Learning to Caricature via Semantic Shape Transform

Wenqing Chu^{1,4} · Wei-Chih Hung² · Yi-Hsuan Tsai³ · Yu-Ting Chang² · Yijun Li⁵ · Deng Cai¹ · Ming-Hsuan Yang^{2,6}

Received: 10 August 2020 / Accepted: 3 June 2021 / Published online: 9 July 2021
© The Author(s) 2021

Abstract

Caricature is an artistic drawing created to abstract or exaggerate facial features of a person. Rendering visually pleasing caricatures is a difficult task that requires professional skills, and thus it is of great interest to design a method to automatically generate such drawings. To deal with large shape changes, we propose an algorithm based on a semantic shape transform to produce diverse and plausible shape exaggerations. Specifically, we predict pixel-wise semantic correspondences and perform image warping on the input photo to achieve dense shape transformation. We show that the proposed framework is able to render visually pleasing shape exaggerations while maintaining their facial structures. In addition, our model allows users to manipulate the shape via the semantic map. We demonstrate the effectiveness of our approach on a large photograph-caricature benchmark dataset with comparisons to the state-of-the-art methods.

Keywords Caricature generation · Dense shape transformation · Semantic map

Communicated by Gang Hua.

✉ Deng Cai
dengcai@cad.zju.edu.cn

Wenqing Chu
wqchu16@gmail.com

Wei-Chih Hung
whung8@ucmerced.edu

Yi-Hsuan Tsai
wasidennis@gmail.com

Yu-Ting Chang
ychang39@ucmerced.edu

Yijun Li
yijli@adobe.com

Ming-Hsuan Yang
mhyang@ucmerced.edu

- ¹ State Key Lab of CAD&CG, Zhejiang University, Hangzhou, Zhejiang, China
- ² Electrical Engineering and Computer Science, University of California at Merced, Merced, CA, USA
- ³ NEC Laboratories America, Santa Clara, CA, USA
- ⁴ Tencent Youtu Lab, Shanghai, China
- ⁵ Adobe Research, San Jose, CA, USA
- ⁶ Yonsei University, Seoul, Korea

1 Introduction

Caricature is a rendered image by abstracting or exaggerating certain facial features (e.g., contour, eyes, ear, eyebrows, mouth, and nose) of a person to achieve humorous or sarcastic effects (Fig. 1). Caricatures are widely used to depict celebrities or politicians for certain purposes in all kinds of media. However, generating visually pleasing caricatures with proper shape distortions usually requires professional artistic skills with creative imaginations, which is a challenging task for common users. Therefore, it is of great interest if caricatures can be generated from normal photos effectively or in a way that users are allowed to flexibly manipulate the output with user controls.

One crucial factor to generate a desirable caricature is to distort facial components properly, i.e., to render personal traits with certain exaggerations. Numerous efforts have been made to perform shape exaggeration by computing warping parameters between photos and caricatures from user-defined shapes (Akleman et al. 2000) or hand-crafted rules (Brennan 2007; Liao et al. 2004). However, such methods may have limitations on generating diverse and visually pleasing results due to inaccurate shape transformations. Recently, image-to-image translation (Isola et al. 2017; Zhu et al. 2017a; Lee et al. 2018) and neural style transfer (Gatys et al. 2016; Johnson et al. 2016; Li et al. 2017a) algorithms have been developed,

Fig. 1 Examples of normal photos, hand-drawn caricatures, and a set of caricature outputs generated by the proposed method. Our approach is able to render a diverse set of visually pleasing caricatures

but most techniques can be applied to two domains with local texture variations, not for scenarios where large shape discrepancy exists.

In this work, we aim to create shape exaggerations on standard photos with shape transformations similar to those drawn by artists. Meanwhile, a rendered caricature should still maintain the facial structure and personal traits. Different from existing methods that only consider facial landmarks or sparse points (Cao et al. 2018; Shi et al. 2019), we use a semantic face parsing map, i.e., a dense pixel-wise parsing map, to guide the shape transformation process. As such, this can provide more accurate mapping for facial details, e.g., shapes of eyebrows, noses, and face contours, to name a few. Specifically, given an unpaired caricature with a normal photo, we leverage the cycle consistency strategy and an encoder–decoder architecture to model the shape transformation. Nevertheless, operating this learning process in the image domain may involve noise from unnecessary information in pixels. Instead, we learn the model directly on the face parsing map, which is the shape transformation of interest. To learn effective shape transformation, we design a spatial transformer network (STN) to allow larger and flexible shape changes, while a few loss functions are introduced to better maintain facial structures.

To evaluate the proposed framework, we conduct experiments on the photo-caricature benchmark dataset (Huo et al. 2018). We perform extensive ablation studies to validate each component of the proposed shape transformation algorithm. We conduct qualitative and quantitative experiments with user studies to demonstrate that the proposed approach performs favorably against existing image-to-image translation and caricature generation methods. Furthermore, our model allows users to select the desired caricature semantic shape, with the flexibility to manipulate the parsing map to generate

preferred shapes and diverse caricatures. The main contributions of the paper are as follows:

- We design a shape transformation model to facilitate the photo-to-caricature generation with visually pleasing shape exaggerations that approach the quality of hand-drawn caricatures.
- We introduce the face parsing map as the guidance for shape transformation and learn a feature embedding space for face parsing, which allows users to explicitly manipulate the degree of shape changes in the rendered caricature.
- We evaluate the proposed algorithm on a large caricature benchmark dataset and demonstrate favorable results against existing methods.

2 Related Work

Image Translation Numerous methods based on GANs (Goodfellow et al. 2014) have been recently developed for image translation for paired (Isola et al. 2017; Chang et al. 2018), unpaired (Zhu et al. 2017a; Kim et al. 2017; Liu et al. 2017), and multimodal (Zhu et al. 2017b; Lee et al. 2018; Huang et al. 2018) settings. While most GANs based methods require large image sets for training, neural style transfer models (Gatys et al. 2015, 2016) only need a single style image as the reference. A number of methods have since been developed (Johnson et al. 2016; Li et al. 2017a; Chen et al. 2017; Liao et al. 2017; Li et al. 2017c) to improve style translation or runtime performance. In the photo-to-caricature task, however, neither GAN-based approaches or neural style transfer methods take the large shape discrepancy across domains into account. Unlike existing image transla-

Table 1 Comparisons of caricature generation methods

Method	Akleman et al. (2000)	Brennan (2007)	WarpGAN Shi et al. (2019)	CariGANs Cao et al. (2018)	Ours
Input requirement	Photo	Line drawing	Photo	Photo	Photo
Transformation space	Landmark points	None	Control points	Landmark points	Face parsing map
Output shape control	User-defined	Rules	Warping coefficient	Warping coefficient	Face parsing map

Existing approaches mainly leverage the sparse control points for controlling shape exaggerations. In contrast, we propose a semantic shape transform method to render visually pleasing and plausible caricatures via using the dense face parsing map. In addition, this allows us to control the shape changes directly on the parsing map, while other methods have no control or can only adjust the shape via the one-dimensional warping coefficient

tion methods, our algorithm enables shape exaggerations by utilizing an encoder–decoder architecture with the guidance of a face parsing map to densely model the shape transformation.

Caricature Rendering Learning to caricature from photos is mainly concerned with modeling shape transformation. However, it has not been widely explored due to the large domain gap between photos and caricatures. Some early methods (Akleman 1997; Akleman et al. 2000) rely on user-defined source and target shapes to compute the warping parameters, but the rendered images do not exhibit caricature styles. Several rule-based methods (Luo et al. 2002; Liao et al. 2004; Brennan 2007) are developed to perform shape exaggeration for each facial component.

Recently, CycleGAN-based image translation methods (Li et al. 2020; Zheng et al. 2019) have been proposed for caricature generation with facial landmarks as conditional constraints. However, the geometric structures of the images generated with these algorithms are still close to the original photos, with limited exaggerated effects. To increase shape exaggeration, Cao et al. (2018) use a geometric exaggeration model by leveraging landmark positions to predict key points in the subspace formed by principal components. The WarpGAN model (Shi et al. 2019) learns to directly predict a set of control points used to warp the input photo based on unpaired adversarial learning. Despite showing promising results, their shape exaggerations are still limited due to the use of sparse landmarks or control points. In contrast, we use a pixel-wise parsing map and model shape transformation, thereby generating plausible exaggerations while retaining the facial traits of the input image. Furthermore, due to the learned representations in the parsing space, users can explicitly manipulate the facial map to generate preferred shapes. In Table 1, we compare our algorithm with existing methods in terms of shape transformation and requirements.

In addition to 2D image-based methods, several 3D geometry-based approaches (Wu et al. 2018; Han et al. 2018; Chen et al. 2020) have been developed to manipulate caricature images. Wu et al. (2018) introduce an intrinsic deformation representation which reconstructs a 3D caricature image from sparse 2D landmarks. On the other hand, Han et al. (2018) utilize user-defined line sketches to edit

a 3D caricature model for synthesizing caricature images. However, it is complex for users to define appropriate and exaggerated sketches. In addition, the generated images do not contain caricature textures. To deform an artist-drawn caricature according to a given normal face expression, Chen et al. (2020) develop a method based on Wu et al. (2018) and Cao et al. (2018) to generate caricatures, which are then further re-rendered with dynamic texture generated from a conditional generative adversarial network.

Spatial Transformer Network The spatial transformer networks (STNs) (Jaderberg et al. 2015) are developed to improve object recognition performance by reducing input geometric variations. Numerous variants have since been developed for a wide range of computer vision applications (Wu et al. 2017; Zhou et al. 2018; Wang et al. 2017; Dai et al. 2017; Ganin et al. 2016; Park et al. 2017; Shu et al. 2018) that require geometric constraints. Close to our work is the method proposed by Lin et al. (2018) in which low-dimensional warping parameters are learned to manipulate foreground objects for image composition. In our task, we also introduce an STN to predict warping parameters to enable shape exaggeration on normal photos. In contrast, we need denser and more complex deformations instead of low-dimensional affine transformation (Jaderberg et al. 2015) or homography transformation (Lin et al. 2018). Furthermore, our shape transformation network leverages the facial parsing maps as an additional input to focus on the semantic facial structure.

3 Algorithmic Overview

In this section, we introduce the overall framework of the proposed caricature generation method (see Fig. 2).

Model Inputs Existing methods (Zheng et al. 2019; Li et al. 2020; Cao et al. 2018) use images or sparse landmarks as the inputs to capture facial structure information for shape exaggeration. However, these approaches are not effective for capturing large shape transformations in caricatures as their appearances of facial components are significantly different from the ones in normal photos. In this paper, we use the facial parsing map which provides dense semantic information to

Fig. 2 Overall framework of the proposed caricature generation method. Given an input photo, we first obtain its face parsing map and then retrieve caricature parsing maps from a large-scale database. Second, we feed these two maps into the proposed shape transformation

network to predict the warping parameters and produce the deformed photo. Third, we utilize a style transfer network to generate the final output with caricatured textures

facilitate computing correspondences between faces of the photo and caricature. In practice, we adopt the adapted parsing model in Chu et al. (2019) to account for the domain shift issue. Most recent parsing models only work on normal faces but not on caricature faces. Therefore, we choose (Chu et al. 2019) as our parsing model, which achieves state-of-the-art performance on caricatures and also performs well on face images. More details about the training strategy and network architecture of the parsing model can be found in Chu et al. (2019). Note that during training the parsing network, only 17 facial landmarks of caricatures are provided like other caricature generation methods (Cao et al. 2018; Li et al. 2020). During the testing stage, we do not need any key-point annotations. Although the parsing quality could not be always satisfactory (e.g., IoU is 86.5% on facial skins), we design loss functions (in Sect. 4.3) to make the deformed shape smooth and plausible.

Caricature Retrieval We note that the selection of target semantic shapes is useful and important for diverse caricature generation. In this work, we develop a parsing map retrieval model with the large-scale dataset WebCaricature (Huo et al. 2018) as the gallery images. Given a photo parsing map $\mathbf{P}_{pho}$, we aim to retrieve suitable caricatures $\mathbf{P}_{cari}$ as inputs to calculate the following shape transformation. The key challenge is how to find the appropriate embedding space to perform retrieval. Here, we assume that if $\mathbf{P}_{pho}$ and $\mathbf{P}_{cari}$ belong to the same identity, $\mathbf{P}_{cari}$ could be a good reference. Thus, they should be close to each other in the embedding space. As shown in Fig. 3, we utilize the contrastive loss $\mathcal{L}_{contrastive}$ (Chopra et al. 2005) to enforce the caricature and photo embeddings of the same person being close to each other, while the reconstruction loss $\mathcal{L}_{rec}$ helps preserve the content of the parsing maps through minimizing the Euclidean distance between the input parsing maps and reconstructed ones.

The encoder consists of five basic dense blocks (Huang et al. 2017) and a flatten layer to obtain a global 128-dimensional vector z_{cari}/z_{pho} , while the decoder is composed of five symmetric dense blocks with transposed convolution layers. The contrastive loss for positive and negative pairs is defined as:

$$\mathcal{L}_{contrastive} = \begin{cases} \|z_{cari} - z_{pho}\|_2, & \text{positive,} \\ \max(m - \|z_{cari} - z_{pho}\|_2, 0), & \text{negative,} \end{cases} \quad (1)$$

where the hyper-parameter m is the margin set to 2 in this work. During testing, we pre-compute the caricature embedding of our gallery caricatures in the training set, and use the photo encoder to compute the embedding of the testing photo. To find multiple caricature parsing maps, we first retrieve the top 5 caricature embeddings that are closest to the photo one based on the Euclidean distance, and then use their associated caricature parsing maps as our final outputs. **Style Transfer** We synthesize output images by considering the input photo as content and the retrieved caricature as style in a reference based approach which is widely used in multimodal image translation (Lee et al. 2018; Huang et al. 2018) and neural style transfer (Li et al. 2017b, 2018). In this work, we use a feed-forward style transfer model (Huang and Belongie 2017) and utilize the deformed photo as the content image, while the caricature is regarded as the style image. We generate the content feature maps from the deformed photo and extract the style features from the reference caricature. The adaptive instance normalization method (Huang and Belongie 2017) is then utilized to modify the content feature maps according to the extracted style features. The content feature maps are further processed by a series of convolutions to render the final caricature image.

Fig. 3 Framework of caricature retrieval. Our goal of training the retrieval model is to learn the photo and caricature embedding on parsing maps, so that given the photo parsing map, the model is able to retrieve proper caricature maps during testing

Overall Pipeline Given the input photo, we first use a caricature retrieval model to automatically recommend the proper caricature parsing map, along with the photo parsing map as inputs to the next phase. Second, to better mimic the process of drawing caricatures, we decompose the caricature generation pipeline into two stages: shape transformation and style transfer. In the first stage, we propose a semantic-aware shape transformation network to learn dense warping parameters that enable shape exaggerations for the input photo. Once obtaining the image with the deformed shapes on facial components, we use a reference based feed-forward style transfer network (Huang and Belongie 2017) to perform the photo-to-caricature texture translation and obtain the final caricature output.

4 Semantic Shape Transformation

Given a portrait photo and a recommended caricature, our algorithm transforms the portrait photo to have a similar facial structure to the recommended caricature. In contrast to most image translation methods that learn pixel mapping in the image space, we learn the dense pixel correspondence between inputs and outputs on facial components through the face parsing map. We use an encoder–decoder architecture, where the encoder extracts feature representations of the parsing map and the decoder composed of an STN module estimates the warping parameters, i.e., the transformation from the parsing map of the photo to the caricature one. It is worth noting that learning in the semantic space is easier than the original image space due to less appearance discrepancy. The overall architecture and designed loss functions are presented in Fig. 4.

4.1 Encoder

Here, we describe how to obtain compact representations for the facial structures of caricatures and photos. We first denote the face parsing maps of the photo and caricature as $\mathbf{P}_{pho}$ and $\mathbf{P}_{cari} \in \mathbb{R}^{C \times H \times W}$, where the image height and weight are denoted as H and W , and C is the number of the facial component category. As a result, in each channel, there is a binary map to describe each facial component. Considering the distributions of facial structures between photos and caricatures are quite different, we use two independent encoders in which each network consists of several dense blocks. As such, the encoded feature is a compact 128-dimensional vector. In the following, we denote them as z_{pho} and z_{cari} .

4.2 Decoder

Once obtaining the latent feature z from the encoder to represent the facial structure, the goal of the decoder is to predict the dense correspondence denoted by a tensor $\mathbf{D} \in \mathbb{R}^{2 \times H \times W}$. Specifically, $(\mathbf{D}_{1,i,j}, \mathbf{D}_{2,i,j})$ indicates the corresponding target position when warping each pixel (i, j) from the recommended caricature to the input photo. To perform warping, we first concatenate the latent encoding z_{pho} and z_{cari} , and then feed this feature to the decoder. Here, we introduce a spatial transformer network module to generate the 2D shape transformation parameters for each pixel. As a result, we can apply the differentiable bilinear sampling operation to the photo parsing map $\mathbf{P}_{pho}$ and obtain the deformed photo parsing map $\mathbf{P}_{fake}$.

The ensuing question is how to enforce the generated $\mathbf{P}_{fake}$ to resemble the parsing map of the real caricature. To this end, we impose three constraints on $\mathbf{P}_{fake}$: (1) the generated parsing map should be densely reconstructed with respect to the recommended one in the semantic space; (2) a cycle consistency is measured between the generated parsing map and the recommended one; (3) a coordinate-based reconstruction is used to regularize alignment at locations of facial components. In the following, we introduce the loss functions based on the above-mentioned constraints.

4.3 Reconstruction in the Semantic Space

Reconstruction Loss First, a natural way to enforce similarity is to require $\mathbf{P}_{fake}$ and $\mathbf{P}_{cari}$ identical at each pixel. Therefore, we minimize the L1 distance between them as below:

$$\mathcal{L}_{rec}^p = \frac{1}{C \times H \times W} \sum_{i=1}^C \sum_{j=1}^H \sum_{k=1}^W \|\mathbf{P}_{cari}^{(i,j,k)} - \mathbf{P}_{fake}^{(i,j,k)}\|_1. \quad (2)$$

However, we find that this function is not effective to reconstruct every facial component. For instance, face skin regions

Fig. 4 Proposed shape transformation network. We first feed face parsing maps to each encoder to extract latent feature encodings z . Second, we concatenate two features as the input of the decoder to predict the warping parameters D and thus produce the parsing map $\mathbf{P}_{fake}$ of the deformed photo. A reconstruction loss $\mathcal{L}_{rec}$ is then computed between $\mathbf{P}_{pho}$ and the parsing map of the caricature $\mathbf{P}_{cari}$. To ensure

local details and cycle consistency, we further incorporate an adversarial loss $\mathcal{L}_{adv}$ on $\mathbf{P}_{fake}$ and a cycle consistency loss $\mathcal{L}_{cyc}$ between $\mathbf{P}_{pho}$ and the reconstructed parsing map of the photo $\mathbf{P}_{cyc}$. In addition, we add a coordinate-based loss $\mathcal{L}_{coo}$ to constrain the alignment based on pixel locations

can have a large overlap between $\mathbf{P}_{cari}$ and $\mathbf{P}_{fake}$, while smaller components such as eyes do not usually have spatial overlaps. To handle this issue, we design a location-aware metric which measures the distance of the center in the same facial component between $\mathbf{P}_{cari}$ and $\mathbf{P}_{fake}$. We average the locations of each pixel to obtain a mean location (x^c, y^c) for each facial component c and obtain:

$$\begin{cases} x^c = \frac{1}{H \times W} \sum_{j=1}^H \sum_{k=1}^W \mathbf{P}^{(c,j,k)} \times j. \\ y^c = \frac{1}{H \times W} \sum_{j=1}^H \sum_{k=1}^W \mathbf{P}^{(c,j,k)} \times k. \end{cases} \quad (3)$$

This location-aware reconstruction loss is defined by:

$$\mathcal{L}_{rec}^l = \frac{1}{C} \sum_{i=1}^C (\|\mathbf{x}_{fake}^i - \mathbf{x}_{cari}^i\|_2 + \|\mathbf{y}_{fake}^i - \mathbf{y}_{cari}^i\|_2). \quad (4)$$

Furthermore, we define a global alignment loss by matching the number of pixels in each facial component:

$$\mathcal{L}_{rec}^n = \frac{1}{C} \sum_{i=1}^C \left\| \left(\sum_{j=1}^H \sum_{k=1}^W \mathbf{P}_{cari}^{(i,j,k)} \right) - \left(\sum_{j=1}^H \sum_{k=1}^W \mathbf{P}_{fake}^{(i,j,k)} \right) \right\|_2. \quad (5)$$

The full objective function for reconstruction is:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec}^p + \lambda_l \mathcal{L}_{rec}^l + \lambda_n \mathcal{L}_{rec}^n, \quad (6)$$

where the hyper-parameters λ_l and λ_n control the importance of each term. In this work, we use $\lambda_l = \lambda_n = 2$ in all experiments. To encourage our loss functions to pay attention to small components, we introduce a weight λ_{comp}^c adaptively computed as the reciprocal pixel-ratio in each facial component c of the entire image, i.e., $\mathcal{L}_{rec} = \sum_{c=1}^C \lambda_{comp}^c \mathcal{L}_{rec}^c$, where $\mathcal{L}_{rec}^c$ indicates the loss $\mathcal{L}_{rec}$ for each semantic category c .

Adversarial Loss The reconstruction-based loss can be used to recover the global structure. On the other hand, the GAN-based adversarial loss (Goodfellow et al. 2014) has been shown to be effective for preserving local details. In this work, we adopt a similar approach based on the GAN loss but in the semantic parsing space. To ensure the generated $\mathbf{P}_{fake}$ to look like a realistic caricature, we employ adversarial learning to match the distribution of $\mathbf{P}_{fake}$ to the one of caricatures, i.e., $\mathbf{P}_{cari}$. We adopt the same training scheme and loss function as the Wasserstein GAN model (Arjovsky et al. 2017). We denote this adversarial loss for the generator as $\mathcal{L}_{adv}$.

4.4 Cycle Consistency

Similar to the CycleGAN model (Zhu et al. 2017a), we add the cycle consistency to make the transformation more stable. Specifically, we first input $\mathbf{P}_{fake}$ to the same caricature encoder and extract the feature z_{fake} . We then concatenate z_{fake} and z_{pho} , feed them into a decoder and recover the original face parsing map of the photo, denoted as $\mathbf{P}_{cyc}$. Here, we utilize the same reconstruction-based loss as in (6), defined as $\mathcal{L}_{cyc}$, in which the only difference is that we compute the loss between $\mathbf{P}_{cyc}$ and $\mathbf{P}_{pho}$.

Coordinate-based Loss The above-mentioned loss functions are all based on the parsing map $\mathbf{P}$ or regularization on $\mathbf{D}$, which constrain the output to be consistent in the semantic space. Nevertheless, the constructed pixels may not be well aligned in the coordinate space. To address this issue, we further introduce a coordinate-based loss when computing the cycle consistency. Instead of only considering the parsing map $\mathbf{P}_{pho}$, we construct a coordinate map $\mathbf{M}_{pho} \in \mathbb{R}^{2 \times H \times W}$, where $\mathbf{M}_{pho}^{(i,j)} = (i, j)$ indicates the spatial location. After obtaining the reconstructed $\mathbf{P}_{cyc}$, we convert it to a coordinate map $\mathbf{M}_{cyc}$. Since this $\mathbf{M}_{cyc}$ has been operated through two decoders with the estimated warping parameters, the newly warped coordinates may not be aligned with $\mathbf{M}_{pho}$. Thus we minimize the following loss in the coordinate space:

$$\mathcal{L}_{coo} = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W \|\mathbf{M}_{pho}^{(i,j)} - \mathbf{M}_{cyc}^{(i,j)}\|_1. \quad (7)$$

This spatially-variant consistency loss in the coordinate space can constrain per-pixel correspondence to be one-to-one and reversible, which reduces the artifacts inside each facial part.

4.5 Overall Objective

The overall objective function for the proposed semantic shape transformation network includes the reconstruction/adversarial loss to help recover the semantic parsing map and the cycle consistency/coordinate-based loss to ensure consistency:

$$\mathcal{L}_{shape} = \lambda_r \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{cyc} + \mathcal{L}_{coo}. \quad (8)$$

In this work, we regard the reconstruction term as a critical one and use $\lambda_r = 500$ in the following experiments.

4.6 Implementation Details

The caricature retrieval model is based on an encoder-decoder architecture. For the encoder, we use five basic dense blocks (Huang et al. 2017) to generate the latent embedding for retrieval, while the decoder contains five symmetric dense

blocks with transposed convolution layers for upsampling to the input size. It is worth mentioning that we only need the encoder during the inference stage. To determine the positive and negative pairs, we leverage the identity information for finding the appropriate reference. Given a photo parsing map, we randomly choose a caricature parsing map that belongs to the same identity as the positive sample, and use others as negatives. Note that the evaluated methods for caricature generation (Li et al. 2020; Shi et al. 2019) also use the same identity information and thus the experimental comparisons are fair.

The shape transformation model shares a similar network structure with the caricature retrieval module but is learned with separate parameters due to different objectives. For the encoder, we utilize four basic dense blocks (Huang et al. 2017) to extract features, followed by a flatten layer to obtain a global 128-d code for representing the semantic parsing map. The decoder consists of four symmetric dense blocks with transposed convolution layers for upsampling. In the inference stage, both the encoder and decoder are used.

We implement our method using PyTorch and train the model with a single Nvidia 1080Ti GPU. During training, we utilize the Adam (Kingma and Ba 2015) optimizer and use the batch size as 32. Similar to CycleGAN, we set the initial learning rate as 0.0001 and fix it for the first 300 epochs, and linearly decay the learning rate for another 300 epochs. The source code of the proposed method is available at <https://github.com/wenqingchu/Semantic-CariGANs>.

5 Results and Analysis

We evaluate the proposed algorithm and relevant methods on the large WebCaricature dataset (Huo et al. 2018). This photo-caricature benchmark dataset contains 5974 photos and 6042 caricatures collected from the web. We use the provided landmarks to crop/align faces and then resize them into 256×256 pixels. In addition, we randomly select 500 photos as the test set and use the rest as the training set. We perform qualitative and quantitative experiments to demonstrate the effectiveness of the proposed algorithm.

5.1 Ablation Study on Shape Transformation

We first analyze the quality of the proposed semantic shape transformation algorithm in this section.

Loss Functions We evaluate the effectiveness of the proposed loss functions quantitatively. We randomly select 200 caricatures as reference images, guiding the test photos to generate transformed caricature-like output (without the style transfer stage). Next, we utilize the parsing map of the reference caricature and verify whether the parsing map of the transformed photo is similar to the original parsing map in caricature.

Fig. 5 Visual comparisons for loss functions. From left to right, we first show the raw image and then results of gradually adding $\mathcal{L}_{rec}$, $\mathcal{L}_{cyc}$, $\mathcal{L}_{coo}$

Table 2 Ablation studies on different loss functions

Loss functions	mIoU	pixAcc
w/o $\mathcal{L}_{rec}$	23.25	85.76
w/o $\mathcal{L}_{adv}$	58.19	94.66
w/o $\mathcal{L}_{cyc}$	58.42	94.70
w/o $\mathcal{L}_{coo}$	60.05	95.49
$\mathcal{L}_{shape}$ (ours)	61.66	95.60

Bold values indicate the best performance compared to other settings

We evaluate the performance with mean intersection-over-union (mIoU) and pixel accuracy (pixAcc), which are the common metrics for semantic segmentation.

Table 2 shows the results of this ablation study. Without the reconstruction loss $\mathcal{L}_{rec}$, the model has no control to transform face shapes to be similar to the reference parsing map, resulting in the lowest accuracy. Without the adversarial loss $\mathcal{L}_{adv}$ or cycle consistency loss $\mathcal{L}_{cyc}$, we observe that the training is less stable and is unable to preserve details. Finally, although removing $\mathcal{L}_{coo}$ only slightly degrades the parsing mIoU, we notice that it is a crucial component to perform more accurate alignment for facial components.

In addition, we also provide visual comparisons in Fig. 5 to verify the effectiveness of the designed loss functions qualitatively. Here, we gradually add $\mathcal{L}_{rec}$, $\mathcal{L}_{cyc}$, and $\mathcal{L}_{coo}$ to our model. We find that only using $\mathcal{L}_{rec}$ leads to severe artifacts, e.g., around the eye and skin regions. This is because this model only learns to change the semantic shapes and overlooks the inherent structure of the facial component. Introducing the additional $\mathcal{L}_{cyc}$ could constrain the mapping function and produce less distortion. Finally, we observe that adding $\mathcal{L}_{coo}$ based on the coordinate space makes the con-

structed pixels well aligned and produces visually pleasing results. The reason is that $\mathcal{L}_{coo}$ penalizes pixels with large distortion and enforces all pixels in the same semantic region to have smoother translations. For λ_r in (8), we set it to 500 because we regard the reconstruction term as a critical one. When λ_r is too small, the shape exaggeration is small and not similar to the reference. When λ_r is too large, other loss terms are less important, leading to artifacts as in Fig. 5.

We note that the proposed loss functions facilitate learning pixel-wise correspondences between the photo and caricature with the parsing maps. First, the reconstruction loss constrains the region-based correspondence and is able to make the generated semantic shape plausible as a good starting point. Second, we add the coordinate-based loss and cycle consistency loss to stabilize the transformation process and reduce the artifacts inside each facial component as shown in Fig. 5. Since the semantic map inside each facial component is spatial-invariant, we introduce the spatial-variant coordinate map, which helps the model to learn better per-pixel correspondence. This approach also encourages that the per-pixel flow-field inside each facial component to be one-to-one and invertible, such that many-to-one or one-to-many mapping for per-pixel correspondence are penalized. Hence, the facial structure in the original photo is preserved.

Reconstruction Loss We evaluate the effectiveness of the proposed components in the reconstruction loss functions qualitatively. In (6), λ_l and λ_n are used for the location-aware reconstruction and global alignment loss in the reconstruction term. These two terms measure the global statistical information of the parsing maps. When λ_l is too small or zero, the shape transformation cannot learn to move the specific facial region to a non-overlap location, while increasing λ_l does not provide obvious improvement. When λ_n is too

Fig. 6 Visual results for different loss components in the reconstruction function

Table 3 Ablation studies on different loss components in the reconstruction loss

Loss functions	mIoU	pixAcc
w/o $\mathcal{L}_{rec}^p$	38.38	86.14
w/o $\mathcal{L}_{rec}^l$	61.19	95.42
w/o $\mathcal{L}_{rec}^n$	61.43	95.49
$\mathcal{L}_{shape}$ (ours)	61.66	95.60

Bold values indicate the best performance compared to other settings

small or zero, the shape transformation is enlarged unstably in some specific facial regions with too much or too small transformations, while increasing λ_n also does not bring obvious improvement. We perform an ablation study to demonstrate the effectiveness of different loss components in the reconstruction loss $\mathcal{L}_{rec}$ and show the quantitative results in Table 3 and visual results in Fig. 6. The results demonstrate that without $\mathcal{L}_{rec}^p$, the segmentation accuracy is very low and the shape transformation network does not work. In addition, without $\mathcal{L}_{rec}^l$ or $\mathcal{L}_{rec}^n$, the segmentation accuracy is slightly worse and there exists obviously visual artifacts.

Comparisons to the Landmark Input In this work, we utilize the parsing maps as the input to learn shape transformation, while previous methods (Cao et al. 2018; Li et al. 2020) mainly leverage sparse facial landmarks for shape exaggeration. We present an ablation study of using different inputs for performance evaluation. Similar to Cao et al. (2018) and Li et al. (2020), we consider two approaches to make use of the labeled facial landmarks provided in the Webcaricature dataset and show visual comparisons in Fig. 7. Since the dataset used in Cao et al. (2018) is not released, we can only use the WebCaricature dataset where each image is annotated with 17 landmarks.

For the first method, we compute the warping parameters between the photo and the caricature using the thin-plate-

spline transform (Jaderberg et al. 2015) by aligning the landmark positions. We denote this baseline method as “Landmark Positions”. Although landmarks are aligned, there are obvious distortions due to the lack of control points. We also experiment with sparse points for learning the shape transformation in our framework. However, the model using sparse points (e.g., 17 landmarks) does not generate reasonable results as it does not contain sufficient information to wrap every pixel in the facial region.

Second, to further increase the control points, we connect 17 sparse landmarks that belong to the same semantic (e.g., both eyes) into a single polygon. We then apply a one-hot encoding on the polygons that contain different facial components, in which this input is used for caricature generation as in the proposed model. In Fig. 7, although the results of “Landmark Maps” are visually more pleasing than the one using landmark positions, the facial contour does not transfer from caricatures to the outputs, since we have no control over the regions that are not covered by the landmark polygons. In addition, we conduct a user study: among 15 participants, 89% of the total 310 votes prefer our results than the ones using landmarks. Compared to the two baseline methods, the proposed algorithm uses face parsing maps to further control the shape transform with semantics and thus produce better results.

Although one can annotate more landmark points and apply our method for shape transformation, it is under a different problem context from our setting where only sparse landmarks are available. We regard this as an advantage of our method since we can leverage fewer landmark annotations and synthesize visually pleasing caricatures, which is of great interest when we do not have denser landmark annotations. Moreover, there are several merits for using pixel-wise semantic parsing maps. It enables dense correspondence matching between photos and caricatures, where every pixel

Fig. 7 Visual comparisons of shape transformation generated by different inputs. Our method using face parsing maps is able to accurately transfer the shape exaggerations from the shape reference, while preserving the facial structure

Fig. 8 Visual results on different inputs

correspondence is predicted by the non-linear deep neural network to guide the shape transformation process. As a result, our method can provide more accurate mapping for facial details, e.g., shapes of eyebrows, noses, and face con-

tours. In addition, the dense map is not only used as input, but also for the designed loss functions which encourage preserving the facial structure in the original photo.

Fig. 9 Visual comparisons with different image translation methods

Table 4 Ablation study on different inputs

Loss functions	mIoU	pixAcc
Image	22.31	90.23
Parsing maps	61.66	95.60
Image and parsing maps	61.94	95.78

Comparisons to the Image Input Taking the image information into consideration, we use different inputs to perform the ablation study and show the quantitative results in Table 4 and visual results in Fig. 8. The results demonstrate that only using image information does not generate sufficient shape transformations because there exists large texture and shape discrepancy between the photo and caricature images. The performance of the method by concatenating the image and parsing maps is almost identical to that by the scheme only using parsing maps. For computational efficiency, we only utilize the parsing map as the input in the following experiments.

Shape-based Methods We present comparisons with a non-rigid registration method (Jian and Vemuri 2010) and the recently developed Neural Best-Buddies (Aberman et al. 2018). Similar to Table 2, we evaluate the quality of the transformed parsing map. The mIoU and pixACC are (46.5%, 84.7%), (45.6%, 61.7%) for Jian and Vemuri (2010) and Aberman et al. (2018) respectively, which are worse than ours as (61.7%, 95.6%). One possible reason is that these methods cannot leverage semantic labels to guide registration, while ours is a learning-based model that considers semantics.

5.2 Comparisons to the State of the Arts

We conduct experiments with comparisons to the state-of-the-art methods by showing visual comparisons and performing user studies.

Image Translation Methods We evaluate our method with the state-of-the-art image translation algorithms, including CycleGAN (Zhu et al. 2017a), Neural Style Transfer (Gatys et al. 2016), MUNIT (Huang et al. 2018), and DRIT (Lee

Fig. 10 Visual comparisons with different caricature generation methods

et al. 2018). For Neural Style Transfer, we randomly select 5 caricatures as style images.

As shown in Fig. 9, conventional image translation methods are able to generate different textures in the generated outputs. However, there are severe artifacts due to the large texture variations in caricatures. More importantly, these methods only render slight shape changes, which is unsatisfactory for caricature generation.

Caricature Generation Methods Next, we perform the evaluation on existing caricature generation methods, including WarpGAN (Shi et al. 2019), CariGAN (Cao et al. 2018), Zheng et al. (2019) and Deep Image Analogy (Liao et al. 2017). We use the images provided by CariGAN as the test set for fair comparisons, where their output results are provided by the authors. In Fig. 10, while the other approaches

show improved results compared to image translation baselines, the exaggerated shape effects may not be natural. Note that WarpGAN does not need additional inputs, but it is less effective in aligning larger shape deformation. Compared to the PCA representation in Cao et al. (2018) that constrains the compact facial geometry and is similar to the idea of blend shapes (Parke 1972; Guenter et al. 1998), we use semantic maps to provide dense transformation parameters and capture the fine-grained shape details. With the dense semantics, we propose the coordinate-based loss and cycle consistency loss to stabilize the transformation process and reduce the artifacts inside each facial component.

Furthermore, we present the FID scores (Heusel et al. 2017) for different methods in Table 5. Our method achieves lower score (better realism) than other algorithms, but we note that this metric may not be effective for evaluation of

Table 5 FID scores on comparing our method with different caricature generation approaches

Method	CycleGAN	Style	MUNIT	Analogy	Zheng et al.	CariGANs	WarpGAN	Ours
Score ↓	111.31	118.15	105.13	102.87	134.93	102.36	101.13	100.98

Table 6 User study on image translation methods, hand-drawn caricatures, and our method

Method	CycleGAN	Style	MUNIT	DRIT	Hand-drawn	Ours
Score ↑	0.54	1.02	−0.82	0.75	2.32	2.02

Table 7 User study on comparing our method with different caricature generation approaches

Method	Analogy	Zheng et al.	CariGANs	WarpGAN	Ours
Score ↑	0.13	−1.71	1.19	0.78	2.06

Fig. 11 Visual comparisons of deformed photos guided by random selected caricatures and the recommended ones. Given a photo, our caricature retrieval model is able to automatically select plausible reference caricatures and obtain diverse results

caricatures since the inception model used for FID is trained on natural images, which are significantly different from caricatures. Also, the FID score mainly focuses on the image textures instead of the shapes of facial components, which might not be the best metric to indicate the quality of generated caricature shapes.

User Study We conduct user studies to evaluate the quality of generated caricatures in the above-mentioned methods. For each subject, we first show a normal face with a few hand-drawn caricatures as the instruction to guide the users. During the study, we randomly choose two of the methods among all the methods, and present one generated caricature for each method. We then ask each subject to select the one that “looks more like a caricature” in terms of the shape exaggeration, artistic style, and image quality. To compare with image translation methods, we collect 2650 pairwise results from a total of 38 participants, while for caricature gener-

ation algorithms, we collect 2220 pairwise outcomes from 31 participants. In Tables 6 and 7, we show the normalized scores computed by the Bradley–Terry model (Bradley and Terry 1952). The results show that the proposed algorithm performs favorably against state-of-the-art methods. In addition, compared with hand-drawn caricatures provided from the WebCaricature dataset (Huo et al. 2018) in Table 6, our score is closest to hand-drawn ones.

5.3 Additional Results and Analysis

Caricature Retrieval Given a photo, our caricature retrieval model is able to automatically select plausible reference caricatures and obtain diverse results (Fig. 1). To demonstrate the effectiveness of this recommendation, we also select reference caricatures by random selection and compare the generated results through a user study. Specifically, we show

Table 8 An example of distances for retrieval models with different margins

Sample	Same-1	Same-2	Same-3	Different-1	Different-2	Different-3
Margin = 0	0.185	0.131	0.134	0.114	0.133	0.110
Margin = 2	0.138	0.173	0.214	0.277	2.504	2.785
Margin = 10	3.180	3.133	3.334	1.497	4.192	3.570
Margin = 100	11.724	11.473	11.726	11.847	11.727	11.471

Fig. 12 An example for caricature retrieval ablation study. Same-1, Same-2 and Same-3 share the same identity with the input photo. The identities of Different-1, Different-2 and Different-3 are different from the identity of the input photo **Fig. 13** Visualization of the encoding space for caricature shapes. We show that each group (denoted in different colors) has certain shapes in facial components, while neighbors in this feature space also share similar shapes

two sets of deformed photos and allow users to select the preferred set based on the visual quality and diversity of generated caricatures. Among 900 pairwise outcomes from 30 participants, 83% of the votes prefer our results. Here, we show more results of deformed photos guided by random selected caricatures and the recommended ones in Fig. 11. For simplicity, we only present results of deformed photos before applying style transfer.

One key factor in training the caricature retrieval model is utilizing the positive and negative pairs. For negative pairs, our method may sample a similar shape that belongs to different identities. However, the overall gap between positive and negative pairs is still large. To understand the learning behavior, we adjust the hyperparameter margin for negative pairs in (1). In Table 8, the margin is set to {0, 2, 10, 100} for training the retrieval models, and we report the distances of samples to the anchor (i.e., we pre-sample 3 positives and

3 negatives in this case) associated with the parsing maps in Fig. 12 (e.g., “Same-1” refers to the first positive sample, while “Different-2” is the second negative sample).

When the margin is very small, the model mainly learns to enforce the photo and caricature belonging to the same identity being close to each other in the embedding space. The effect of negative pairs is decreased. The distance between the anchor and the same identity (positive) is very close to that between the anchor and the different identity (negative). When the margin is set to 2 or 10, the gap between positive pairs and negative pairs are mostly satisfactory (i.e., we use 2 in the experiments). When the margin is very large, the effect of negative pairs is increased. The distance of positive and negative pairs are both very large, which is not distinguishable. In addition, from Fig. 12, parsing maps of positive samples (Same-1, Same-2, Same-3) have large nose and similar pose to the anchor one (Photo), which are quite different from the ones of different identities.

Shape Embedding Space We analyze the latent shape embedding space extracted from the encoder. We randomly select 1000 caricatures and extract their shape embedding vectors. To verify whether the shape embedding features could capture meaningful facial structure information, we first apply the mean shift clustering method (Comaniciu and Meer 2002) to group caricature shapes and then apply the t-SNE (van der Maaten and Hinton 2008) scheme for visualization. Figure 13 shows that caricatures within the same cluster share a similar facial structure, while neighboring clusters are also similar to each other in certain semantic parts.

We note that the embedding space shows smooth transitions between different shape deformation clusters, which can be exploited to generate different caricatures through a simple interpolation between caricature references as shown in Fig. 14. As a result, the degree of shape exaggeration could be controlled through interpolating between the input facial

Fig. 14 Shape interpolations between the input image and caricature in our learned shape encoding space.

Fig. 15 Results of manipulating on the parsing map. Here, for simplicity, we only present results of deformed photos before applying style transfer

shape and the reference one, whenever users prefer to preserve more identity of the input portrait.

User Control Given a caricature/photo pair, our approach is not only as simple as a one-click shape transformation, but is also flexible to accommodate fine-grained facial structure refinements from users via providing grid controls on the parsing map. We show an example for results controlled by the grid in Fig. 15. Through adjusting the positions of control points, we are able to change the face contour, move the facial component positions, or adjust the shape of facial components. Therefore, users could manipulate the desired shape easily and create diverse caricatures.

In addition, these results demonstrate that it is plausible to feed new caricature parsing maps into our shape transforma-

tion model, e.g., the parsing map could be flexibly modified by the users as shown in Fig. 15.

Identity Preservation In addition, we conduct a user study similar to CariGAN (Cao et al. 2018) to evaluate the degree of identity preservation. The users are asked to choose the correct subject from 5 portraits, given the results generated by each method. We collect 650 votes with 26 participants and the accuracy are 70% (our method), 53% CariGAN (Cao et al. 2018), 53% (Zheng et al. 2019), 68% WarpGAN (Shi et al. 2019), and 45% Deep Image Analogy (Liao et al. 2017). The results show that our model is the best one to preserve the identity.

Limitation Although using the proposed shape transformation for caricature generation is simple and effective, we find that it may render unsatisfactory results for some scenarios. For example, when the eyebrows and eyes are very close to each other in the photo, there would be some artifacts around the eyes when the transformer tries to change the eyebrow shapes as shown in Fig. 16. In the future work, we plan to use multiple shape sub-transformers for each facial component, followed by a global refinement network to integrate different sub-transformers together.

Runtime Performance In the proposed framework, we use a single Nvidia 1080Ti GPU and the runtime on a 256×256 input photo is around 0.65 s, including 0.1 s for face parsing, 0.1 s for caricature retrieval, 0.3 s for shape transformation, and 0.15 s for style transfer.

Fig. 16 Limitation of the proposed method

6 Conclusions

In this paper, we propose a semantic dense shape transformation algorithm for learning to caricature. Specifically, we utilize a face parsing map to densely predict warping parameters, such that the shape exaggerations are effectively transferred while the facial structure is still maintained. Visual comparisons and user studies demonstrate that the proposed algorithm is able to generate high-quality caricatures against state-of-the-art methods. In addition, we show that the learned embedding space of semantic parsing map allows us to directly manipulate the parsing map and generate shape changes according to the user preference.

Acknowledgements This work was supported in part by The National Key Research and Development Program of China (Grant Nos: 2018AAA0101400), in part by The National Nature Science Foundation of China (Grant Nos: 62036009, 61936006), in part by Innovation Capability Support Program of Shaanxi (Program No. 2021TD-05). W.-C. Hung, Y.-T. Chang, Y. Li, and M.-H. Yang were supported in part by National Science Foundation CAREER Grant 1149783.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

Aberman, K., Liao, J., Shi, M., Lischinski, D., Chen, B., & Cohen-Or, D. (2018). Neural best-buddies: Sparse cross-domain correspondence. *ACM TOG*.

Akleman, E. (1997). Making caricatures with morphing. In *ACM SIG-GRAPH* (p. 145).

Akleman, E., Palmer, J., & Logan, R. (2000). Making extreme caricatures with a new interactive 2d deformation technique with simplicial complexes. In *Proceedings of visual*.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *ICML*.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39, 324–345.

Brennan, S. E. (2007). Caricature generator: The dynamic exaggeration of faces by computer. *Leonardo*, 40(4), 392–400.

Cao, K., Liao, J., & Yuan, L. (2018). Carigans: Unpaired photo-to-caricature translation. *ACM TOG*.

Chang, H., Lu, J., Yu, F., & Finkelstein, A. (2018). Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *CVPR*.

Chen, D., Yuan, L., Liao, J., Yu, N., & Hua, G. (2017). Stylebank: An explicit representation for neural image style transfer. In *CVPR*.

Chen, K., Zheng, J., Cai, J., & Zhang, J. (2020). Modeling caricature expressions by 3d blendshape and dynamic texture. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 3228–3236).

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity discriminatively, with application to face verification. In *CVPR*.

Chu, W., Hung, W.-C., Tsai, Y.-H., Cai, D., & Yang, M.-H. (2019). Weakly-supervised caricature face parsing through domain adaptation. In *ICIP*.

Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *PAMI*, 24(5), 603–619.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In *ICCV*.

Ganin, Y., Kononenko, D., Sungatullina, D., & Lempitsky, V. (2016). Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *ECCV*.

Gatys, L., Ecker, A. S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. In *NIPS*.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *CVPR*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NIPS*.

Guenther, B., Grimm, C., Wood, D., Malvar, H., & Pighin, F. (1998). Making faces. In *Proceedings of the 25th annual conference on computer graphics and interactive techniques* (pp. 55–66).

Han, X., Hou, K., Du, D., Qiu, Y., Cui, S., Zhou, K., & Yu, Y. (2018). Caricatureshop: Personalized and photorealistic caricature sketching. *IEEE Transactions on Visualization and Computer Graphics*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS* (pp. 6626–6637).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *CVPR*.

Huang, X., & Belongie, S. J. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In *ECCV*.

Huo, J., Li, W., Shi, Y., Gao, Y., & Yin, H. (2018). Webcaricature: A benchmark for caricature recognition. In *BMVC*.

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *CVPR*.

Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. In *NIPS*.

Jian, B., & Vemuri, B. C. (2010). Robust point set registration using gaussian mixture models. *PAMI*.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *ECCV*.

Kim, T., Cha, M., Kim, H., Lee, J. K., & Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. In *ICML*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR*.

- Lee, H. Y., Tseng, H. Y., Huang, J. B., Singh, M. K., & Yang, M. H. (2018). Diverse image-to-image translation via disentangled representations. In *ECCV*.
- Li, W., Xiong, W., Liao, H., Huo, J., Gao, Y., & Luo, J. (2020). Carigan: Caricature generation through weakly paired adversarial learning. *Neural Networks*, 132, 66–74.
- Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M. H. (2017a). Diversified texture synthesis with feed-forward networks. In *CVPR*.
- Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M. H. (2017b). Universal style transfer via feature transforms. In *NIPS*.
- Li, Y., Wang, N., Liu, J., & Hou, X. (2017c). Demystifying neural style transfer. In *IJCAI*.
- Li, Y., Liu, M. Y., Li, X., Yang, M. H., & Kautz, J. (2018). A closed-form solution to photorealistic image stylization. In *ECCV*.
- Liao, J., Yao, Y., Yuan, L., Hua, G., & Kang, S. B. (2017). Visual attribute transfer through deep image analogy. *ACM TOG*.
- Liao, P. Y., Chiang, W. H., & Li, T. Y. (2004). Automatic caricature generation by analyzing facial features. In *ACCV*.
- Lin, C. H., Yumer, E., Wang, O., Shechtman, E., & Lucey, S. (2018). Stgan: Spatial transformer generative adversarial networks for image compositing. In *CVPR*.
- Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. In *NIPS*.
- Luo, W. C., Liu, P. C., & Ouhyoung, M. (2002). Exaggeration of facial features in caricaturing. In *Proc. int. computer symposium*.
- Park, E., Yang, J., Yumer, E., Ceylan, D., & Berg, A. C. (2017). Transformation-grounded image generation network for novel 3d view synthesis. In *CVPR*.
- Parke, F. I. (1972). Computer generated animation of faces. In *Proceedings of the ACM annual conference* (Vol. 1, pp. 451–457).
- Shi, Y., Deb, D., & Jain, A. K. (2019). Warpgan: Automatic caricature generation. In *CVPR*.
- Shu, Z., Sahasrabudhe, M., Alp Guler, R., Samaras, D., Paragios, N., & Kokkinos, I. (2018). Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *ECCV*.
- van der Maaten, L. J. P., & Hinton, G. E. (2008). Visualizing high-dimensional data using t-SNE. *JMLR*, 9, 2579–2605.
- Wang, X., Shrivastava, A., & Gupta, A. (2017). A-fast-rcnn: Hard positive generation via adversary for object detection. In *CVPR*.
- Wu, Q., Zhang, J., Lai, Y. K., Zheng, J., & Cai, J. (2018). Alive caricature from 2d to 3d. In *CVPR* (pp. 7336–7345).
- Wu, W., Kan, M., Liu, X., Yang, Y., Shan, S., & Chen, X. (2017). Recursive spatial transformer (rest) for alignment-free face recognition. In *CVPR*.
- Zheng, Z., Wang, C., Yu, Z., Wang, N., Zheng, H., & Zheng, B. (2019). Unpaired photo-to-caricature translation on faces in the wild. *Neurocomputing*, 355, 71–81.
- Zhou, E., Cao, Z., & Sun, J. (2018). Gridface: Face rectification via learning local homography transformations. In *ECCV*.
- Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.
- Zhu, J. Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., & Shechtman, E. (2017b). Toward multimodal image-to-image translation. In *NIPS*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.