RD-GAN: Few/Zero-Shot Chinese Character Style Transfer via Radical Decomposition and Rendering

Yaoxiong Huang[1,2], Mengchao He[2], Lianwen Jin *[1], and Yongpan Wang[2]

[1] School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China
[2] Alibaba Group, Hangzhou, China
hwang.yaoxiong@gamil.com, mengchao.hmc@alibaba-inc.com
lianwen.jin@gamil.com, yongpan@taobao.com

Abstract. Style transfer has attracted much interest owing to its various applications. Compared with English character or general artistic style transfer, Chinese character style transfer remains a challenge owing to the large size of the vocabulary(70224 characters in GB18010-2005) and the complexity of the structure. Recently some GAN-based methods were proposed for style transfer; however, they treated Chinese characters as a whole, ignoring the structures and radicals that compose characters. In this paper, a novel radical decomposition-and-rendering-based GAN(RD-GAN) is proposed to utilize the radical-level compositions of Chinese characters and achieves few-shot/zero-shot Chinese character style transfer. The RD-GAN consists of three components: a radical extraction module (REM), radical rendering module (RRM), and multi-level discriminator (MLD). Experiments demonstrate that our method has a powerful few-shot/zero-shot generalization ability by using the radical-level compositions of Chinese characters.

Keywords: GAN, Style Transfer, Radical Decomposition, Few-Shot/Zero-Shot Learning

1 Introduction

With the development of deep learning, character recognition has reached an unprecedented stage of development; however, it is very data-dependent. In many cases, such as in historical documents, character samples are expensive/difficult to obtain. One of the most efficient ways to obtain character samples is to generate character data via style transfer. Unfortunately, character generation remains a relatively under-explored problem compared with the automatic recognition of characters [24, 34]. This unbalanced progress is detrimental to the development of optical character recognition.

Recently, there have been many attempts to generate simple characters such as English and Latin characters [13, 21]; however, Chinese character generation

* Corresponding author

has not been explored extensively. Compared with English or Latin character generation, Chinese character generation is much more challenging owing to the following characteristics. First, Chinese characters share an extremely large vocabulary. To address this issue, [61] generated Chinese characters by introducing a Recurrent Neural Network (RNN)-based model and learned a low-dimensional character label embedding. However, this method can only generate characters that the model has seen. Unfortunately, it is almost impossible to obtain all of the categories' samples from a fixed style. Generating unseen Chinese characters remains an urgent problem.

Moreover, Chinese characters contain a large number of glyphs with complicated content and characteristic style that vary from the shapes of the component and the stroke styles. Recent works such as "Rewrite" [2] and its advanced version "zi2zi" [3] generated Chinese characters by learning to map the source style to a target style with thousands of character pairs for strong supervision. However, these methods still cannot generate unseen Chinese characters.

Finally, unlike the photo-to-artwork task, Chinese characters have a complex and flexible structure. Subtle errors in the skeleton and stroke are obvious and unacceptable. Some attempts have been made in Chinese character generation by assembling components of radicals and strokes [52, 48, 55]. However, these performed poorly for two reasons: 1) they are largely dependent on the performance of radical/stroke extraction while perfect automatic radical/stroke extraction is almost impossible in real applications; and 2) they pay more attention to the rendering of the radical/stroke while ignoring their internal relationship.

Although Chinese characters comprise an extremely large vocabulary, more than 10,000 characters can be composed by approximately 1000 radicals [46]. Meanwhile, all Chinese characters can be decomposed into a unique radical string. When people learn Chinese characters, they first learn the radicals and structures that form characters. By learning radicals and structures, the difficulty of learning to read and write Chinese characters decreases significantly.

Compelled by the above observations, we propose a novel radical decomposition-and-rendering-based GAN(RD-GAN) for Chinese character style transfer that can efficiently generate unseen Chinese characters with a few samples. The RD-GAN consists of three components: a radical extraction module (REM) to extract the radical roughly, radical rendering module (RRM) that learns how to render the radical with stroke details in the target style, and multi-level discriminator (MLD) that guarantees the global structure and local details of the generated character images. The advantages of the proposed RD-GAN can be summarized as follows:

- Owing to the specificity of the relationship between characters and radicals, we can use only a few samples to generate unseen Chinese characters efficiently. This can largely reduce the difficulty and labor of collecting training data.
- By decomposing Chinese characters into radicals, the rendering difficulty decreases significantly.

– Owing to the multi-level discriminator, we can generate stylized Chinese characters that not only have good details but also have more realistically combined components.
– RD-GAN can generate realistic character samples for training character classifiers with few real data. Experiments show that our method can effectively transfer unseen Chinese characters and obtain better performance than recent state-of-the-art methods.

2 Related Work

2.1 Image-to-Image Translation

Image-to-image translation learns the mapping from the input image to the output image and covers many tasks such as edge/contour extraction [38, 50], semantic segmentation [28, 36], artistic style transfer [19, 9], and image colorization [31, 59]. Pix2pix [16] used a conditional GAN based network that needs a significant amount of paired data for training. To alleviate the problem of obtaining data pairs, unpaired image-to-image translation frameworks [26, 27, 63] have been proposed. Liu et al. [26] made a shared-latent space assumption that a pair of corresponding images in different domains can be mapped to the same latent representation in a shared-latent space. After that, authors [27] extended [26] to an unsupervised image-to-image translation problem. Then, authors [63] proposed the cycle-consistent adversarial network (CycleGAN), which performs well for many vision and graphics tasks. Meanwhile, supervised GAN-based methods require numerous image pairs, while unsupervised methods often cause blurred and incorrect construction. In this paper, we propose a novel radical decomposition-and-rendering GAN that focuses on training a generative model with as few samples as possible.

2.2 Character Style Transfer

Recent studies considered character style transfer as an image translation task. A popular project named "Rewrite" [2] implemented a simple traditional flavor top-down Convolutional Neural Network (CNN) to transfer a standard font to another stylized font. After that, its advanced version, named "zi2zi" [3], implemented the font style transfer of Chinese characters by learning to map the source style to a target style with thousands of character pairs. Upchurch et al. [43] adopted a supervised method and assigned each character a one-hot label. In addition, Lyu et al. [32] proposed an auto-encoder-guided GAN network (AEGN) to synthesize calligraphy images with specified styles from standard Chinese font images. Easyfont [23] extracted strokes from given Chinese characters and learned to generate corresponding strokes for other characters in the same style. Jiang et al. [18] integrated the domain knowledge of Chinese characters with deep generative networks to ensure that high-quality glyphs with correct structures can be synthesized. MC-GAN [4] synthesizs ornamented glyphs from images of a few example glyphs in the same style by predicting the coarse glyph shapes and texture of the final glyphs.

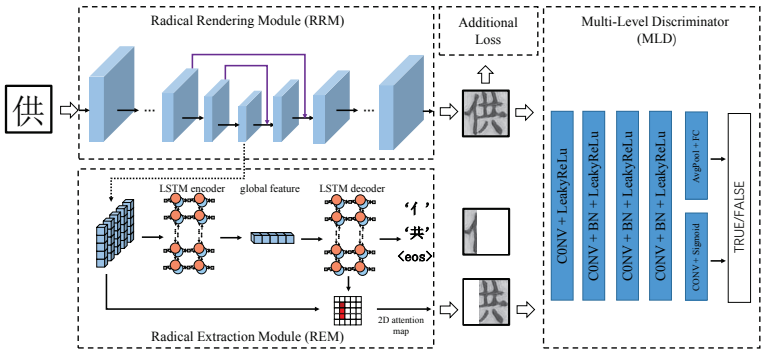

Fig. 1. Overview of the proposed method for Chinese character style transfer. Standard font Chinese character images are firstly fed into the radical rendering module to obtain the output stylized images. The output stylized images and the corresponding target images are then transmitted to the radical extraction module to obtain a 2D-attention map. With the 2D-attention map, we crop the corresponding radical regions from the output images. Finally, output images and cropped images are fed to the multi-level discriminator to improve the distributional similarity between the output images and the corresponding target images.

2.3 Attention Mechanism

The attention mechanism was first proposed in machine translation [5, 44] to enable a model to automatically search for parts of a source sentence for prediction. Then, the method rapidly became popular in applications such as (visual) question answering [30, 54], image caption generation [54, 51, 29], speech recognition [6, 20], and scene text recognition [40, 7, 22]. Most important, the attention mechanism can also be applied to 2D predictions, such as mathematical expression recognition [56, 57], paragraph recognition [49, 8], and radical recognition [45]. Thanks to the characteristics of the attention mechanism, we implement a 2D attention mechanism for rough radical extraction.

3 Proposed Methodology

3.1 Overview

Given a standard-font Chinese character image $\boldsymbol{I_C}$ with content C, our proposed system f aims to generate another stylized character image $\boldsymbol{I'_C}$ with the same content as realistically as possible. The proposed RD-GAN is a network for few-shot/zero-shot Chinese character image generation. As illustrated in Figure 1, the proposed RD-GAN consists of three components. The **Radical Extraction Module** splits an image into different parts to empower our system with few-shot/zero-shot learning. The **Radical Rendering Module** outputs stylized character images based on stroke/radical details. Finally, a **multi-level discriminator** is adopted to pay more attention to both local details and global context.

3.2 Radical Extraction Module

As presented in Figure 1, the radical extraction module consists of two main parts: a weight-shared CNN for feature extraction and a 2D-attention based encoder-decoder model. It takes a character image as input and outputs a varying length sequence of radicals. Meanwhile, the learned 2D-attention map can be used for radical decomposition.

Encoder Inspired by SAR [22], we implement a two-layer Bi-directional LSTM (BLSTM) as an encoder to handle the 2D feature maps from the weight-shared CNN. As shown in Figure 2, we compress each column feature along the vertical direction by average-pooling at each time step, and use the compressed feature to update the hidden state $\boldsymbol{h_t}$. After T steps, which is the width of the 2D feature maps, the final hidden state of the encoder $\boldsymbol{f_g}$ is output as the global feature of our input character image, and is fed to the following decoder.

Decoder The decoder is another BLSTM model with two layers. Most traditional 2D attention models [8, 56] consider only local information and treat each location independently, neglecting the relationship between pixels in adjacent areas. We follow the concept of [22] to take neighborhood information into consideration.

Initially, the global feature $\boldsymbol{f_g}$ is fed into the decoder BLSTM. The decoder iteratively updates the attention mechanism and outputs the current prediction(radical) according to the previous output y_{t-1} and hidden state s_{t-1}:

$$\hat{s_t} = BLSTM(y_{t-1}, s_{t-1}) \tag{1}$$

$$c_t = f_{attn}(\hat{s_t}, \boldsymbol{F}) \tag{2}$$

$$s_t = BLSTM(c_t, \hat{s_t}) \tag{3}$$

$$y_t = \psi(s_t) \tag{4}$$

where $\psi(.)$ is a linear transformation, and f_{attn} is a neighborhood-considered 2D attention mechanism as follows:

$$e_{ij} = W^{attn} \cdot tanh(W^s \hat{s_t} + W^p p_{ij} + \sum_{x=i-1}^{i+1} \sum_{y=j-1}^{j+1} W^h_{xy} p_{xy}) \tag{5}$$

$$\alpha_{ij} = \frac{e_{ij}}{\sum_{ij} e_{ij}} \tag{6}$$

$$c_t = \sum_{ij} \alpha_{ij} p_{ij} \tag{7}$$

Note that, W^{attn}, W^s, W^p, and W^h are all learnable parameters; p_{ij} is the local feature vector at position (i, j) in the input 2D feature map $\boldsymbol{F}$; and i, j are range from 0 to the width and height of $\boldsymbol{F}$, respectively. According to the response of 2D attention map, we can roughly separate the corresponding radicals for subsequent processes.

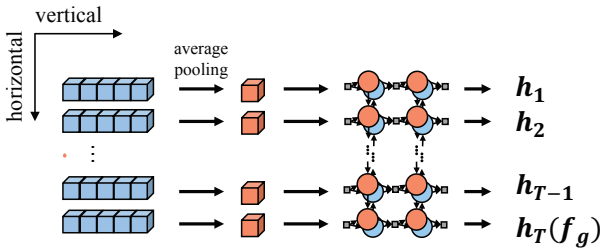

Fig. 2. Illustration of the BLSTM encoder. T represents the width of 2D feature map. At each time step, column feature is compressed by a vertical average-pooling and the final hidden state of the encoder is fed to the following decoder.

3.3 Radical Rendering Module

The overall network architecture of the radical rendering module is shown in Figure 3. The RRM is composed of a downsampling module and an upsampling module. Lateral connections are employed to preserve more details. Fed with a standard-font Chinese character image, the RRM generates a stylized character image with complete strokes and radicals.

It is typically considered that higher-level features share stronger semantics, while lower-level features exhibit semantically weak features but include more detailed information [39] such as texture and position information. Therefore, we introduce a lateral connection to mix up semantics features with detailed geometric features. As shown in Figure 4, this lateral connection contains two pathways, named SE-pathway and up-sample pathway, respectively.

index	Layer(Weight-Shared CNN)	output shape (C×H×W)
0	input: character image	1×96×96
1	num: 64, kernel: 3×3, pad: (h:1, w:1), stride: (h:1, w:1)	64×96×96
2	64, 3×3, (1,1), (1,1)	64×96×96
3	kernel: 2×2, stride: 2×2	64×48×48
4	[128, 3×3, (1,1), (1,1)] ×2	128×48×48
5	2×1, 2×2	128×24×24
6	[256, 3×3, (1,1), (1,1)] ×2	256×24×24
7	2×1, 2×2	256×12×12
8	512, 3×3, (1,1), (1,1)	512×12×12

index	Layer(Up-Sample)	output shape (C×H×W)
18	output: stylized image	1×96×96
17	1, 1×1, (0,0), (1,1)	1×96×96
16	ratio: 2×2	64×96×96
15	[64, 3×3, (1,1), (1,1)] ×2	64×48×48
14	ratio: 2×2	128×48×48
13	[128, 3×3, (1,1), (1,1)] ×2	128×24×24
12	ratio: 2×2	256×24×24
11	[256, 3×3, (1,1), (1,1)] ×2	256×12×12
10	ratio: 2×2	512×12×12
9	2×1, 2×2	512×6×6

convolution layer | max pooling layer | up-sample layer | ⟺ lateral connection

Fig. 3. Illustration of the radical rendering module. It is a convolutional neural network with an encoder decoder structure and lateral connection. Specially, the encoder part is shared with radical extraction module as feature extractor.

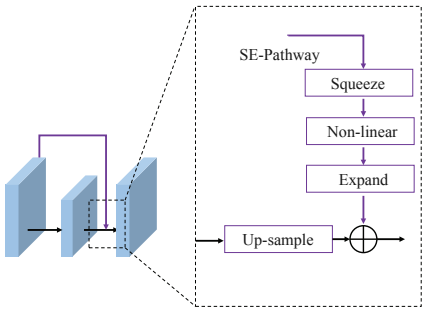

Fig. 4. Illustration of the lateral connection.

SE-Pathway Unlike [15], our operation is performed in the channel dimension. First, the squeeze layer reduces the feature dimensions by using a 1×1 convolution for information fusion from different channels and less computation. Then, we implement a non-linear layer including two deformable convolutions and ReLU activation function to achieve large receptive fields. Last, an expand layer is used to enlarge the feature map channels by a 1×1 convolution as the reverse of the squeeze layer.

Up-sample Pathway To enlarge the feature map, we utilize a bilinear sampling operation followed by two same-size (3×3) convolutional layers. This can avoid checkerboard artifacts more effectively than deconvolution. The up-sampled feature maps are then element-wise summed with the corresponding ones from the SE-Pathway. Inspired by [10], we use ELU as non-linear activation function because it can handle large negative responses and thereby stabilize the training process.

3.4 Multi-Level Discriminator

To differentiate fake images from real ones, the original GANs [12] discriminate the results based on the entire image level. However, as mentioned before, Chinese characters have complex and flexible structures, which make the discriminator difficult to focus on. To determine whether the images we generate are realistic enough to lead the RRM to generate more realistic images, we proposed a multi-level discriminator. Here, "multi-level" includes both the geometry level and structure level.

Geometry Level As noted in [41], any local patch sampled from the generated image should have statistics similar to those of a real image patch. Specifically, following the PatchGAN [16], we execute our discriminator to classify whether each $N \times N$ patch in an image is real or fake. We run this discriminator convolutationally across the image, averaging all responses to provide the ultimate

output of the discriminator. We define the loss as follows:

$$L_D^g = -\frac{1}{N^2}\sum_{n=1}^{N^2}(1-y)log(\xi_n) \tag{8}$$

where y and ξ are the label of the image and the discriminator prediction, respectively. In our experiments, N is set to 6.

Structure Level All Chinese characters can be decomposed into a unique radical set. A perfectly generated Chinese character image should have well refined strokes and more realistically combined components. Therefore, we penalize both the entire image and the radical patches. Given the 2D-attention map $\mathcal{M}_t$ at timestep t and input image $\mathcal{I}$, we define the structure level loss as follows:

$$L_D^s = -(1-y)[log(D^s(\mathcal{I})) + \sum_{t=1}^{T} log(D^s(\mathcal{I} \otimes \mathbb{I}\{\mathcal{M}_t > \theta\}))] \tag{9}$$

where $\mathbb{I}\{\cdot\} = 1$ when the condition is true and zero otherwise. Besides, D^s represents a convolutional network with binary classification outputs, θ denotes the threshold set to 0.5 and $\otimes$ is element-wise multiplication.

By combining the above two loss terms, we come to the overall adversarial training objective:

$$L_a = L_D^g + L_D^s \tag{10}$$

3.5 Additional Loss Function

We aim to generate stylized Chinese character images as realistically as possible. This includes both per-pixel reconstructed accuracy as well as radical composition, i.e. how smoothly the radical regions can harmonize with their surrounding context. Inspired by recent image processing tasks (neural style transfers [11] and text eraser [60]), three additional losses are applied to our system as follows:

L2 Loss($\boldsymbol{L_2}$) To obtain an output image that is both invariant to content and complete in describing a Chinese character, we can simply minimize the difference between the output image $\boldsymbol{I'_C}$ and target image $\hat{\boldsymbol{I}}_{\boldsymbol{C}}$ in an explicit fashion:

$$L_2 = \|\boldsymbol{I'_C} - \hat{\boldsymbol{I}}_{\boldsymbol{C}}\|_2 \tag{11}$$

Total Variation Loss($\boldsymbol{L_{tv}}$) For image generation, a common problem is that model tends to generate noisy images. To solve this problem, we adopt L_{tv} [19] for global denoising, as defined below:

$$L_{tv} = \sum_{ij} \|\boldsymbol{I'_C}[\boldsymbol{i},\boldsymbol{j}] - \boldsymbol{I'_C}[\boldsymbol{i+1},\boldsymbol{j}]\|_1 + \|\boldsymbol{I'_C}[\boldsymbol{i},\boldsymbol{j}] - \boldsymbol{I'_C}[\boldsymbol{i},\boldsymbol{j+1}]\|_1 \tag{12}$$

Here, i, j indicates the position of the pixel.

Content Loss(L_c) As noted in [19], the loss function measured for different high-level features is effective for feature reconstruction. To better generate images from different levels, we introduce content loss to penalize the discrepancy between the features of the output images and the corresponding ground truth images on certain layers in the CNN. We feed the output images and ground truth images to a pre-trained model and force the response in the corresponding layer to be matched. The content loss can be formulated as follows:

$$L_c = \sum_{n=1}^{N-1} \|\phi_n(\boldsymbol{I'_C}) - \phi_n(\hat{\boldsymbol{I_C}})\|_1 \tag{13}$$

where N and $\phi_n(.)$ are the layer index we choose and the feature responding in layer n, respectively. Following [60], we compute the content loss at layers pool1, pool2, and pool3 of a pretrained VGG16 [42].

Combining all of the above loss terms, we come to the overall training objective for our Chinese character style transfer model:

$$L = L_a + \alpha_1 L_2 + \alpha_2 L_{tv} + \alpha_3 L_c \tag{14}$$

where α_1, α_2, and α_3 are weighting coefficients that are empirically set to 0.5, 0.5, and 0.1 in our experiments, respectively.

4 Experiments

4.1 Dataset

In the following experiments, a historical document dataset named TKH Dataset [53] is used to quantitatively and qualitatively evaluate the performance of Chinese character style transfer. TKH has 1000 manually annotated Tripitaka paragraph images composed of approximately 320,000 character instances. There are two benefits to using this dataset: 1) The Chinese characters in historical documents are much closer to handwritten characters, which vary in the shapes of the components and the stroke styles. This is more useful in verifying the robustness and effectiveness of our method for difficult styles. 2) There are many strange and unusual characters in historical documents and they are an ideal testbed for few-shot/zero-shot learning.

To meet the requirements of different experiments and metrics, the entire dataset is partitioned into three subsets: $\boldsymbol{\mathcal{D}1}$, where both the training set and test set have the same category (1473 classes in the same style) with 50 samples for each category; $\boldsymbol{\mathcal{D}2}$, a subset of $\boldsymbol{\mathcal{D}1}$ with only 5 samples for each category in the training set; and $\boldsymbol{\mathcal{D}3}$, where images in the test set (213 classes) are never seen in the training set (1260 classes) but share the same radical sets. The three datasets represent different levels of challenges, e.g., supervised learning, few-shot learning, and zero-shot learning.

4.2 Experiment Setup

We use TrueType fonts to render the corresponding characters in black with font style Song as standard font character images. Both the standard font character images and target character images are resized to 96×96 before being fed to the model. We set the initial learning rate as 0.0001 and train the model end-to-end with the Adam optimization method until the output is stable.

Radicals are viewed as a part of the semantics and are shared by different characters [35]. Many studies have examined the reasonable splitting of Chinese characters into radical sets. [33] extracted 1118 substructures from 4284 characters to build a radical lexicon, and [46] extended this lexicon to 9820 characters. In our experiments, we adopt the radical lexicon in project [1] and filter out the symbols and single-structure characters in the dataset that cannot be decomposed into smaller parts. There are 1473 characters with 576 radicals for experiments. In our experiments, we train an REM with 150,000 samples generated by [17] and achieve an accuracy of 98.7% as tested on synthetic data.

Although the most commonly used metric for determining the quality of generative models is the inception score [37], it is not suitable for Chinese character style transfer [25]. To impartially compare the proposed method with other recent works, we calculate the L1 loss, Root Mean Square Error (RMSE) and the structual similarity (SSIM) [47] between the generated images and target images. In addition, one of the most important purposes of Chinese character style transfer is to improve the classifier performance. Therefore, we compare the performance among classifiers trained with different generation methods. We adopt a character recognizer with 1473 classes as follows:

$76 \times 76 Input - 32C3 - MP2 - 64C3 - MP2 - 128C3 - 128C3 - MP2 - 256C3 - 256C3 - MP2 - 384C3 - 384C3 - FC1024 - FC1473 - Output$

where xCy represents a convolutional layer with kernel number of x and kernel size of $y \times y$, MPx denotes a max-pooling layer with kernel size of x, and FCx is a fully connected layer of kernel number of x.

4.3 Experimental Results

Comparison with State-of-the-Art Methods In this subsection, we compare our model with the following methods for Chinese character style transfer from the perspective of supervised learning, few-shot learning, and zero-shot learning:

1) Pix2pix [16]: Pix2pix is a conditional GAN based image translation network and is optimized by L1 distance loss and adversarial loss.

2) Cycle-GAN [63]: Cycle-GAN not only learns the mapping from the input image to the output image, but also learns a loss function to reverse this mapping. It is noted that Cycle-gan only requires unpaired data.

3) MC-GAN [4]: MC-GAN is the first end-to-end solution to synthesizing ornamented glyphs from images of a few example glyphs in the same style.

4) Zi2zi [3]: Zi2zi is an application and extension of the pix2pix model to Chinese characters with the addition of category embedding.

5) EMD [62]: EMD is a generalized style transfer network that attempts to separate the representations for style and content.

Supervised Learning: In this part, all methods are trained with paired images in $\mathcal{D}\mathbf{1}$. The results are displayed in Figure 5 (a). We observe that for supervised learning, our proposed method outperforms Pix2pix and Cycle-GAN and is slightly better than the results of Zi2zi and EMD. It is noted that Cycle-GAN has the worst performance, as it can only generate parts of characters or sometimes unreasonable structures. This may be because it only learns the domain mappings without the domain knowledge [62]. Zi2zi and EMD can learn how to map standard fonts to stylized fonts through an abundant number of paired images, as in our method. For quantitative analysis, we conducted experiments three times with different initializations. The average results are displayed in the last three columns in Figure 5 (a). We can observe that our method performs best and achieves the lowest L1 loss, RMSE and the highest SSIM.

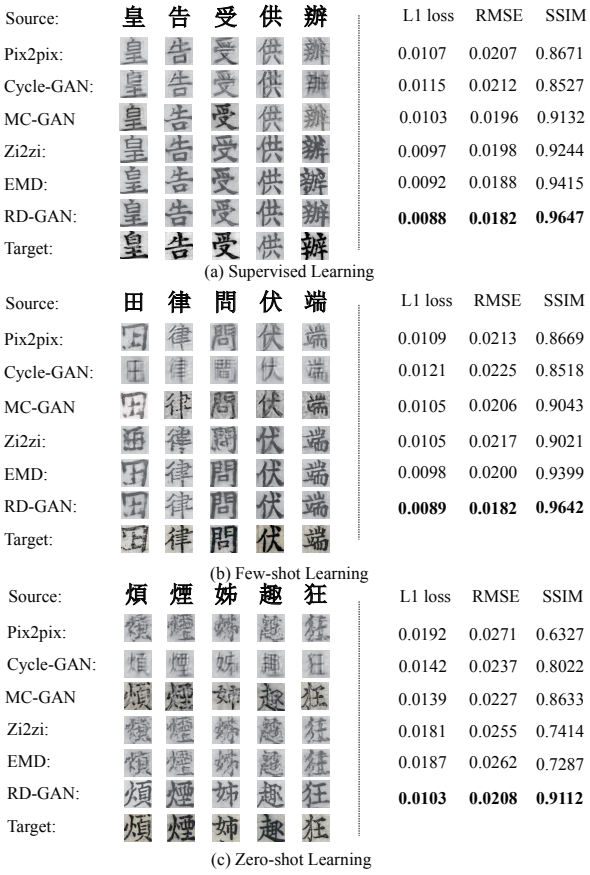

Fig. 5. Comparison among previous methods and proposed method.

Table 1. Character recognition accuracy trained with different generation methods

Generation Methods	Accuracy(%)
$\boldsymbol{\mathcal{D}1}$(Real data only)	82.45
$\boldsymbol{\mathcal{D}1}$+synthtext2014 [17]	83.25
$\boldsymbol{\mathcal{D}1}$+pix2pix [16]	84.74
$\boldsymbol{\mathcal{D}1}$+zi2zi [3]	85.63
$\boldsymbol{\mathcal{D}1}$+EMD [62]	87.21
$\boldsymbol{\mathcal{D}1}$+RD-GAN	88.11

Few-Shot Learning: We train our model and other methods with $\boldsymbol{\mathcal{D}2}$, in which there are only five samples for each category. The results are presented in Figure 5 (b). As shown in the figure, our method and EMD still exhibit good performance, while the performance of Zi2zi drops sharply. It is unrealistic for Zi2zi to transfer font style for 1473 categories trained with only 5 samples for each category. However, by decomposing Chinese characters into radicals, the number of categories we need to learn is reduced from 1473 to 576. This significantly reduces the difficulty of our task. In addition, all Chinese characters share the same radical lexicon, which means a radical will appear in both character A and character B. Therefore, a sufficient number of radical samples can be used for our training. The quantitative comparison results including the L1 loss, RMSE, and SSIM are also shown in the last three columns of Figure 5 (b). Note that our model achieves the best performance among all of the methods and there is almost no degradation in performance even when training with only five samples, which demonstrates the effectiveness of our method.

Zero-Shot Learning: As an extreme case, all methods are trained with $\boldsymbol{\mathcal{D}3}$. The categories in the test set were never seen during training. Both qualitative and quantitative analyses are presented in Figure 5 (c). As shown in Figure 5 (c), our model can still deal with unseen categories in the training set. However, images generated by other methods are messy, and their content may not be recognized. This is because these methods treat Chinese characters as a whole and cannot generate unseen Chinese characters. Differently, we cannot see the entire characters but all radicals in the test set can be explored during training, giving our method the ability of zero-shot learning.

In conclusion, most of the state-of-the-art methods require many paired images to train, which may difficult to collect images for some special fonts or categories such as historical documents. In addition, these methods can only transfer character font styles for categories appearing in the training set, and with no ability to generate unseen categories. However, our method can generalize stylized characters given only a few reference images. In addition, the experiments indicated the strong few-shot/zero-shot learning ability of our method owing to the relationship between characters and radicals.

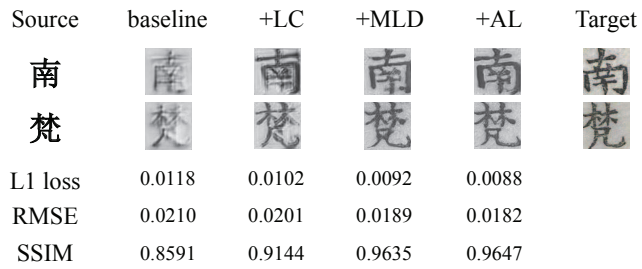

Fig. 6. Effect of different components in our method. LC, MLD and AL represent lateral connection, multi-level discriminator and additional loss, respectively.

Classifier Performance In this part, we train a character classifier using 50,000 synthetic character images generated by different methods and test the classifier on $\mathcal{D}1$. Table 1 lists the classification accuracy using different generation methods. We can observe that the character classifier trained with samples generated by our method performs much better than others. This is also consistent with the results of the quantitative and qualitative analyses mentioned above. It further reflects the outstanding ability of our method to generate data to promote the classifier performance.

4.4 Ablation Studies

In this section, we analyze the influence of the factors influencing the model performance, including the lateral connection, multi-level discriminator, and additional loss.

Lateral Connection: To evaluate the effectiveness of the lateral connection during image generation, we compare the results with and without lateral connections in Figure 6. As shown in the figure, images generated with lateral connections exhibit much better details and obtain a lower L1 loss. This indicates that the lateral connections can effectively learn more detailed information to reconstruct stylized images.

Multi-level Discriminator: Radical decomposition and reconstruction are the key features of the proposed RD-GAN model. The multi-level discriminator is one of the indispensable components. To evaluate the influence of the discriminator, we conduct experiments using the multi-level discriminator and single-level discriminator, which treat images as a whole. The results are displayed in Figure 6. We can observe that images generated with a multi-level discriminator have a better stroke/radical rendering. Besides, they obtain higher SSIM, which indicates that the character structure is better reconstructed.

Additional Loss: In addition, we conduct experiments with additional loss. Figure 6 displays the image generation results with and without additional loss. It it noted that the generated images have more local details, less noise, and better reconstruction with lower L1 Loss, RMSE and higher SSIM.

Fig. 7. Experimental results on face transfer. Images from top to buttom: input images and output results.

4.5 Generalization to Face Transfer

As Chinese characters are composed of multiple radicals, the face is also composed of multiple parts including the eyes, nose, and mouth. In this section, an experiment on face transfer is conducted to test the generalization ability of RD-GAN. We evaluate our method on a well-known face dataset named Ms-celeb-1m [14]. In the experiment, we choose pairs of photos taken for the same person but from different perspectives for training. Besides, we implement MTCNN [58] instead of REM to extract the components of the face. The qualitative results are shown in Figure 7. We can observe that the generated images effectively retain the face information and details, which further reflects the generalization and effectiveness of our method.

5 Conclusion

In this paper, we proposed a novel Chinese character style transfer model named RD-GAN that shows powerful few-shot/zero-shot generalization ability. The main idea is that all Chinese characters share the same radical lexicon, and that the REM decomposes Chinese characters into radical parts. Then, according to the radical parts, the RRM renders the radicals with stroke details in the target style. Finally, an MLD was proposed to guarantee the global structure and local details of the generated characters. To the best of our knowledge, RD-GAN is the first method that can generate Chinese character images of unseen categories with roughly radical decomposition. We evaluated the proposed method on Chinese character style transfer task, and extensive experiment demonstrated its effectiveness.

Acknowledgement

This research is supported in part by NSFC (Grant No.: 61936003), GD-NSF (no. 2017A030312006), Alibaba Innovative Research Foundation (no. D8200510), and Fundamental Research Funds for the Central Universities (no. D2190570).

References

1. Cjkvi. https://github.com/cjkvi/cjkvi-ids
2. Rewrite. https://github.com/kaonashi-tyc/Rewrite
3. Zi2zi. https://github.com/kaonashi-tyc/zi2zi
4. Azadi, S., Fisher, M., Kim, V., Wang, Z., Shechtman, E., Darrell, T.: Multi-content gan for few-shot font style transfer (2017)
5. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
6. Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., Bengio, Y.: End-to-end attention-based large vocabulary speech recognition. In: ICASSP (2016)
7. Bai, F., Cheng, Z., Niu, Y., Pu, S., Zhou, S.: Edit probability for scene text recognition. In: CVPR (2018)
8. Bluche, T.: Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. In: NIPS (2016)
9. Chen, T.Q., Schmidt, M.: Fast patch-based style transfer of arbitrary style. arXiv preprint arXiv:1612.04337 (2016)
10. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
13. Graves, A.: Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850 (2013)
14. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV (2016)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
16. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
17. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014)
18. Jiang, Y., Lian, Z., Tang, Y., Xiao, J.: Scfont: Structure-guided chinese font generation via deep stacked networks. In: AAAI (2019)
19. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
20. Kim, S., Hori, T., Watanabe, S.: Joint ctc-attention based end-to-end speech recognition using multi-task learning. In: ICASSP (2017)
21. Lake, B.M., Salakhutdinov, R., Tenenbaum, J.B.: Human-level concept learning through probabilistic program induction. Science (2015)
22. Li, H., Wang, P., Shen, C., Zhang, G.: Show, attend and read: A simple and strong baseline for irregular text recognition. In: AAAI (2019)
23. Lian, Z., Zhao, B., Chen, X., Xiao, J.: Easyfont: a style learning-based system to easily build your large-scale handwriting fonts. ACM Transactions on Graphics (TOG) (2018)
24. Lian, Z., Zhao, B., Xiao, J.: Automatic generation of large-scale handwriting fonts via style learning. In: SIGGRAPH ASIA 2016 Technical Briefs (2016)
25. Lin, Q., Liang, L., Huang, Y., Jin, L.: Learning to generate realistic scene chinese character images by multitask coupled gan. In: PRCV (2018)

26. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: NIPS (2017)
27. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: NIPS (2016)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
29. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: CVPR (2017)
30. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: NIPS (2016)
31. Luan, Q., Wen, F., Cohen-Or, D., Liang, L., Xu, Y.Q., Shum, H.Y.: Natural image colorization. In: Proceedings of the 18th Eurographics conference on Rendering Techniques (2007)
32. Lyu, P., Bai, X., Yao, C., Zhu, Z., Huang, T., Liu, W.: Auto-encoder guided gan for chinese calligraphy synthesis. In: ICDAR. vol. 1 (2017)
33. Ma, L.L., Liu, C.L.: A new radical-based approach to online handwritten chinese character recognition. In: ICPR (2008)
34. Miyazaki, T., Tsuchiya, T., Sugaya, Y., Omachi, S., Iwamura, M., Uchida, S., Kise, K.: Automatic generation of typographic font from small font subset. IEEE computer graphics and applications (2019)
35. Myers, J.: Knowing chinese character grammar. Cognition (2016)
36. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: ICCV (2015)
37. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NIPS (2016)
38. Shen, W., Zhao, K., Jiang, Y., Wang, Y., Zhang, Z., Bai, X.: Object skeleton extraction in natural images by fusing scale-associated deep side outputs. In: CVPR (2016)
39. Shen, X., Chen, Y.C., Tao, X., Jia, J.: Convolutional neural pyramid for image processing. CVPR (2017)
40. Shi, B., Yang, M., Wang, X., Lyu, P., Yao, C., Bai, X.: Aster: An attentional scene text recognizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence (2018)
41. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
42. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
43. Upchurch, P., Snavely, N., Bala, K.: From a to z: supervised transfer of style and content using deep neural network generators. arXiv preprint arXiv:1603.02003 (2016)
44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
45. Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. AAAI (2020)
46. Wang, T.Q., Yin, F., Liu, C.L.: Radical-based chinese character recognition via multi-labeled learning of deep residual networks. In: ICDAR (2017)
47. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing (2004)
48. Wen, C., Chang, J., Zhang, Y.: Handwritten chinese font generation with collaborative stroke refinement. arXiv preprint arXiv:1904.13268 (2019)

49. Wigington, C., Tensmeyer, C., Davis, B., Barrett, W., Price, B., Cohen, S.: Start, follow, read: End-to-end full-page handwriting recognition. In: ECCV (2018)
50. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
51. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML (2015)
52. Xu, S., Jin, T., Jiang, H., Lau, F.C.: Automatic generation of personal chinese handwriting by capturing the characteristics of personal handwriting. In: IAAI (2009)
53. Yang, H., Jin, L., Huang, W., Yang, Z., Lai, S., Sun, J.: Dense and tight detection of chinese characters in historical documents: Datasets and a recognition guided detector. IEEE Access (2018)
54. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR (2016)
55. Yiming Gao, J.W.: Gan-based unpaired chinese character image translation via skeleton transformation and stroke rendering. AAAI (2020)
56. Zhang, J., Du, J., Dai, L.: Track, attend, and parse (tap): An end-to-end framework for online handwritten mathematical expression recognition. IEEE Transactions on Multimedia pp. 221–233
57. Zhang, J., Du, J., Zhang, S., Liu, D., Hu, Y., Hu, J., Wei, S., Dai, L.: Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. Pattern Recognition (2017)
58. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters (2016)
59. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
60. Zhang, S., Liu, Y., Jin, L., Huang, Y., Lai, S.: Ensnet: Ensconce text in the wild. In: AAAI (2019)
61. Zhang, X.Y., Yin, F., Zhang, Y.M., Liu, C.L., Bengio, Y.: Drawing and recognizing chinese characters with recurrent neural network. IEEE transactions on pattern analysis and machine intelligence (2017)
62. Zhang, Y., Zhang, Y., Cai, W.: Separating style and content for generalized style transfer. In: CVPR (2018)
63. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)