

Fig. 6. Failure cases: The reason could be that facial landmarks do not capture well enough details of micro-emotions.

of abundant variations of expressions in data. For example, in Fig. 4 of the paper, the 1st-row achieves face-slimming as a result of pulling landmarks for left (right) cheeks inward, even though there is no slim-face groundtruth as training data. Similarly, the 2nd-row of Fig. 4 shows the mouth fully-closed by merging landmarks for upper and down lips. These two results were obtained with a web tool we developed for interactive portrait manipulation[1], where users can manipulate facial landmarks manually and evaluate the model directly.

Another example for continuous edits is face interpolation. Our model is capable of generating new facial expressions unseen for a certain person. For example, given two canonical expressions (e.g. surprise and smile), we can interpolate[2] a neutral expression in between through interpolating their facial landmarks. The granularity of face edits depends on the gap between two facial landmarks. Here we show a more challenging case, where we interpolate five intermediate transitions given only two real faces. In this case, the quality of the 3rd face is dependent on previous generations (i.e. after the 2nd and 4th fake faces are generated). In Fig. 5, our model can gradually transition a surprise emotion to a smile emotion, beyond canonical emotions.

Compared to discrete conditional labels, facial landmark gives full freedom for continuous shape editing. As can be seen, our model integrates two functions into a single

[1]The tool will be released at: https://github.com/davidsonic/Flexible-Portrait-Manipulation.

[2]Please refer to supplementary material for more details.

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

Fig. 7. Left: original image; Right: generated image.

Fig. 8. Left: original image; Right: generated image.

model: shape edits (when modality is fixed) and style transfer (when landmark is fixed). Not only that, our model supports bidirectional transfer using a single generator, i.e. from natural domain to stylistic domain (1st column to 3rd column or from 5th to 6th) or from stylistic domain to natural domain (3rd column to 5th column). The user can manipulate in any domain and can generate edited shapes in another domain immediately. For example, the 1st row successfully performed simultaneous face-slimming and stylistic transfer.

However, there does exist some failure cases, which generally happen in iCV dataset. In Fig. 6, we tried to manipulate landmark in order to change the original expression (1st column) into groundtruth (4th column) but failed. The closest generated result we can get is shown in the 3rd column. As can be seen, the generated picture fails to

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

Fig. 9. Left: original image; Right: generated image.

Fig. 10. Left: original image; Right: generated image.

mimic the intricate expression displayed in groundtruth. Given that iCV is a micro-emotion dataset, our guess is that 68 landmark is not sufficient for capturing subtle expressions.

An overview of manipulation results are shown in Fig. 1. Some interesting generations were observed. For example, our model seems to be capable of learning some common knowledge, i.e. teeth is hallucinated when mouth is open (1st row, 4th-6th column), after we manipulate the landmarks along the edge of mouth. It is also surprising that our model can preserve obscure details such as earrings (5th row, 4th–6th column). We also notice some artifacts during translation (3rd–4th row, 8th column). The reason is due to the challenge in handling emotion changes and multi-modal transfer with a single model. Having said that, our framework shows promising results in trying to address

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

Fig. 11. Left: original image; Right: generated image.

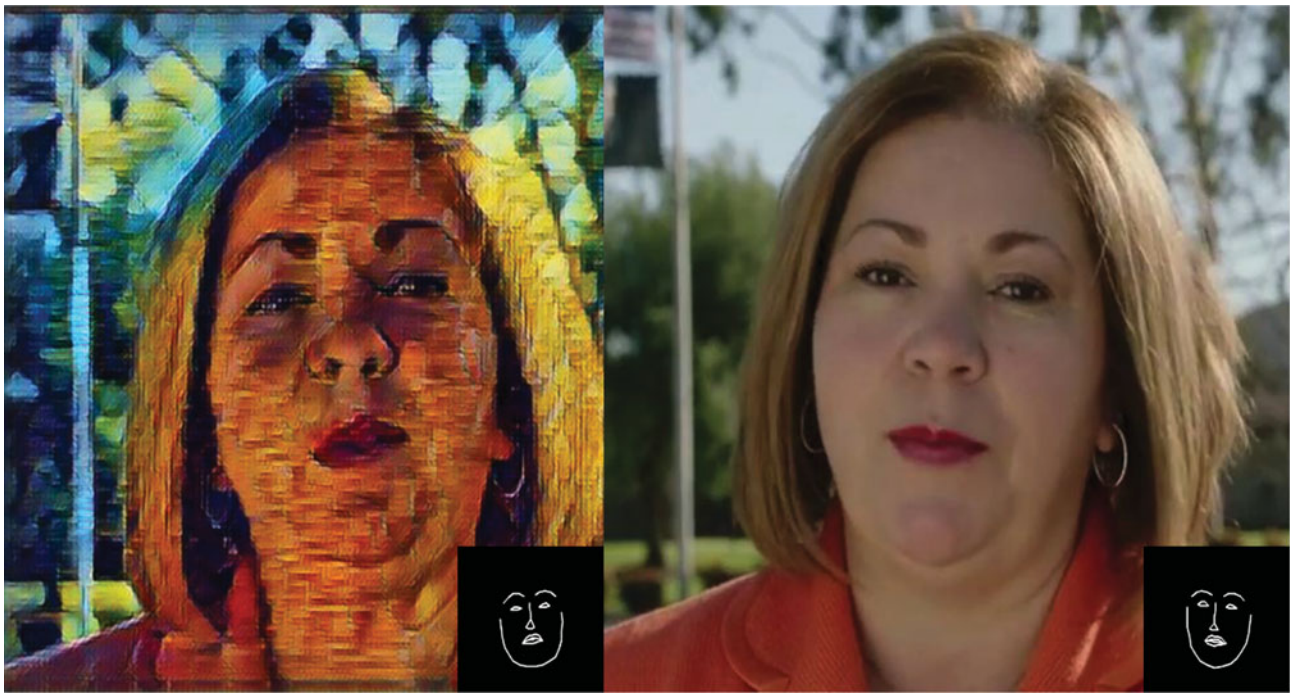

Fig. 12. Left: original image; Right: generated image.

both simultaneously. For high-resolution (512x512) synthesis, please refer to Figs. 7, 8, 9, 10, 11, 12, 13, 14. As can be seen, our model is able to manipulate expression and style based on landmark prior of the target emotion with photo-realistic effect.We refer readers to the supplementary material for more qualitative results. We will also release a website showcasing more results in original resolution (512*512) on Github[3].

[3]https://github.com/davidsonic/Flexible-Portrait-Manipulation.git.

Ablation study

Each component is crucial for the proper performance of the system, which we demonstrate through qualitative figures and quantitative numbers in Table 1. First multi-level adversarial loss is essential for high-resolution generation. As can be seen in Fig. 15, face generated with this design exhibits more fine-grained details and thus more realistic. In Table 1, SSIM drops 1.6% without this loss. Second, texture loss is crucial for pattern similarity during modality

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

Fig. 13. Left: original image; Right: generated image.

Fig. 14. Left: original image; Right: generated image.

transformation. As shown in Fig. 16, PortraitGAN generates more consistent textures compared to StarGAN and CycleGAN. In Table 1, SSIM drops 3.6% if without. Last but not least, $\mathcal{L}_{\text{cyc}}$ and $\mathcal{L}_{id}$ help preserve identity.

B) Perceptual quality

Quantitative analysis

We incorporated 1000 images (500 stylized and 500 natural) to conduct quantitative analysis. For generative adversarial network, two widely used metric for image quality is MSE and SSIM, between the generated image and groundtruth. For MSE, the lower means more fidelity to groundtruth, and for SSIM the higher the better. Table 1 shows quantitative results between CycleGAN, StarGAN, and our approach. As can be seen, our method achieves the best MSE and SSIM score while maintaining relatively fast speed.

Subjective user study

As pointed out in [14], traditional metrics should be taken with care when evaluating GAN, therefore we adopt the

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

Fig. 15. Effect of multi-level adversarial supervision. Left/Right: wo/w multi--level adversarial supervision. Please also refer to the supplementary material for the high-resolution (512 × 512) version.

Table 1. Quantitative evaluation for generated image. Our model is slightly slower than StarGAN but achieves the best MSE and SSIM.

Method	MSE↓	SSIM↑	inference time(s)↓
CycleGAN	0.028	0.473	0.365
StarGAN	0.029	0.483	**0.263**
Ours wo/ $\mathcal{L}_{multi} + \mathcal{L}_{texture}$	0.028	0.472	0.271
Ours wo/ $\mathcal{L}_{multi}$	0.011	0.639	0.277
Ours wo/ $\mathcal{L}_{texture}$	0.013	0.619	0.285
Ours	**0.011**	**0.655**	0.290

same evaluation protocol as in [14, 15, 26, 27] for human subjective study generated images. We collect responses from 10 users (5 experts, 5 non-experts) based on their preferences about images displayed at each group in terms of perceptual realism and identity preservation. Each group consists of one photo input and three randomly shuffled manipulated images generated by CycleGAN [15], StarGAN [26], and our approach. We conducted two rounds of user study where the 1st round has a time limit of 5 s while 2nd round is unlimited. There are in total 100 images and each user is asked to rank three methods on each image twice. Our model gets the best score among three methods as shown in Table 2.

V. CONCLUSIONS

We present a flexible portrait manipulation framework that integrates continuous shape edits and modality transfer into a single adversarial framework. To overcome the technical challenges, we proposed to condition on facial landmark

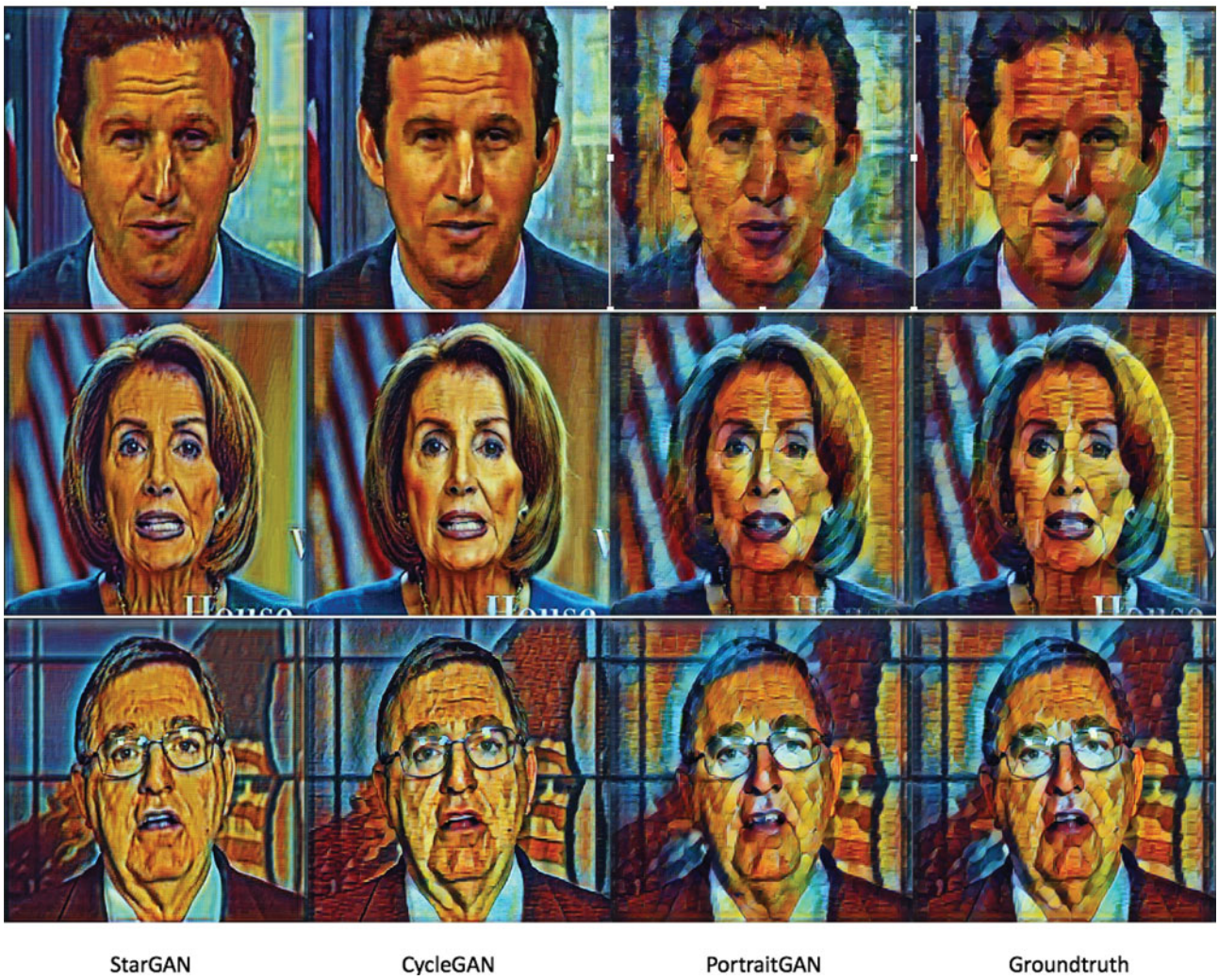

Fig. 16. Comparison with StarGAN and CycleGAN. Images generated by our model exhibit closer texture proximity to groundtruth, due to adoption of texture consistency loss.

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

Table 2. Subjective ranking for different models based on perceptual quality. Our model is close to CycleGAN but is much better than StarGAN.

Method (%)	1st round	2nd round	Average
StarGAN	31.2	32.3	31.75
CycleGAN	33.0	33.5	33.25
Ours	**35.8**	**34.2**	**35.0**

as input and designed a multi-level adversarial supervision structure for high-resolution synthesis. Beyond photo quality, our loss function also takes into account identity and texture into consideration, verified by our ablation studies. Experimental results show the promise of our framework in generating photo-realistic and supporting flexible manipulations. For future work, we would like to improve on the stability of training.

STATEMENT OF INTEREST

None.

REFERENCES

[1] Averbuch-Elor, H., Cohen-Or, D.; Kopf, J.; Cohen, M.F.: Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, **36** (6) (2017), 196.

[2] Chen, Y.-C. *et al.*: Facelet-bank for fast portrait manipulation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 3541–3549.

[3] Fried, O.; Shechtman, E.; Goldman, D.B.; Finkelstein, A.: Perspective-aware manipulation of portrait photos. *ACM Transactions on Graphics (TOG)*, **35** (4) (2016), 128.

[4] Korshunova, I.; Shi, W.; Dambre, J.; Theis, L.: Fast face-swap using convolutional neural networks. In *The IEEE International Conference on Computer Vision*, 2017.

[5] Suwajanakorn, S.; Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, **36** (4) (2017), 95.

[6] Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, 2387–2395.

[7] Levin, A.; Lischinski, D.; Weiss, Yair.: Colorization using optimization, in *ACM SIGGRAPH 2004 Papers*, 2004, 689–694.

[8] Gatys, L.A.; Ecker, A.S.; Bethge, M.: Image style transfer using convolutional neural networks, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016, 2414–2423.

[9] Huang, X.; Belongie, Serge: Arbitrary style transfer in real-time with adaptive instance normalization. *CoRR, abs/1703.06868*, 2017.

[10] Johnson, J.; Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution, in *European Conference on Computer Vision*. Springer, 2016, 694–711.

[11] Luan, F.; Paris, S.; Shechtman, E.; Bala, K.: Deep photo style transfer. *CoRR, abs/1703.07511*, 2017.

[12] Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C.: Image inpainting, in *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, 2000, 417–424.

[13] Liao, J.; Yao, Y.; Yuan, L.; Hua, G.; Kang, S.B.: Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017.

[14] Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A.: Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[15] Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017.

[16] Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y.: Semantic image synthesis with spatially-adaptive normalization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, 2337–2346.

[17] Bau, D. *et al.*: Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018.

[18] Chen, Q.; Koltun, V.: Photographic image synthesis with cascaded refinement networks, in *The IEEE International Conference on Computer Vision (ICCV)*, vol. **1** 2017.

[19] Karras, T.; Aila, T.; Laine, S.; Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Tzeng, E.; Hoffman, J.; Saenko, K.; Darrell, T.: Adversarial discriminative domain adaptation. *Computer Vision and Pattern Recognition (CVPR)*, **1** (2017), 4.

[21] Yang, C.; Lu, X.; Lin, Z.; Shechtman, E.; Wang, O.; Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. **1**, 2017, 3.

[22] Yi, Z.; Zhang, H.; Tan, P.; Gong, M.: Dualgan: Unsupervised dual learning for image-to-image translation. *arXiv preprint*, 2017.

[23] Zhu, J.-Y. *et al.*: Toward multimodal image-to-image translation, in *Advances in Neural Information Processing Systems*, 2017, 465–476.

[24] Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; Van Gool, L.: Pose guided person image generation, in *Advances in Neural Information Processing Systems*, 2017, 405–415.

[25] Pumarola, A.; Agudo, A.; Sanfeliu, A., Moreno-Noguer, F.: Unsupervised person image synthesis in arbitrary poses, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 8620–8628.

[26] Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.

[27] Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. *arXiv preprint arXiv:1711.11585*, 2017.

[28] Zhang, H. *et al.*: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *arXiv: 1710.10916*, 2017.

[29] Blanz, V.; Vetter, T.: 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., 1999, 187–194.

[30] Lau, M.; Chai, J.; Xu, Y.-Q.; Shum, H.-Y.: Face poser: Interactive modeling of 3d facial expressions using facial priors. *ACM Transactions on Graphics (TOG)*, **29** (1) (2009), 3.

[31] Sucontphunt, T.; Mo, Z.; Neumann, U.; Deng, Z.: Interactive 3d facial expression posing through 2d portrait manipulation, in *Proceedings of Graphics Interface 2008*. Canadian Information Processing Society, 2008, 177–184.

[32] Kemelmacher-Shlizerman, I.; Suwajanakorn, S.; Seitz, S.M.: Illumination-aware age progression, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, 3334–3341.

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press

[33] Blanz, V.; Basso, C.; Poggio, T.; Vetter, T.: Computer Graphics Forum, Vol. 22. No. 3. Oxford, UK: Blackwell Publishing, Inc, 2003.

[34] Hoffman, J. *et al.*: Cycada: Cycle-consistent adversarial domain adaptation. *arXiv preprint arXiv:1711.03213*, 2017.

[35] Liu, M.-Y.; Breuel, T.; Kautz, J.: Unsupervised image-to-image translation networks, in *Advances in Neural Information Processing Systems*, 2017, 700–708.

[36] Qian, S. *et al.*: Make a face: Towards arbitrary high fidelity face manipulation, in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, 10033–10042.

[37] Wang, W.SEP, Alameda-Pineda, X.; Xu, D.; Fua, P.; Ricci, E.; Sebe, N.: Every smile is unique: Landmark-guided diverse smile generation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, 7083–7092.

[38] Lassner, C., Pons-Moll, G.; Gehler, P.V.: A generative model of people in clothing. *arXiv preprint arXiv:1705.04098*, 2017.

[39] Siarohin, A.; Sangineto, E.; Lathuiliere, S.; Sebe, N.: Deformable gans for pose-based human image generation. *arXiv preprint arXiv:1801.00055*, 2017.

[40] Walker, J.; Marino, K.; Gupta, A.; Hebert, M.: The pose knows: Video forecasting by generating pose futures, in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, 3352–3361.

[41] Reed, S.E.; Akata, Z.; Mohan, S.; Tenka, S.; Schiele, B.; Lee, H.: Learning what and where to draw, in *Advances in Neural Information Processing Systems*, 2016, 217–225.

[42] Choi, Y.; Uh, Y.; Yoo, J; Ha, J.-W.: Stargan v2: Diverse image synthesis for multiple domains, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 8188–8197.

[43] Huang, X.; Liu, M.-Y.; Belongie, S.; Kautz, J.: Multimodal unsupervised image-to-image translation, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, 172–189.

[44] Goodfellow, I.: Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[45] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X.: Improved techniques for training gans, in *Advances in Neural Information Processing Systems*, 2016, 2234–2242.

[46] Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.K.; Wang, Z.; Smolley, S.P.: Least squares generative adversarial networks, in *2017 IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2017, 2813–2821.

[47] Simonyan, K.; Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.J.; Hawk, S.T.; Van Knippenberg, A.D.: Presentation and validation of the radboud faces database. *Cognition and emotion*, **24** (8) (2010), 1377–1388.

[49] Lüsi, I. *et al.*: Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases, in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 2017, 809–813.

[50] King, D.E.: Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, **10** (2009), 1755–1758.

Jiali Duan is currently a PhD student in MCL, Viterbi Engineering School of USC, under the supervision of Prof. C.-C. Jay Kuo. He received his Master degree in 2017 with Presidential Award from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include generative adversarial learning and deep metric learning.

Xiaoyuan Guo is currently a third-year PhD at Emory University supervisied by Prof. Ashish Sharma. Her research involves object detection, object segmentation, and medical image analysis(whole-slide microscopy image and radiology image) using deep neural networks.

Dr. C.-C. Jay Kuo received the B.S. degree from the National Taiwan University, Taipei, in 1980 and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively, all in Electrical Engineering. From October 1987 to December 1988, he was Computational and Applied Mathematics Research Assistant Professor in the Department of Mathematics at the University of California, Los Angeles. Since January 1989, he has been with the University of Southern California (USC). He is presently USC Distinguished Professor of Electrical Engineering and Computer Science and Director of the Multimedia Communication Laboratory. Dr. Kuo is a Fellow of AAAS, IEEE and SPIE.

https://doi.org/10.1017/ATSIP.2020.20 Published online by Cambridge University Press