G2L-CariGAN: Caricature Generation from Global Structure to Local Features

Xin Huang[1], Yunfeng Bai[1], Dong Liang[1], Feng Tian[2], Jinyuan Jia[1]*

[1]Tongji University
[2]Duke Kunshan University
{huangxin0124, 2131480, sse_liangdong, jyjia}@tongji.edu.cn, feng.tian978@dukekunshan.edu.cn

Abstract

Existing GAN-based approaches to caricature generation mainly focus on exaggerating a character's global facial structure. This often leads to the failure in highlighting significant facial features such as big eyes and hook nose. To address this limitation, we propose a new approach termed as G2L-CariGAN, which uses feature maps of spatial dimensions instead of latent codes for geometric exaggeration. G2L-CariGAN first exaggerates the global facial structure of the character on a low-dimensional feature map and then exaggerates its local facial features on a high-dimensional feature map. Moreover, we develop a caricature identity loss function based on feature maps, which well retains the character's identity after exaggeration. Our experiments have demonstrated that G2L-CariGAN outperforms the state-of-arts in terms of the quality of exaggerating a character and retaining its identity.

Introduction

Drawing caricatures involves exaggerating distinctive facial features that convey comedy and sarcasm. The early approaches to automatic caricature generation are mainly based on graphics and image-to-image generation. Graphic-based methods (Akleman 1997a; Akleman, Palmer, and Logan 2000) pay attention to exaggeration, while image-to-image conversion methods (Huang et al. 2018; Zhu et al. 2017) mainly focus on color transfer. With the advancement of artificial neural networks, several methods based on generative adversarial networks (GAN) have been proposed (Cao, Liao, and Yuan 2018; Shi, Deb, and Jain 2019; Chu et al. 2021; Gu et al. 2021), combing both exaggeration and color transfer. However, the quality of caricatures generated by these methods is generally not high, as exaggeration through 2D image warping tends to cause distortion.

With the development of generative AI, a recent work StyleGAN (Karras, Laine, and Aila 2019) was developed to produce high-quality realistic portraits. A few methods based on StyleGAN have been proposed to generate caricatures without abnormal distortion by exaggerating in StyleGAN's latent space. StyleCariGAN (Jang et al. 2021) has made fine adjustments to StyleGAN, but can achieve only one style of exaggeration at a time. Subsequently, DualStyleGAN (Yang et al. 2022) proposed an example-based method, which can generate high-quality caricatures and achieve various exaggeration styles based on reference caricatures. However, both StyleCariGAN and DualStyleGAN exaggerate geometry shapes in vector-based latent space that makes it difficult for an encoder to compress an image's local semantics in a disentangled way. As a result, some prominent local facial features, such as a hook nose or wide mouth, cannot be exaggerated properly (Fig. 1 (a)). The other problem is that the caricatures these methods generated fail to preserve the identity of the character (Fig. 1 (b)).

In order to address these limitations, in this paper, we put forward G2L-CariGAN, a new caricature generation approach based on reference caricatures to exaggerate a character's global and local features. G2L-CariGAN consists of a global exaggeration module and three local exaggeration modules, which exaggerate the (global) facial structure (shape) and (local) facial features, respectively. G2L-CariGAN uses feature maps instead of vector-based latent codes as the latent representation of face's shape because (a) feature maps retain the spatial information of the photo, which helps to maintain the facial details; (b) feature maps can easily enhance distinguishing features for local exaggeration; (c) feature maps make it easy to exaggerate using multiple references on different regions with spatial masks. As a result, feature maps with different spatial dimensions can control the geometric features such as a round face or squared face, from coarse to fine. As for color style, G2L-CariGAN uses the vector-based latent code for representation. The reason is that if feature maps are used to extract the color style, the spatial distribution of color in the reference caricature will be *completely* reflected in the generated caricatures, which often results in undesirable effects. Moreover, we develop a new caricature identity loss based on the principle of "exaggerating the difference from the mean" (EDFM) (Brennan 1985). The loss has well balanced the exaggeration of a character's local features and the preservation of its identity in the generated caricature.

The main contributions of the paper include:

- We propose a new GAN-based neural network G2L-CariGAN for extracting both a character's overall facial structure and its facial features, enabling exaggerations with multiple references.

*Corresponding author

- We propose a new latent space, Global-Local-Style (GLS), to encode spatial information, which is more capable of exaggerating local facial features.
- We propose a new identity loss function based on feature maps. It shows a higher recognition rate on the photo-to-caricature dataset and exaggerates a character's prominent features more obviously.

Related Work

In this section, we review some related works, focusing on traditional caricature generation and deep caricature generation.

Traditional Caricature Generation

Traditional caricature creation methods relied heavily on digital image processing and computer graphics. They could be divided into three categories: interactive methods, rule-based methods, and instance-based methods. Interactive methods (Akleman 1997a; Akleman, Palmer, and Logan 2000) allowed for interactive exaggeration of photos but typically required artists with extensive expertise and experience. The majority of rule-based methods (Brennan 1985; Lai, Chung, and Edirisinghe 2006; Chen et al. 2009; Mo, Lewis, and Neumann 2004) obeyed the principle of "exaggerating the difference from the mean" (Brennan 1985). Automatically, instance-based techniques (Liang et al. 2002; Liu, Chen, and Gao 2006; Liu et al. 2009) extracted matching criteria from facial expression databases. Without taking into account the color differences between caricatures and photos, they all placed more emphasis on exaggerating facial contours.

Deep Caricature Generation

In recent years, deep neural networks have made great progress in image-to-image translation (Huang et al. 2018)(Zhu et al. 2017)(Liu, Breuel, and Kautz 2017). CariGANs (Cao, Liao, and Yuan 2018) achieved unpaired photo-to-caricature translation by learning both geometric exaggeration and appearance stylization respectively with two sub-networks. In a combined learning framework, WarpGAN (Shi, Deb, and Jain 2019) covered both style transfer and facial deformation. However, these techniques generated only one geometry exaggeration style. CariME (Gu et al. 2021) created a network for learning multiple styles from caricatures to address this issue. Chu et al. (Chu et al. 2021) used facial segmentation maps to learn the deformed style. Nevertheless, these methods do not support users' editing. CariPainter (Huang et al. 2022) applied sketch and segmentation map to perform diverse exaggeration. All these methods mentioned above based on deep learning produced limited image quality because distortion is often brought on by exaggeration in 2D images.

Through hierarchical style control, StyleGAN (Karras, Laine, and Aila 2019) created high-resolution face images that closely resemble genuine photos. Recent works based on StyleGAN have successfully embedded real photos into the latent space of StyleGAN, and enabled semantic editing in the latent space of StyleGAN. Pinkney and Adler (Pinkney and Adler 2020) realized the conversion from real face to cartoon face by fine-tuning StyleGAN on the limited cartoon data. StyleCariGAN (Jang et al. 2021) solved the cross-domain problem from photos to caricatures by exchanging the layers of two StyleGANs and attaching the learnable shape exaggeration blocks to the coarse layers copied from the StyleGAN trained on the photo dataset. Only color styles can be migrated using the above techniques, while shape styles cannot be controlled. DualStyleGAN (Yang et al. 2022) added an external style path, which can be trained to control both color and shape style. However, this method's control on the geometry style was largely limited to the exaggeration of the overall facial structure, and could not exaggerate the facial features. In contrast, our method can not only learn the overall structural style of the reference caricature but also exaggerate the local features.

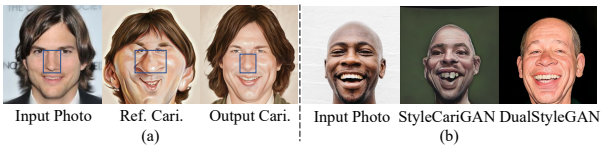

Figure 1: (a) A case which fails to enhance the character's characteristic features (big nose) (Yang et al. 2022). (b) A failure case where neithor StyleCariGAN nor DualStyleGAN preserves the identity of the input photo.

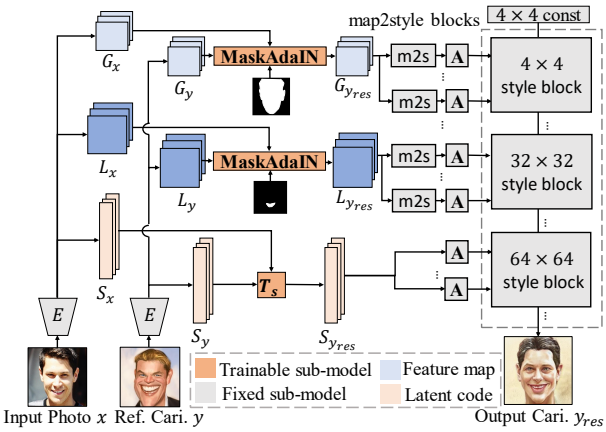

Figure 2: Outline of G2L-CariGAN. The input photo $\boldsymbol{x}$ and reference caricature $\boldsymbol{y}$ are fed into the pre-trained encoders (E) to obtain the global feature maps ($\boldsymbol{G}_x$, $\boldsymbol{G}_y$), local feature maps ($\boldsymbol{L}_x$, $\boldsymbol{L}_y$) and latent codes ($\boldsymbol{S}_x$, $\boldsymbol{S}_y$). Style transfer module (T_s) consists of 3 fully connected layers, which take latent codes as input to generate style latent codes ($\boldsymbol{S}_{y_{res}}$) for the output caricature ($\boldsymbol{y_{res}}$). The geometry exaggeration module uses MaskAdaIN, i.e., mask-based AdaIN (Adaptive Instance Normalization (Huang and Belongie 2017)) blocks, to exaggerate feature maps. $\boldsymbol{G}_x, \boldsymbol{G}_y, \boldsymbol{G}_{y_{res}} \in \mathbb{R}^{512\times16\times16}$ controls the exaggeration of global face shape, while $\boldsymbol{L}_x, \boldsymbol{L}_y, \boldsymbol{L}_{y_{res}} \in \mathbb{R}^{512\times32\times32}$ controls the exaggeration of local facial features.

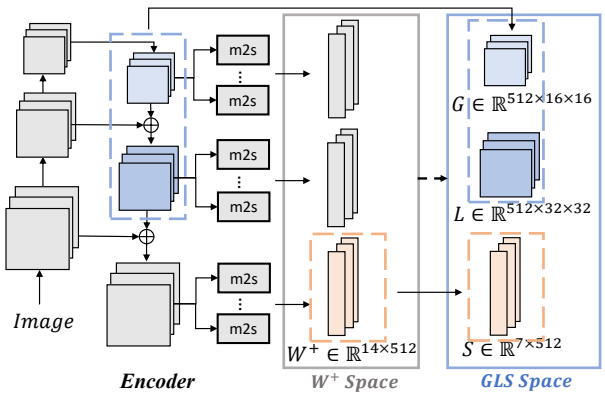

Figure 3: The relation between our Global-Local-Style (GLS) space and W^+ space. The encoder (E in the pipeline in Fig. 2) first extracts feature maps from an input image using a standard feature pyramid (Richardson et al. 2021) over a ResNet backbone. The first k ($k = 7$) blocks of the $W+$ code are then replaced by the feature maps (blue dashed boxes) including a global feature map (G) and a local feature map (L), while the remaining W^+ codes (orange dashed boxes) are utilized as a style code (S).

G2L-CariGAN

Fig. 2 shows the overall structure of G2L-CariGAN. It consists of a pre-trained StyleGAN generator on the webcaricature dataset (Huo et al. 2018), a style transfer module T_s, and a geometric exaggeration module MaskAdaIN.

GLS Space

Before exaggeration, we need to map images (including input photos and reference caricatures) to an appropriate latent space. Unlike StyleCariGAN (Jang et al. 2021) and DualStyleGAN (Yang et al. 2022) where the exaggeration takes place in W^+ space, we propose Global-Local-Style (GLS) space to address the issues attributed to W^+ space such as the compromised identity of the character in the input image. Fig. 3 shows the connection between the GLS space and the W^+ space. We adopt the structure of the psp encoder (Richardson et al. 2021), which first extracts feature maps of coarse ($512\times16\times16$), middle ($512\times32\times32$), and fine ($512\times64\times64$) dimensions. Then, the encoder maps the three feature maps to the low, middle, and high layers of the W^+ space through map2style (m2s in Fig. 3) modules. In the W^+ space, the low layer controls face shapes, the middle layer controls facial features, and the high layer controls color styles. Among them, face shapes and facial features contain spatial information, which is more suitable for using feature maps for exaggeration while the color styles are not affected by shapes so are more suitable for using latent codes for style transfer. Therefore, our GLS space consists of three parts: $\boldsymbol{G} \in \mathbb{R}^{512\times16\times16}$, which controls face shape exaggeration, $\boldsymbol{L} \in \mathbb{R}^{512\times32\times32}$, which controls facial feature exaggeration, and $\boldsymbol{S} \in \mathbb{R}^{7\times512}$, which controls color styles.

Style Transfer Module

The caricature style code $\boldsymbol{S}_y$ (obtained from the reference caricature or just a random style) undergoes a style transfer module T_s to obtain style deviation $\boldsymbol{\delta}_S$. The photo $\boldsymbol{x}$ undergoes a pre-trained encoder to obtain the photo style code $\boldsymbol{S}_x$. The two are merged through a weight w_s to get $\boldsymbol{S}_{res}$ ($\boldsymbol{S}_{res} = (1-w_s)\boldsymbol{S}_x + w_s\boldsymbol{\delta}_S$) and input into an affine transformation A. T_s is composed of several trainable fully connected layers.

We define a content loss between the generated caricature $\boldsymbol{y}_{res}$ and $\boldsymbol{x}$ to ensure that $\boldsymbol{y}_{res}$ still retains the facial region of the photo:

$$L_{con} = \|\phi_{5\_3}(\boldsymbol{y}_{res}) - \phi_{5\_3}(\boldsymbol{x})\|_2 \tag{1}$$

where ϕ_{5_3} represents $relu5\_3$ feature map in VGG-19 (Simonyan and Zisserman 2015) pre-trained on ImageNet dataset.

We define a style loss between $\boldsymbol{y}_{res}$ and $\boldsymbol{y}$ to ensure that T_s can extract caricatures' styles:

$$L_{sty} = \|Gram(\phi_i(\boldsymbol{y}_{res})) - Gram(\phi_i(\boldsymbol{y}))\|_2 \tag{2}$$

where each ϕ_i represents a layer in VGG-19 (Simonyan and Zisserman 2015). We use $relu1\_1$, $relu2\_1$, $relu3\_1$, $relu4\_1$, $relu5\_1$ layers with equal weights in our experiments.

Finally, the whole loss for optimizing L_{style} is:

$$L_{style} = \lambda_{con}L_{con} + \lambda_{sty}L_{sty} \tag{3}$$

where the parameters λ_{con} and λ_{sty} balance multiple objectives.

Geometry Exaggeration Module

The feature map $\boldsymbol{G} \in \mathbb{R}^{512\times16\times16}$ and the feature map $\boldsymbol{L} \in \mathbb{R}^{512\times32\times32}$ are used for exaggerating the facial structure and facial features, respectively. We denote $r \in \{face, eyes, nose, mouth\}$ as the facial region. Taking facial structure exaggeration ($r = face$) as an example, the MaskAdaIN module is shown in Fig.4. The caricature feature map $\boldsymbol{G}_y$ is processed by the γ_r module and the β_r module to obtain the exaggeration coefficient map $\boldsymbol{G}_{coff}$ and the bias map $\boldsymbol{G}_{bias}$ to get $\boldsymbol{G}_x^{'}$:

$$\boldsymbol{G}_x^{'} = \boldsymbol{G}_{coff} \otimes norm(\boldsymbol{G}_x) + \boldsymbol{G}_{bias} \tag{4}$$

Where $norm(.)$ represents normalization.

The exaggerated $\boldsymbol{G}_{res}$ is an alpha blending of $\boldsymbol{G}_x$ and $\boldsymbol{G}_x^{'}$:

$$\boldsymbol{G}_{res} = w_g\boldsymbol{m}_r \otimes \boldsymbol{G}_x^{'} \oplus (1-\boldsymbol{m}_r) \otimes \boldsymbol{G}_x \tag{5}$$

Where $r \in \{face, eye, nose, mouth\}$. $\boldsymbol{m}_r$ is the mask of the region r. w_g controls the degree of exaggeration for the global facial structure. Finally, we obtain an exaggerated feature map $\boldsymbol{G}_{y_{res}} \in \mathbb{R}^{512\times16\times16}$. We obtain exaggerated feature maps for facial features $\boldsymbol{L}^r_{y_{res}} \in \mathbb{R}^{512\times32\times32}$ in the same way.

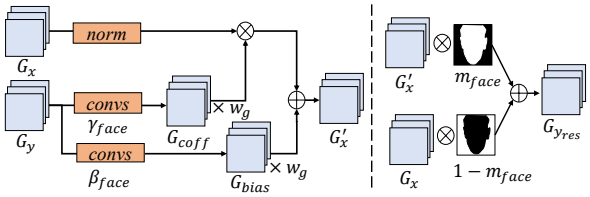

Figure 4: Mask-based AdaIN block (MaskAdaIN).

Method	Photo-to-Photo(%) ↑	Photo-to-Cari(%) ↑
ArcFace	99.46	70.92
Ours	n/a	**85.73**

Table 1: The accuracy of facial recognition from ArcFace (Deng et al. 2019) and ours (using the loss defined in Eq. 8) on the photo-to-photo dataset and the photo-to-caricature dataset (Huo et al. 2018) which contains photos and hand-drawn caricatures of 252 identities in total.

We take the L_2 norm of feature maps as a reconstruction loss.

$$L^r_{rec} = \|\boldsymbol{L}^r_{y_{res}} - \boldsymbol{L}^r_y\|_2 + \|\boldsymbol{G}_{y_{res}} - \boldsymbol{G}_y\|_2 \tag{6}$$

We also define a masked version of LPIPS loss (Zhu et al. 2021) between $\boldsymbol{y}^r_{res}$ and $\boldsymbol{y}$ to ensure that $\boldsymbol{y}^r_{res}$ still retains the facial feature of the photo:

$$L^r_c = L_{maskedPIPS}(\boldsymbol{y}^r_{res}, \boldsymbol{y}, \boldsymbol{m}_r) \tag{7}$$

where $\boldsymbol{y}^r_{res}$ is the caricature generated with region r exaggerated.

Unlike other face-generation tasks, caricature-generation tasks produce images with significant differences in geometric shapes compared to real human faces. Currently, most caricature generation methods (Gu et al. 2021), (Shi, Deb, and Jain 2019) use facial recognition networks to calculate identity loss. Since facial recognition networks are pre-trained on a large number of real human face images, and their performance in photo-to-caricature decreases, as shown in Tab. 1. Obviously, using an identity loss function based on real face recognition methods is not effective enough for caricature generation.

Hence, we aim to develop a new identity loss function base on "Exaggerating the Difference From the Mean" (EDFM) (Brennan 1985), which is one of the fundamental rules of drawing a caricature. It emphasizes those features that make a person unique, i.e., different from the average face. In other words, as long as the prominent features of the character are exaggerated, the identity characteristics of the character can be highlighted. In the past, researchers (Brennan 1985; Akleman 1997b) often used the difference between feature points and average facial feature points to implement this theory, which is too sparse and easily affected by rotation angles in 2D images. Since our feature map is suitable for extracting prominent features, we define the character feature as the difference between the photo feature map and the average photo feature map, and the caricature feature as the difference between the caricature feature map and the average caricature feature map. We used 5974 caricatures and 6042 photos from the WebCaricature dataset(Huo et al. 2018) to train the psp encoder (Richardson et al. 2021) and then generate the feature maps of the same size for each photo and caricature. Then we obtain the average feature maps of photos and caricatures, respectively. The cosine similarity between the photo feature and the caricature feature is then used as the identity loss to emphasize the features that make the photo's subject unique:

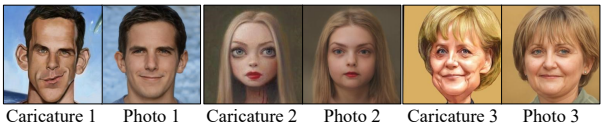

Figure 5: Examples of our training dataset.

$$L_{cariidt} = 1 - cos\_sim((\boldsymbol{G}_x - \boldsymbol{G}_{x_{avg}}), (\boldsymbol{G}_y - \boldsymbol{G}_{y_{avg}})) \tag{8}$$

Where $cos\_sim(.)$ represents cosine similarity. $\boldsymbol{G}_{x_{avg}}$ and $\boldsymbol{G}_{y_{avg}}$ represent the average feature maps of the photo and the caricature, respectively. Cosine similarity can make their directions to be as similar as possible but not approaching completely the same if using Euclidean distance which reduces exaggeration. As a result, exaggeration and identity preservation are well balanced. We calculate our method's accuracy on the same photo-to-caricature dataset (Huo et al. 2018) which contains photos and caricatures of 252 identities as Arcface(Deng et al. 2019) and the result is shown in Tab. 1.

Finally, the whole loss for optimizing L_{geo} is:

$$L^r_{geo} = \lambda_{rec} L^r_{rec} + \lambda_c L^r_c + \lambda_{cariid} L^r_{cariid} \tag{9}$$

Where λ_{rec}, λ_c and λ_{cariid} are chosen to balance multiple objectives.

Experiments

In this section, we compare our G2L-CariGAN to state-of-the-art methods and evaluate its performance. In Eq. 3, we set $\lambda_{con} = 0.01$ and $\lambda_{sty} = 20$. In Eq. 9, we set $\lambda_{rec} = 1$, $\lambda_c = 1$, $\lambda_{cariid} = 0.01$. w_r can be set to different values to control the degree of exaggeration. We use an RTX 3060 GPU for all experiments. The learning rate, number of epochs, and batch size are as 0.01, 2000, and 1, respectively.

Datasets and Training Procedure

To train the style transfer module T_s, we use Webcaricature (Huo et al. 2018), which is a large unpaired photo-caricature dataset consisting of 6042 caricatures and 5974 photos from 252 persons in total. To train the geometry exaggeration module, we use the destylization method in DualStyleGAN(Yang et al. 2022) to convert caricatures from Webcaricature dataset into corresponding photos to obtain paired photo-to-caricature dataset. Using a paired dataset can ensure local region alignment and learn the exaggeration of shapes. Examples of generated training pairs are shown

Method	FID ↓	Identity(%) ↑
CariGANs	n/a	95.29
WarpGAN	61.96	93.33
AutoToon	103.46	52.94
Semantic-CariGANs	75.22	78.82
CariME	53.64	95.29
StyleCariGAN	64.98	85.49
CariPainter	63.23	76.47
DualStyleGAN	96.30	82.35
G2L-CariGAN(ours)	**46.73**	**97.65**

Table 2: FID Score and identity evaluation results. A lower FID indicates higher image quality. (n/a: CariGANs does not provide open-source code)

in Fig.5. During the training process, we follow the principle of starting from the whole to the parts, similar to the way a painter draws by hand. We begin by training the T_s module. The global MaskAdaIN module is then trained for overall facial structure exaggeration, and each local MaskAdaIN module is trained independently.

Comparison to State-of-the-Art Methods

As shown in Fig. 6 and Fig. 7, we compare G2L-CariGAN with the state-of-arts, including those with reference caricatures and those without. All implementations of the methods are based on their default settings, except for CariGANs (Cao, Liao, and Yuan 2018), which has no published code, so we use the result published on their project website[1]. AutoToon and Semantic-CariGANs do not change the style (so we add AdaIN (Huang and Belongie 2017) on them for style transfer). Among them, only CariME, Semantic-CariGANs, and DualStyleGAN provide the shape exaggeration of reference caricatures. We will use the same reference caricatures to compare with these methods.

Comparison to Caricature Generation Methods without Reference The methods based on GAN, such as CariGANs (Cao, Liao, and Yuan 2018) and WarpGAN (Shi, Deb, and Jain 2019), have difficulty to generate caricatures with high-quality and they are prone to produce texture defects, which is obvious in the second row (output) of Fig. 6. Neither CariGANs nor WarpGAN can generate diversified deformation styles by referring to caricatures. AutoToon (Gong, Hold-Geoffroy, and Lu 2020) can produce only geometric deformation. StyleCariGAN (Jang et al. 2021) is capable of producing crisper caricatures, but its identity retention is insufficient, as shown in the second row (output) of Fig. 6. Our method not only has the advantages of StyleGAN in generating high-quality pictures but also creatively proposes exaggeration for local facial features, which produces exaggerated caricatures with more prominent photo features. For example, the first image of the last column in Fig. 6 is a high-quality caricature (generated by our method) with an exaggerated face and big eyes.

Comparison to Caricature Generation Methods with Reference We further compare our approach with caricature generation methods with the same reference caricature. As shown in Fig. 7, CariME (Gu et al. 2021) and Semantic-CariGAN (Chu et al. 2021) can refer to the reference caricatures to provide different exaggeration styles, but the quality of the generated caricatures is low. For example, in the second row, the subject's texture is blurry. CariME cannot generate color styles according to the reference caricature, so we provide random color styles. Caricatures generated by DualStyleGAN (Yang et al. 2022) can generate high-quality caricatures, while it focuses more on the overall deformation, ignoring the exaggeration of facial details. As shown in the sixth row of Fig. 7, DualStyleGAN's result ignores the prominent feature of the hook nose. Compared to these methods, G2L-CariGAN produces high-quality caricatures that exaggerate not only facial contour but also facial details. For example, in the first row of Fig. 7, our method can generate a high-quality caricature with a big nose that is similar to the reference.

[1]https://ai.stanford.edu/kaidicao/cari-gan/index.html

Quantitative Analysis

Through analysis of Frchet Inception Distance (FID) (Heusel et al. 2017), we quantitatively assessed the fidelity to caricature image distribution. Tab. 2 compares the FID of ours and other state-of-the-art methods. This demonstrates that caricatures generated by our approach are the most similar to the distribution of caricatures. The FID values are calculated between the generated caricatures and all caricatures in the WebCaricature dataset (Huo et al. 2018). The caricature dataset for calculating FID of all the methods are generated from the same photo dataset, which contains 1800 random selected photos from the CelebA (Liu et al. 2015) dataset that were not used in training. The reference caricatures for methods which needs reference are random selected from WebCaricature dataset (Huo et al. 2018).

Identity Evaluation

In this part, we quantify identity preservation accuracy. We use our caricature identity loss defined in 3.3 as the distance d between a photo and a caricature. We define a threshold K to distinguish whether the photo and the caricature belong to the same identity, i.e., they do if $d < K$ (K is set to 0.82 in this paper). Then we calculate the accuracy on the photo-to-caricature dataset with 294 pairs in total with each of the 9 methods. The 294 photos are from the test dataset of CariGANs on their website because their open-source codes are not available. Tab. 2 (right) shows the results for identity preservation. Obviously, G2L-CariGAN can retain better identification than other approaches.

Ablation Study

Loss Functions We train 2 variants of G2L-CariGAN for comparison by eliminating L_{cariid} in Eq. 8, to examine the effectiveness of our identity loss function. As shown in Tab. 3, L_{cariid} can greatly improve the identity retention effect of the generated image.

Figure 6: Comparison of G2L-CariGAN with other state-of-the-art caricature generation methods without reference caricatures.

Global and Local Feature Maps To validate the effectiveness of using feature maps as geometric features, we replace the global feature map and the local feature map with W+ latent codes, respectively. The results are shown in Fig. 8. The global feature map can exaggerate the face shape with the facial segmentation map. This ensures that other areas are not affected by the reference caricature. As shown in the example in the third column, using vector-based latent codes instead of global feature maps results in a hairstyle that does not match the photo. The local feature map can better maintain the local expressions of the photo and exaggerate them.

Local Exaggeration Modules We compare the caricature generations with and without the local feature map $\boldsymbol{L} \in \mathbb{R}^{512\times32\times32}$. Fig. 9 shows that the caricature with local exaggeration can better highlight the local characteristics (such as the hook nose) of the input photo.

Applications

Multiple References Unlike other methods that can only exaggerate shapes based on one reference caricature, our method can take additional references to exaggerate a particular part of the input image. As shown Fig. 10, the first line's output refer to the left reference's facial structure and nose and the right reference's mouth, highlighting the features of a slim face, narrow nose, and big mouth in the photo.

Exaggeration Scale Unlike other methods that can only control the exaggeration of the overall shape, our method can easily control the exaggeration of both the overall shape and facial features. As shown in Fig. 11, the first row on the right shows the results of gradually increased exaggeration of the face shape from small to large. The second row shows the results of gradually increased exaggeration of the mouth in the reference caricature.

User Perceptual Evaluation

In order to further evaluate our method, we invited 72 volunteers to participate in our perceptual evaluation. The volunteers were divided into two groups: A (37 ordinary users) and B (35 experts in painting). We conducted two evaluations: one on identity study and the other on qualitative study.

Figure 7: Comparison of G2L-CariGAN with other state-of-the-art caricature generation methods with reference caricatures. Our method can highlight the local characteristics of the photo better than other methods in terms of exaggeration of facial features.

Figure 8: The effectiveness of using feature maps instead of vectors as geometric features. Global feature map G ensure that only the face shape is exaggerated without affecting other parts (such as the hairstyle). Similarly, local feature map L can better exaggerate local facial features.

Identity Study For the identity study, we invited 37 volunteers to answer 45 questions. There are 5 questions for each of 9 caricature generation methods including CariGANs (Cao, Liao, and Yuan 2018), WarpGAN (Shi, Deb, and Jain 2019), CariME (Gu et al. 2021), Semantic-CariGAN (Chu et al. 2021), Autotoon (Gong, Hold-Geoffroy, and Lu 2020), StyleCariGAN (Jang et al. 2021), CariPainter (Huang et al. 2022), DualStyleGAN (Yang et al. 2022) and ours. Each question displays a caricature pro-

	w/o L_{cariid} **(%)** ↑	**with** L_{cariid} **(%)** ↑
Recognition Rate	66.83	**72.55**

Table 3: Face recognition rates with and without caricature identity loss function.

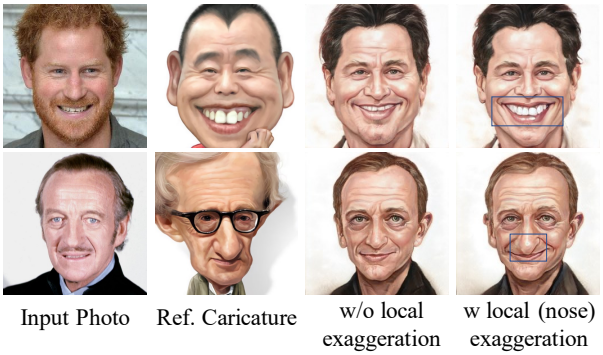

Figure 9: Results with and without local exaggeration.

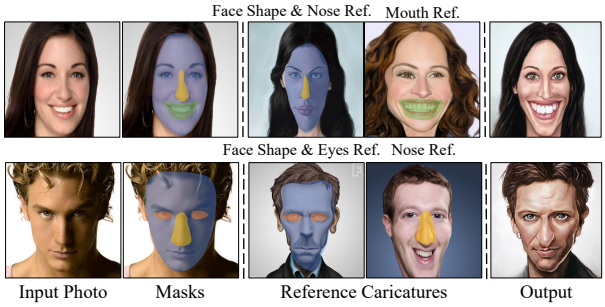

Figure 10: Our results with multiple reference caricatures. We can refer to multiple parts of reference images (for example, the face region and nose region in the first reference and the mouth region in the second reference of the first row) to exaggerate the corresponding region of the input photo.

duced using one of the 9 approaches and asks participants to select a photo from 4 options that represent the same person as the caricature. Among the 4 options, only one photo has the same identity as the caricature, and the other 3 options have similar identity characteristics as interference. The results of the identity evaluation are shown in Fig. 12. It is clear that our method produces a higher recognition accuracy (74.59%) than others indicating a stronger ability to maintain identity.

Qualitative Study We invited 35 artists with rich painting experience. We design 10 questions and divided into two parts: with reference caricature and without reference caricature, each with 5 questions. The questions without reference provide a photo and the results of 5 caricature generation methods (CariGANs (Cao, Liao, and Yuan 2018), WarpGAN (Shi, Deb, and Jain 2019), AutoToon (Gong, Hold-Geoffroy, and Lu 2020), StyleCariGAN (Jang et al. 2021), and CariPainter (Huang et al. 2022)) which can only generate random geometry styles and the results of our method. The questionnaires with reference provide an additional reference caricature and the results of 3 caricature generation methods (CariME (Gu et al. 2021), Semantic-CariGAN(Chu et al. 2021), and DualStyleGAN (Yang et al. 2022)) and our method. Each question requires users to select an option that is closest to a real hand-drawn caricature. The results in Fig.

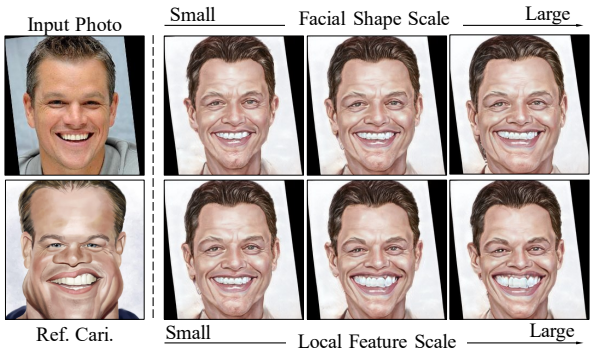

Figure 11: The exaggeration scale of both the overall and facial features. The input photo and the reference caricature are shown on the left of the dashed line. The first row shows different exaggeration scale of face shape and the second row shows the exaggeration scales of the mouth region. Our method can significantly enhance the local features of photos with multiple degrees of exaggeration.

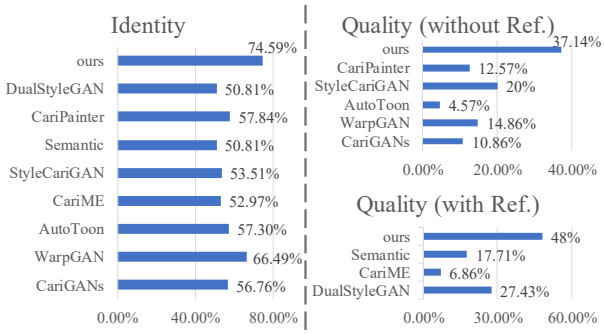

Figure 12: User perceptual evaluation.

12 show that artists agree that those caricatures generated by ours are closer to ones drawn by artists.

All the perception evaluation questions are included in the supplementary material.

Conclusion

In this paper, we have proposed a new GAN-based caricature generation approach G2L-CariGAN. By extracting the caricature's geometry characteristics into two-dimensional feature maps, we combine the overall exaggeration and local exaggeration to clearly highlight a character's local features. To balance the exaggeration of a character and the retaining of its identification, we proposed a new identity loss based on feature maps. Our experiments and ablation studies have demonstrated that G2L-CariGAN can produce caricatures of a greater caliber than existing approaches. We believe our idea of global-to-local caricature generation can be potentially applied to other tasks such as more general image-to-image translation.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.6207071897), and the National Natural Science Regional Joint Foundation of China (U19A2063).

References

Akleman, E. 1997a. Making caricatures with morphing. In *SIGGRAPH Visual Proceedings*, 145. ACM.

Akleman, E. 1997b. Making caricatures with morphing. In *International Conference on Computer Graphics and Interactive Techniques*.

Akleman, E.; Palmer, J.; and Logan, R. 2000. Making extreme caricatures with a new interactive 2D deformation technique with simplicial complexes. In *Proceedings of Visual*, volume 1, 2000. Citeseer.

Brennan, S. E. 1985. Caricature generator: The dynamic exaggeration of faces by computer. *Leonardo*, 18(3): 170–178.

Cao, K.; Liao, J.; and Yuan, L. 2018. CariGANs: unpaired photo-to-caricature translation. *ACM Trans. Graph. (TOG)*, 37(6): 244:1–244:14.

Chen, W.; Yu, H.; Shi, M.; and Sun, Q. 2009. Regularity-Based Caricature Synthesis. In *2009 International Conference on Management and Service Science*, 1–5. IEEE.

Chu, W.; Hung, W.-C.; Tsai, Y.-H.; Chang, Y.-T.; Li, Y.; Cai, D.; and Yang, M.-H. 2021. Learning to caricature via semantic shape transform. *International Journal of Computer Vision (IJCV)*, 1–17.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Gong, J.; Hold-Geoffroy, Y.; and Lu, J. 2020. AutoToon: Automatic Geometric Warping for Face Cartoon Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 360–369.

Gu, Z.; Dong, C.; Huo, J.; Li, W.; and Gao, Y. 2021. CariMe: Unpaired Caricature Generation with Multiple Exaggerations. *IEEE Transactions on Multimedia (TMM)*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1501–1510.

Huang, X.; Liang, D.; Cai, H.; Zhang, J.; and Jia, J. 2022. CariPainter: Sketch Guided Interactive Caricature Generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 1232–1240. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, 172–189.

Huo, J.; Li, W.; Shi, Y.; Gao, Y.; and Yin, H. 2018. WebCaricature: a benchmark for caricature recognition. In *BMVC*, 223. BMVA Press.

Jang, W.; Ju, G.; Jung, Y.; Yang, J.; Tong, X.; and Lee, S. 2021. StyleCariGAN: caricature generation via StyleGAN feature map modulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–16.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4401–4410.

Lai, K.; Chung, P.; and Edirisinghe, E. 2006. Novel approach to neural network based caricature generation.

Liang, L.; Chen, H.; Xu, Y.-Q.; and Shum, H.-Y. 2002. Example-Based Caricature Generation with Exaggeration. In *PG*, 386–393. IEEE Computer Society.

Liu, J.; Chen, Y.; and Gao, W. 2006. Mapping learning in eigenspace for harmonious caricature generation. In *ACM Multimedia*, 683–686. ACM.

Liu, J.; Chen, Y.; Xie, J.; Gao, X.; and Gao, W. 2009. Semi-supervised learning of caricature pattern from manifold regularization. In *International Conference on Multimedia Modeling*, 413–424. Springer.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised Image-to-Image Translation Networks. In *NIPS*, 700–708.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 3730–3738.

Mo, Z.; Lewis, J. P.; and Neumann, U. 2004. Improved automatic caricature by feature normalization and exaggeration. In *ACM SIGGRAPH 2004 Sketches*, 57. ACM.

Pinkney, J. N.; and Adler, D. 2020. Resolution dependent gan interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*.

Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; and Cohen-Or, D. 2021. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, Y.; Deb, D.; and Jain, A. K. 2019. Warpgan: Automatic caricature generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10762–10771.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Yang, S.; Jiang, L.; Liu, Z.; and Loy, C. C. 2022. Pastiche Master: Exemplar-Based High-Resolution Portrait Style Transfer. In *CVPR*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2223–2232.

Zhu, P.; Abdal, R.; Femiani, J.; and Wonka, P. 2021. Barbershop: GAN-based Image Compositing using Segmentation Masks. arXiv:2106.01505.