GAN-Based Unpaired Chinese Character Image Translation via Skeleton Transformation and Stroke Rendering

Yiming Gao, Jiangqin Wu
College of Computer Science, Zhejiang University, Hangzhou, China
{gor_yatming, wujq}@zju.edu.cn

Abstract

The automatic style translation of Chinese characters (CH-Char) is a challenging problem. Different from English or general artistic style transfer, Chinese characters contain a large number of glyphs with the complicated content and characteristic style. Early methods on CH-Char synthesis are inefficient and require manual intervention. Recently some GAN-based methods are proposed for font generation. The supervised GAN-based methods require numerous image pairs, which is difficult for many chirography styles. In addition, unsupervised methods often cause the blurred and incorrect strokes. Therefore, in this work, we propose a three-stage Generative Adversarial Network (GAN) architecture for multi-chirography image translation, which is divided into skeleton extraction, skeleton transformation and stroke rendering with unpaired training data. Specifically, we first propose a fast skeleton extraction method (ENet). Secondly, we utilize the extracted skeleton and the original image to train a GAN model, RNet (a stroke rendering network), to learn how to render the skeleton with stroke details in target style. Finally, the pre-trained model RNet is employed to assist another GAN model, TNet (a skeleton transformation network), to learn to transform the skeleton structure on the unlabeled skeleton set. We demonstrate the validity of our method on two chirography datasets we established.

Introduction

Chinese characters (CH-Chars) are a complicated and ancient art with the carrier of Chinese culture, of which aesthetic value attracts calligraphy lovers to imitate the works of famous calligraphers. To achieve visually-pleasing imitation, amounts of time and repetitive training are necessary. Unlike English, CH-Chars contain thousands of glyphs and various chirography styles that have great differences in the overall structure and stroke details. Therefore, automatic multi-chirography style translation of CH-Chars from some observed instances is a meaningful and challenging task.

Early researches on CH-Char synthesis mainly include the brush model-based methods(Wong and Ip 2000; Wu et al. 2006), the rendering manuscript methods(Yu and Peng 2005; Zhang, Wu, and Yu 2010) and the image assembly-based methods(Xu et al. 2005; Du, Wu, and Xia 2016). However, these methods are inefficient and dedicated, requiring manual intervention, which also lead to unrealistic results.

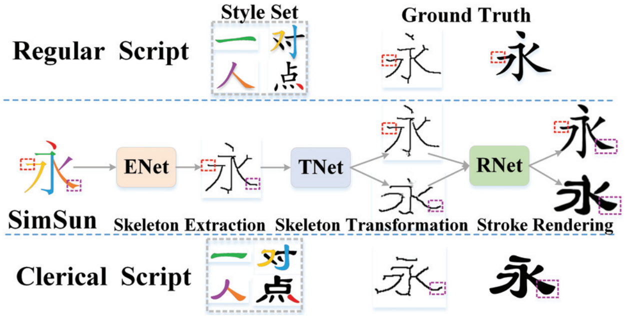


Figure 1: The overview of our method. The middle shows the translation of ”yong” in our method. The top and bottom show the ground truths of ”yong”. The color boxes mean the corresponding stroke details during the process. To visually distinguish different strokes, we utilize different colors to mark the relevant strokes in the figure.

Recently, many deep learning-based methods (Gatys, Ecker, and Bethge 2016; Li and Wand 2016; Johnson, Alahi, and Fei-Fei 2016) achieve satisfactory results in texture features transfer tasks such as photography to artwork, but fail in large geometric variations. Moreover, a variety of GAN-based methods, e.g. PixPix (Isola et al. 2017) and CycleGAN (Zhu et al. 2017), offer the general-purpose solution for image-to-image translation. Subsequently, some font generation methods based on them are proposed, such as zi2zi(Tian 2017), MC-GAN(Azadi et al. 2018), Calli-GAN(Gao and Wu 2019) etc.

However, it is difficult and expensive to collect adequate training pairs for the Pix2Pix-based methods in many cases, e.g. calligraphy fonts, due to the damage and loss in the long history and the complexity and diversity of CH-Chars. Moreover, unlike the photo-to-artwork task, CH-Chars is only black and white, so the subtle errors of skeleton and stroke are obvious and unacceptable, while the CycleGAN-based methods often cause falseness and blur. In order to improve the quality of details, SCFont(Jiang et al. 2019), ReenactGAN(Wu et al. 2018), EverybodyDance(Chan et al. 2018), etc., divide the task into several subtasks according to task character.

In this paper, we propose ChiroGAN, a novel framework for multi-chirography Chinese character image translation based on skeleton transformation and stroke rendering with unpaired training data. We define the chirography style as the skeleton structure and stroke style, because the skeleton contains the basic information of character, such as the composition and position of strokes, writing direction, etc., while the stroke style means the deformation of the skeleton, such as the thickness, shape, writing strength, etc.

To address the aforementioned problems, we divide the task into three stages showed in Fig.1. First, we improve the skeletonization algorithm (ENet) based on mathematical morphology(Jian-ping et al. 2005) by using convolution operations to represent erosion and dilation, and apply it to the images in the dataset. Then, we utilize a skeleton transformation network (TNet) to transfer the structure of the source skeleton to the target style. Finally, the transformed skeleton is rendered with the shape and details of the target style via a stroke rendering network (RNet).

In this manner, the skeleton-image pairs automatically extracted in the first stage are leveraged to train RNet to render the stylistic and clear strokes on the skeleton. For TNet, the correlation between skeletons in different styles can be learned from unpaired skeleton style sets via CycleGAN-based methods. We also propose to train TNet jointly with the pre-trained RNet to compensate for the lack of labeled information in training data, and further constrain the consistency of content. Similar to StarGAN(Choi et al. 2018), we develop a novel cGAN(Mirza and Osindero 2014) architecture to learn the translation of multi-chirography styles with only a single model.

To sum up, our major contributions are as follows:

(1)We propose a novel stacked cGAN models for unpaired multi-chirography Chinese character image translation based on skeleton transformation and stroke rendering.

(2)We improve the skeletonization algorithm based on mathematical morphology such that the process can be greatly accelerated via representing by the neural network.

(3)We build both standard font and real calligraphy datasets, and compare our model with the baseline methods on them to demonstrate the effectiveness of our method.

Related work

Chinese Character Synthesis

Chinese character synthesis is a long studied problem. The brush model-based methods(Wong and Ip 2000; Wu et al. 2006) allow users to manually create calligraphy by modeling the realistic writing environment. The rendering manuscript methods(Yu and Peng 2005; Zhang, Wu, and Yu 2010) utilize the specific stroke texture patches to render the skeleton. The image-based methods(Xu et al. 2005; Du, Wu, and Xia 2016) deform and assemble the corresponding radicals and strokes in dataset to generate the target characters. However, these methods are slow and require manual extraction and intervention.

Neural Style Transfer

(Gatys, Ecker, and Bethge 2016) first uses pre-trained convolutional neural network(CNN) to extract the content and style features. Whereafter, (Johnson, Alahi, and Fei-Fei 2016) propose perceptive loss to end-to-end train style transfer network while retaining the content. These methods have a poor performance in geometric style transfer.

Image-to-Image Translation

Recently, a series of GANs(Goodfellow et al. 2014), e.g. cGAN(Mirza and Osindero 2014), WGAN(Arjovsky, Chintala, and Bottou 2017), Pix2Pix(Isola et al. 2017), CycleGAN(Zhu et al. 2017), StarGAN(Choi et al. 2018) are widely used in the field of image generation.

Like Pix2Pix, various supervised methods are proposed for Chinese font style translation with thousands of training pairs, such as Rewrite(Tian 2016) and zi2zi(Tian 2017) (font pairs), AE-GAN(Lyu et al. 2017) (feature pairs), and SCFont(Jiang et al. 2019) (skeleton flow pairs and stroke semantic map of OptSet). MC-GAN(Azadi et al. 2018) synthesizes the 26 letters from a few examples by learning the correlation between glyphs, but is not suitable for Chinese due to numerous glyphs. CalliGAN(Gao and Wu 2019) achieves unpaired chirography translation, but ghosting artifacts and blur often appear in the results.

ReenactGAN(Wu et al. 2018) utilizes CycleGAN with a PCA shape constrain loss to transfer facial movements and expressions by boundary latent space, then uses Pix2Pix to rebuild the target face. EverybodyDance(Chan et al. 2018) represents the motion with pose stick figures, applys a rigid normalization, and reconstructs the target appearance. For our task, we achieve the non-rigid skeleton transformation with TNet which is trained with pre-trained RNet and a content loss, and then apply RNet for stroke rendering.

Method

People distinguish CH-Chars by the combination of radicals and the topology of strokes. We define that CH-Chars with same content must contain same radicals with similar strokes and layout. For example, the simplified and traditional characters are viewed as different contents. Moreover, we divide the chirography style into the skeleton structure (e.g. the aspect ratio, radical interval, stroke density) and the stroke style (e.g. the thickness, inclination, writing strength, the starting and ending shape). In our task, the content of generated image ought to be similar to the original one and the style is consistent with the target style sets.

Skeleton Extraction (ENet)

According to the definition, the skeleton should contain only the information representing the content without redundant stroke style. The skeletonization algorithm based on mathematical morphology(Jian-ping et al. 2005) leverages a set of mask matrices to erode and dilate the binarized CH-char images with the following equation:

$$X \otimes \{S\} = ((...((X \otimes E_1) \oplus D_2)...) \oplus D_4) \otimes E_8, \quad (1)$$

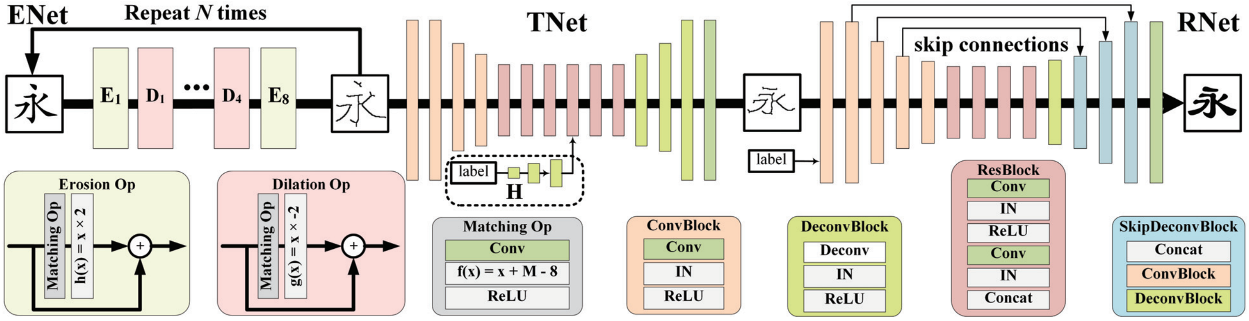


Figure 2: All networks in our model. We use different colors to represent different kinds of components in each network. We define Conv as the convolution layer, IN as the instance normalization(Ulyanov, Vedaldi, and Lempitsky 2016), ConvBlock as Conv-IN-ReLU, DeconvBlock as transposed Conv-IN-ReLU, ResBlock as the residual block(He et al. 2016), and SkipDeconvBlock as ConvBlock-DeconvBlock after the concatenating operation.

0	0	0
*	1	*
1	1	1

E₁

*	0	0
1	1	0
1	1	*

E₂

1	*	0
1	1	0
1	*	0

E₃

1	1	*
1	1	0
*	0	0

E₄

1	1	1
*	1	*
0	0	0

E₅

*	1	1
0	1	1
0	0	*

E₆

0	*	1
0	1	1
0	*	1

E₇

0	0	*
0	1	1
*	1	1

E₈

0	0	0
1	0	1
1	1	1

D₁

1	1	0
1	0	0
1	1	0

D₂

1	1	1
1	0	1
0	0	0

D₃

0	1	1
0	0	1
0	1	1

D₄

Figure 3: Mask matrixes for mathematic morphology. E is the mask matrix for erosion, while D is for dilation. The ”0”, ”1” and ”*” mean white, black and arbitrary respectively.

where X is the CH-char image, $\{S\} = \{E_1, D_1, E_2, E_3, D_2, E_4, E_5, D_3, E_6, E_7, D_4, E_8\}$ are the mask matrices in Fig.3, $\otimes$ is erosion operation and $\oplus$ is dilation operation.

Since matching operation is similar to the convolution, we improve the method as follows. We let $\hat{X}^{xy}$ be the 3×3 image patch centered on point X_{xy} and $\hat{S}$ be the matrix in S. The matching operation $\odot$ for two elements is defined as

$$a \odot b = \begin{cases} 1, & if\ (a=b) \vee (a=*) \vee (b=*) \\ 0, & otherwise \end{cases}. \qquad (2)$$

Thus the original matching operation $\circledcirc$ for image patch $\hat{X}^{xy}$ and mask matrix $\hat{S}$ is expressed as:

$$\hat{X}^{xy} \circledcirc \hat{S} = \prod_{j=1}^{3}\prod_{i=1}^{3}(\hat{X}^{xy}_{ij} \odot \hat{S}_{ij}). \qquad (3)$$

For image X and mask matrices $\{S\}$, if the white, black, and arbitrary in Fig.3 are respectively set as ”-1”, ”1”, ”0” instead of ”0”, ”1”, ”*”, the sum of $\hat{X}^{xy}_{ij} \times \hat{S}_{ij}$ will equal to $P - M$ when $\hat{X}^{xy}$ and $\hat{S}$ match, where P is the number of elements in $\hat{S}$ and M is number of ”*”. The matching operation can be expressed as:

$$\begin{aligned} \hat{X}^{xy} \circledcirc \hat{S} &= (\sum_{j=1}^{3}\sum_{i=1}^{3} \hat{X}^{xy}_{ij} \times \hat{S}_{ij} + M) \odot P \\ &= max((\hat{X}^{xy} \circ \hat{S} + M - 3 \times 3) + 1, 0) \\ &= ReLU(\hat{X}^{xy} \circ \hat{S} + M - 8), \end{aligned} \qquad (4)$$

where $\circ$ means convolution, and $ReLU(x) = max(x, 0)$. In this way, the matching operation can be represented as a convolution layer with the kernel $\hat{S}$ followed by $ReLU$ (see Fig.2). To keep the size, the zero padding is used before it. As Equ.3, the result is still 1 when matching, otherwise 0.

For erosion, X_{xy} is black(1) when $\hat{X}^{xy}$ matches, which has to become white(-1). And similarly for dilation, X_{xy} need to be black(1) from white(-1) while matching. Therefore, the erosion $\otimes$ and dilation $\oplus$ are expressed as:

$$X \otimes \hat{S} = X - 2 * (X \circledcirc \hat{S}), \qquad (5)$$

$$X \oplus \hat{S} = X + 2 * (X \circledcirc \hat{S}), \qquad (6)$$

where these operations can be viewed as a residual mapping(He et al. 2016) $\mathcal{H}(X) = X + \mathcal{F}(X)$. Next, we take each element of $\{S\}$ as a subnet and stack them into a network, called ENet (see the left part of Fig.2). Finally, we input $Y^{(0)} \equiv X$ into ENet to obtain the skeletonized image $Y^{(1)}$, then input $Y^{(i)}$ into the ENet repeatedly until $Y^{(i+1)}$ no longer changes, i = 1,2...N, and $Y \equiv Y^{(N)}$ is the skeleton of X. We try 35,280 characters in 4 styles, and the maximum of i is 12. Therefore, we set $N = 15$ in all experiments. Since the features of single-pixel skeleton is difficult to extract by CNNs, we broadcast it to 4 pixels. Through formula derivation, the improved operations can be easily accelerated in parallel using the existing deep learning framework.

Stroke Rendering Network (RNet)

As Fig.1 shows, the goal of RNet is to render the stroke details of the target style on the skeleton through a cGAN consisting of a generator G_R and a discriminator D_R. Feeding in a skeleton y conditioned on the one-hot vector of the specified style $c \in C$ (C is chirography style set), G_R generates a CH-Char image $G_R(y, c)$ with complete strokes.

In RNet, G_R is composed of a downsampling module, four ResBlocks, and a upsampling module(as shown in the right part of Fig.2). Skip connections are employed to preserve more spatial details. For D_R, we leverage the PatchGANs D^{adv}_R and the auxiliary classifier D^{cls}_R in StarGAN(Choi et al. 2018) to guide the single model G_R to achieve realistically multi-chirography style rendering.

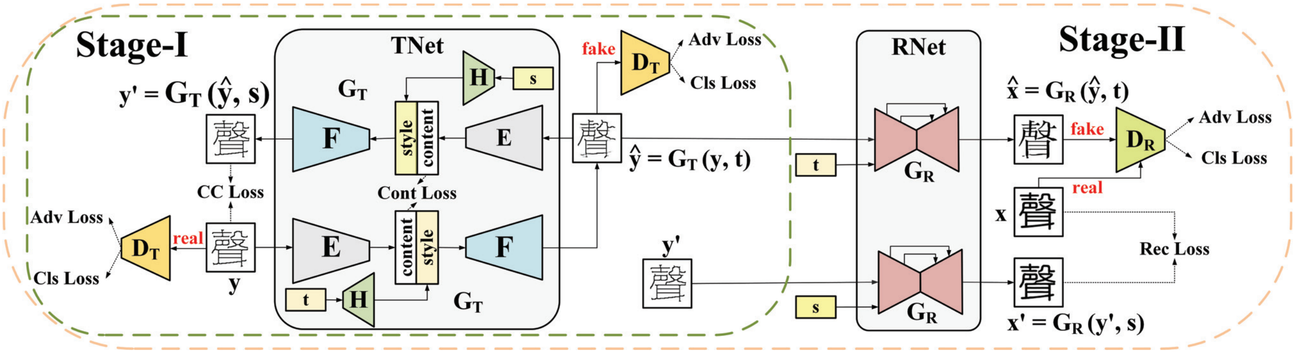


Figure 4: Schematic of the joint training for TNet and RNet. In Stage-I, We first train TNet individually, following the process in the green box. Then, in Stage-II, we utilize the pre-trained RNet (pink box) to further assist the training process of TNet in Stage-I. Moreover, both G_T, D_T and D_R are trained in Stage-II, while freezing G_R.

In order to guide the training of RNet, we first construct a skeleton-image pairs dataset by ENet. And we pre-train RNet on them by optimizing the loss function, which consists of adversarial loss $\mathcal{L}_{adv}$, chirography classification loss $\mathcal{L}_{cls}$ and reconstruction loss $\mathcal{L}_{rec}$. We use WGAN-GP(Gulrajani et al. 2017) loss for adversarial training of both RNet and TNet. Also, the L_1 loss function is used as $\mathcal{L}_{rec}$ on the rendered image and the ground truth.

For G and D, we define the general form of $\mathcal{L}_{adv}$ as:

$$\begin{aligned}\mathcal{L}_{adv}(D) =& \mathbb{E}_{z,c}[D(G(z,c))] - \mathbb{E}_z[D(z)] \\ &+ \lambda_{gp}\mathbb{E}_{\hat{z}}[(\parallel \bigtriangledown_{\hat{z}} D(\hat{z}) \parallel_2 -1)^2],\end{aligned} \tag{7}$$

$$\mathcal{L}_{adv}(G) = -\mathbb{E}_{z,c}[D(G(z,c))], \tag{8}$$

and $\mathcal{L}_{cls}$ as:

$$\mathcal{L}_{cls}(D) = \mathbb{E}_{z,c}[-\log D(c \mid z)], \tag{9}$$

$$\mathcal{L}_{cls}(G) = \mathbb{E}_{z,c}[-\log D(c \mid G(z,c))], \tag{10}$$

where z is the input image, c is the style label, and $\hat{z}$ is sampled uniformly along straight lines between a pair of real and generated images. We define the $\mathcal{L}_{rec}$ for G_R as:

$$\mathcal{L}_{rec}(G_R) = \mathbb{E}_{x,y,c}[\parallel x - G_R(y,c) \parallel_1]. \tag{11}$$

Here, y is the input skeleton and x the target CH-Char.

Therefore, the total loss functions for G_R and D_R are:

$$\mathcal{L}(G_R) = \alpha^r \mathcal{L}_{adv}(G_R) + \beta^r \mathcal{L}_{cls}(G_R) + \lambda^r \mathcal{L}_{rec}(G_R), \tag{12}$$

$$\mathcal{L}(D_R) = \alpha^r \mathcal{L}_{adv}(D_R^{adv}) + \beta^r \mathcal{L}_{cls}(D_R^{cls}). \tag{13}$$

Skeleton Transformation Network (TNet)

TNet is designed for transforming skeleton structure to the particular style by learning from the unpaired multi-chirography style sets. Since it is impossible to learn the mapping between different stroke styles directly from the training pairs, our TNet needs to extract the content features of skeleton first, and then restore them conditioned on specified style label. The transformed skeleton ought to keep the consistency of the contents.

To achieve better extraction effect of content feature, we adjust the conditional GAN by concatenating the style information with content features rather than input images, as shown in the middle of Fig.2. Given the input skeleton y, we utilize the encoder E to extract the content features $E(y)$, and then use the style decoder H to map the 1D one-hot style vector c to the 2D feature maps $H(c)$ with the same size as $E(y)$. After that, we employ the decoder F to reconstruct the $E(y)$ conditioned on $H(c)$ to the corresponding skeleton $G_T(y,c) \equiv F(E(y), H(c))$, where G_T is generator in our cGAN model. In this way, the encoder E is more focused on learning the content of the skeleton.

In G_T, the E contains a downsampling module and three ResBlocks, while the H is stacked by three DeconvBlocks, and the F is composed of three ResBlocks and a upsampling module. The D_T is the same as D_R, including D_T^{adv} to distinguish true from false and D_T^{cls} to classify.

With the purpose of enabling E to extract the complete content attributes and F to reconstruct the stylized skeleton with unpaired training data, we apply the training process of Stage-I in Fig.4 to TNet. First, we input the skeleton image y in style $s \in C$ and the target label $t \in C$ into G_T to generate the $\hat{y} = G_T(y,t)$, and constrain the stylistic correctness of $\hat{y}$ with $\mathcal{L}_{adv}$ and $\mathcal{L}_{cls}$. Next, $\hat{y}$ and s are inputted into the network to obtain the restored $y' = G_T(\hat{y}, s)$. The content attributes consistency of input and output is constrained by cycle consistency loss $\mathcal{L}_{cc}$, which signifies enforcing $y' \approx y$. Finally, based on $\mathcal{L}_{cc}$, we utilize the L_1 loss function on $E(y)$ and $E(\hat{y})$ as content loss $\mathcal{L}_{cont}$ for further improving the capability of E and enforcing the content consistency of input and output in feature-level. For G_T, we define the $\mathcal{L}_{cc}$ as:

$$\mathcal{L}_{cc}(G_T) = \mathbb{E}_{y,c}[\parallel y - G_T(G_T(y,t),s) \parallel_1], \tag{14}$$

and the $\mathcal{L}_{cont}$ as:

$$\mathcal{L}_{cont}(G_T) = \mathbb{E}_{y,c}[\parallel E(y) - E(G_T(y,t)) \parallel_1]. \tag{15}$$

Thus the total loss functions for G_T and D_T in Stage-I are written as:

$$\begin{aligned}\mathcal{L}(G_T) =& \alpha^t \mathcal{L}_{adv}(G_T) + \beta^t \mathcal{L}_{cls}(G_T) \\ &+ \lambda^t \mathcal{L}_{rec}(G_T) + \gamma^t \mathcal{L}_{cont}(G_T),\end{aligned} \tag{16}$$

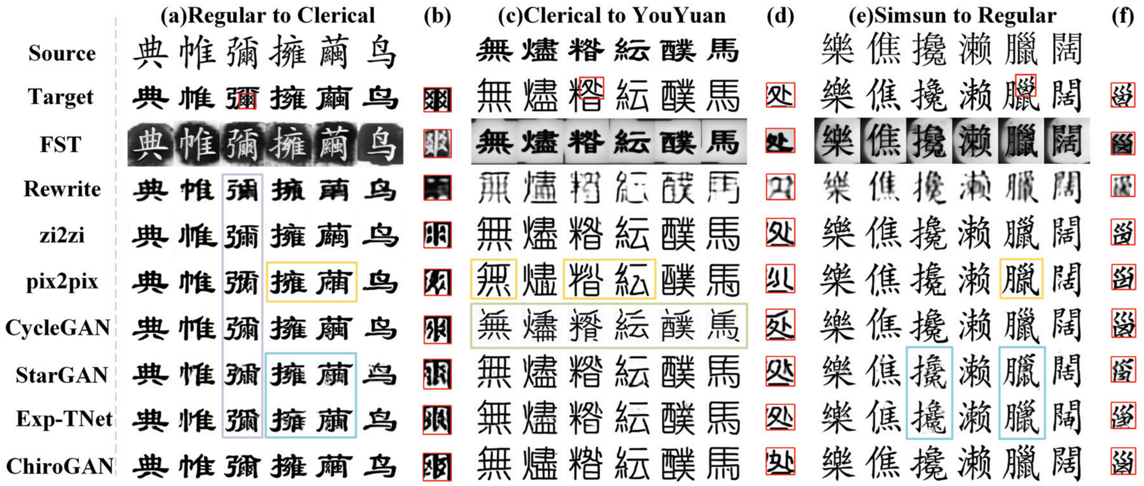


Figure 5: Comparison of the baselines and our method. The yellow box means the error stroke, the purple box indicates the missing detail, the green box shows incorrect translation, and the blue box represents the blurred result. (b), (d) and (e) are respectively the close-up views of the part in the red boxes of (a), (c), and (e).

$$\mathcal{L}(D_T) = \alpha^t \mathcal{L}_{adv}(D_T^{adv}) + \beta^t \mathcal{L}_{cls}(D_T^{cls}). \tag{17}$$

In stage-I, TNet can preliminarily learn the skeleton transformation between different chirography style.

Joint Training

The pre-trained TNet of Stage-I offen causes the ghosting artifact, intermittent skeletons or missing strokes. If we directly render these results by RNet, the generated CH-Chars images will contain obvious flaws such as incorrect strokes and noise patchs. The principal reason is the insufficient guiding information for training, so it is difficult to learn the mapping relationship between skeletons, which is a common problem in unpaired image translation.

To address this problem, we propose a low-cost solution that leveraging the pre-trained RNet to refine TNet, called Stage-II(see Fig.4). Aiming to reduce ghosting artifacts, blur and excess strokes in the generated skeletons, we jointly fine-tune the G_T, D_T and D_R together while frozing the G_R. We should enforce not only the generated skeleton $\hat{y}$ but also its rendered result $\hat{x} = G_R(\hat{y}, t)$ to be realistic and well-classified through the adversarial training between G_T and both D_T and D_R. Meanwhile, with the help of G_R, we enforce the rendered result $x' = G_R(y', s)$ of the cycle reconstructed skeleton y' to be the same as the ground truth y by optimizing the joint cycle reconstruction loss $\mathcal{L}_{rec}^j$.

As illustrated in Fig.4, both Stage-I and Stage-II are both executed for the joint optimization of TNet and RNet simultaneously to further constrain the training of G_T in terms of content and style. In joint training, the loss functions for D_T and D_R are still the same as $\mathcal{L}_{adv}(D_T)$ and $\mathcal{L}_{adv}(D_R)$ in pre-training. For G_T, the total loss functions is

$$\begin{aligned}\mathcal{L}_j(G_T) =& \mathcal{L}(G_T) + \alpha^r \mathcal{L}_{adv}^j(G_T) \\ &+ \beta^r \mathcal{L}_{cls}^j(G_T) + \lambda^r \mathcal{L}_{rec}^j(G_T),\end{aligned} \tag{18}$$

where $\mathcal{L}_{rec}^j = \mathbb{E}_{x,y',s}[\| \ x - G_R(y', s) \ \|_1]$, $\mathcal{L}_{adv}^j = -\mathbb{E}_{\hat{x},t}[D_R(\hat{x})]$ and $\mathcal{L}_{cls}^j = \mathbb{E}_{\hat{x},t}[-\log D_R(t \mid \hat{x})]$.

Experiments

Dataset & Implementation Details

We establish the standard font dataset with detailed annotation for the comparison, called StdFont-4, and the real calligraphy dataset with only author information for the perception experiments, called Calli-5. For StdFont-4, we utilize the expert-designed font libraries, including regular script, clerical script, Simsun, and YouYuan, to create about 6,700 CH-Char images each font based on the GB2312. For Calli-5, we collect about 1,200 famous calligraphers' digital works for each style, including regular script of Ouyang Xun and Yan Zhenqing, clerical script of Deng Shiru and Cao Quan tablet and semi-cursive script of Zhao Mengfu.

In our experiment, the input and output are 128 × 128 grayscale. During pre-training, α^r, β^r and λ^r of RNet are set to 1, 1 and 100, while α^t, β^t, λ^t, γ^t of TNet are set to 1, 1, 10 and 5, respectively. During joint training, we reset λ^t, β^t, and λ^r as 20, 2 and 10, respectively, and reduce learning rates by a factor of 10. We employ the history pool(Shrivastava et al. 2017) to improve quality.

Baseline Models

Six existing methods are chosen as baselines to compare with our method, including fast style transfer (FST)(Johnson, Alahi, and Fei-Fei 2016), Rewrite(Tian 2016), zi2zi(Tian 2017), Pix2Pix, CycleGAN and StarGAN(Choi et al. 2018). Among them, the former 5 methods are proposed for the translation between two domains, while StarGAN and our method are for the multi-domain. Therefore, we select 3 typical tasks from StdFont-4 for comparison, including Simsun-to-Regular similar skeleton (SK) but different stroke style (SS), Regular-to-Clerical with dissimilar SK but semblable SS, and clerical-to-YouYuan with large difference in both SK and SS. In addition, Rewrite, zi2zi and Pix2Pix require image pairs for training. We find the corresponding image pairs from StdFont-4, and build the pairs

Table 1: Quantitative evaluations and the average score(AVG Score) of test-I in user study for our method and baselines. R2C, C2Y and S2R represent the Regular2Clerical, Clerical2YouYuan and Simsun2Regular, respectively. MS ACC stands for the percentage classification accuracy of InceptionV3 pre-trained on StdFont-4. The ground truth is described as GT.

Method	Style Error Rate(%)				Content Error Rate(%)				IOU			MS	AVG
	R2C	C2Y	S2R	Total	R2C	C2Y	S2R	Total	R2C	C2Y	S2R	ACC	Score
GT	0.00	0.07	0.04	0.04	2.70	3.32	2.93	2.98	0.188	0.202	0.191	–	–
Rewrite	52.91	99.42	0.56	50.60	0.14	35.92	1.38	12.33	0.189	0.021	0.014	–	0.04
zi2zi	**0.07**	**0.25**	**0.04**	**0.12**	**0.28**	**1.91**	**0.21**	**0.80**	**0.482**	**0.359**	**0.266**	–	**2.84**
PixPix	0.00	0.25	0.11	0.12	0.57	1.01	0.32	0.63	0.548	0.349	0.331	–	1.48
CycleGAN	0.07	0.29	0.07	0.14	0.00	0.00	0.04	0.01	0.499	0.124	0.307	–	1.76
StarGAN	0.07	0.00	9.91	3.36	0.43	1.30	0.56	0.76	0.386	0.326	0.243	85.46	0.92
Exp-TNet	0.07	0.04	3.39	1.18	0.43	1.33	0.88	0.88	0.451	0.344	0.260	94.15	0.82
ChiroGAN	**0.00**	**0.00**	**0.42**	**0.14**	**0.11**	**0.87**	**0.00**	**0.32**	**0.477**	**0.346**	**0.278**	99.72	**2.13**

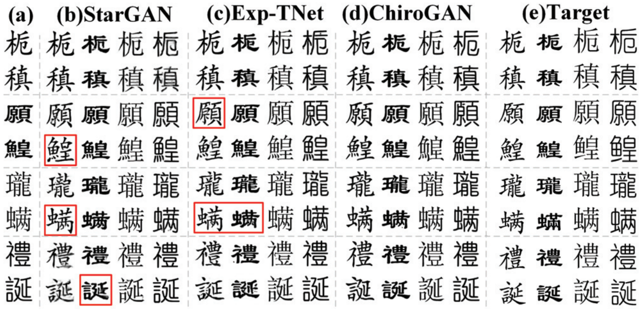


Figure 6: Comparison of the StarGAN, Exp-TNet and ChiroGAN on StdFont-4. The red box represents poor results.

dataset StdFontPair. Moreover, We train CycleGAN on StdFontPair without the paired labels, and train multi-domain methods on StdFont-4.

Qualitative Evaluation

We compare the baselines with our two scenarios, including the three-stage architecture ChiroGAN, and the Exp-TNet which directly employs TNet to translate the CH-Char from source to target style like CalliGAN(Gao and Wu 2019).

Comparison for Two Chirographies Translation. The qualitative results are presented in Fig.5, visibly demonstrating the validity of our unpaired method with realistic and clear results as zi2zi. More specifically, Fig.5b, d, f show the close-up views of the details, proving that our method can generate finer stroke details than others.

The results of FST reveal that the pre-trained CNN is incapable of learning the geometric style of CH-Char.Although Pix2Pix seems to transfer the style well while preserving the content, the results often contain erroneous and missing strokes in terms of complex tasks and high-density strokes (yellow boxes). Besides, CycleGAN produces poor-quality results for the tasks with great differences like Clerical2YouYuan (green box). The generated strokes appear to be directly modified on the source skeleton without transformation, and the unreasonable strokes are added to the top and bottom to make the overall structure look like the target style. Furthermore, StarGAN and Exp-TNet often cause ghosting artifacts and blur (blue boxes).

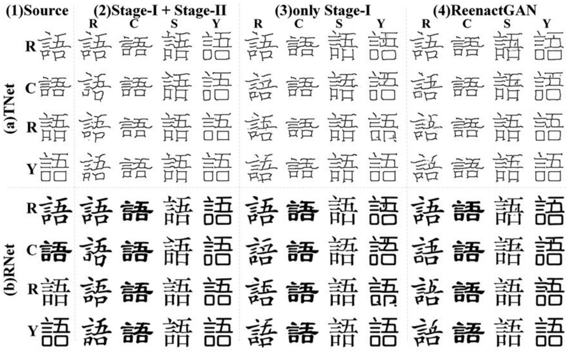


Figure 7: Comparison of ChiroGAN, ChiroGAN without joint training, and ReenactGAN on StdFont-4.

Comparison for Multi-chirography Translation. In Fig.6b, StarGAN causes fuzzy strokes. Comparing with it, the superiority of our method in unpaired multi-chirography translation is demonstrated, which produces clearer and more stylized results with precise content(see Fig.6d).

Quantitative Evaluation

In qualitative experiments, we calculate the SER(style error rate), CER(content error rate) and IoU(the intersection of black pixels between the generated image and ground truth over their union) for each method. The SER refers to the misclassified rate by InceptionV3(Szegedy et al. 2016) pre-trained in StdFont-4. And the CER represents the rate of the original and generated images that are recognized as different CH-Chars by InceptionV3. The InceptionV3 for CER is trained on the 10,000 matched and 10,000 unmatched pairs collected from StdFont-4. As shown in Table1, our method achieves the low SER, CER and high IoU as the supervised method and far exceeds other unpaired methods. Compared with CycleGAN, ChiroGAN is more stable on various tasks, which learns the commonality of multiple chirographies. In addition, we employ pre-trained InceptionV3 on StdFont-4 to classify all results of StarGAN, Exp-TNet and ChiroGAN, while ChiroGAN achieves the best performance.

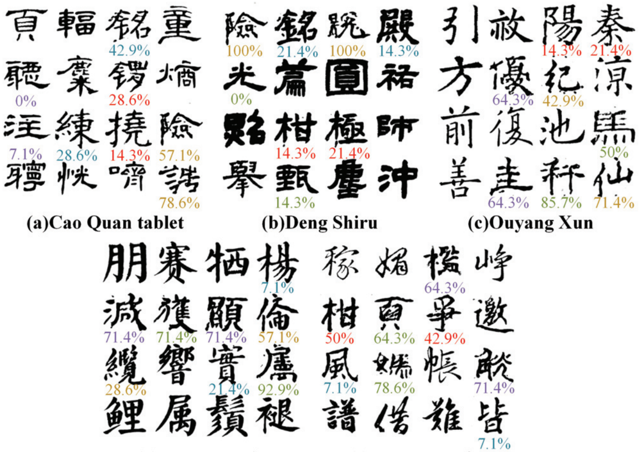


Figure 8: The examples in test-II. The green, purple, blue, red and yellow respectively represent the source style of the results as Cao Quan tablet, Deng Shiru, Ouyang Xun, Yan Zhenqing and Zhao Meng. The number is the percentage of results that users consider to be unrealistic, while others are real calligraphy.

User Study

We also conduct the user study to compare the results with baselines and to test the realism and style-similarity of ChiroGAN on Calli-5 dataset. In test-I, we randomly choose 3 samples from results in StdFontPair for each task and each method. Users are asked to select and sort the top 4 results according to the ground truths. For the top-4 results and the rest, we assign 4, 3, 2, 1 and 0 points, respectively. In test-II, we randomly select 8 real calligraphy from each style in Calli-5 and mix them with 8 generated ones to build the test matrix(as shown in Fig.8). The users are firstly presented with some real samples, and then are required to choose the unrealistic ones. Finally, we calculate the average score for test-I and the accuracy for the test-II respectively. We design several groups of experiments according to the above rules.

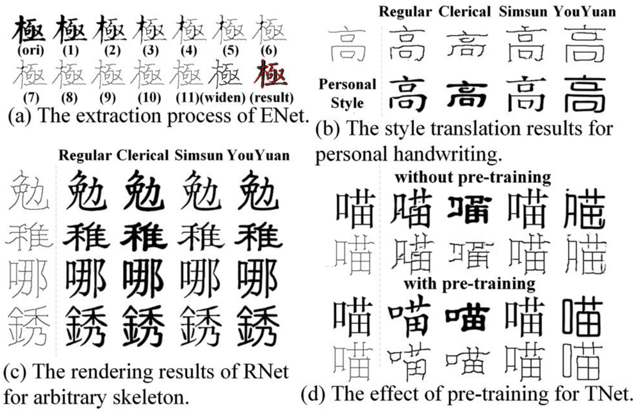
For test-I and test-II, 42 and 29 people familiar with Chinese are involved, including 13 calligraphers. Our method achieve a high score close to zi2zi(see Table1) in test-I. Furthermore, the accuracy in Fig.2 is lower than random selection(50%), proving the difficulty for users to distinguish the generated results from the ground truth.

Table 2: The accuracy of test-II in user study.

Style	Ou	Yan	Deng	Cao	Zhao	Total
ACC(%)	45.7	47	35.8	31.5	42.7	40.5

Analysis of Each Component

Effect of TNet. By comparing the performance of our two scenarios in Fig.6c, d, we prove the superiority of our three-stage architecture, which implements character-to-skeleton by ENet, skeleton-to-skeleton by TNet and skeleton-to-character by RNet. Exp-TNet directly implements character-to-character, leading to the fuzzy details similar to StarGAN.



(a) The extraction process of ENet.

(b) The style translation results for personal handwriting.

(c) The rendering results of RNet for arbitrary skeleton.

(d) The effect of pre-training for TNet.

Figure 9: Analysis of all the components in our method.

Effect of ENet. Fig.9a shows the extraction process of ENet. The structure and content of the example can be well represented by the skeleton extracted in 11 steps. We also reproduce the original algorithm to test the efficiency, which takes about 10 seconds to extract the skeleton for one image on the Intel Xeon e5-2640 v4 CPU. However, using ENet, it only takes 0.02 seconds on the NVIDIA TITAN Xp GPU. Moreover, ENet only requires 0.27 seconds and 800M memory for 1,000 images simultaneously.

Effect of RNet. When training RNet, we add label information to achieve multi-style translation. To verify it, arbitrary skeletons are rendered in various styles by RNet. Fig.9c manifests that RNet has the ability to render any skeleton in the specified stroke style according to the label.

Effect of Joint Training. In Fig.7(3), we use the TNet only trained on Stage-I and pre-trained RNet for comparison. We also try the two-phase method ReenactGAN(Wu et al. 2018) on our task in Fig.7(4). It separately trains boundary transformer (like TNet) and encoder-decoder (like ENet-RNet), and utilizes a PCA shape loss to constrain content consistency instead of our content loss $\mathcal{L}_{cont}$.

Fig.7(2) shows our demonstration of the noticeable improvement in TNet from joint training. There are many incorrect strokes in the generated skeleton without joint training, being likely to affect the realism.

Effect of Content Loss. In order to verify whether content loss $\mathcal{L}_{cont}$ can guide the encoder E to capture the content features, we use the handwriting in personal style as input to observe the results. Fig.9b indicates that handwriting is successfully transferred to the target style, preserving the content consistency, which proves the validity of $\mathcal{L}_{cont}$.

Effect of the Pre-training for TNet. As shown in Fig.9d, without pre-training, although TNet learns the features of the overall structure, the results are far from expectation. Therefore, the pre-training for TNet plays a significant role in learning the skeleton details.

Conclusion

In this paper, we propose ChiroGAN, a novel three-stage framework for multi-chirography Chinese character translation with unpaired training data. First, we improve a skeletonization algorithm (ENet) to efficiently extract the skeleton. Then, the novel stacked cGAN based on skeleton transformation (TNet) and stroke rendering (RNet) is employed to transfer the skeleton structure and stroke style for the input. To compensate for the lack of labeled information in training data, we propose a low-cost solution by training TNet jointly with the pre-trained RNet. The experiments demonstrate that the performance of our method in visual perceptions and quantitative evaluations is comparable to the supervised baseline methods and superior to other unpaired baseline methods.

Acknowledgments

The work is supported by the National Nature Science Foundation of China (No.61379073), the Fundamental Research Funds for the Central Universities (2018FZA5012), the Engineering Research Center of Digital Library, Ministry of Education, the China Research Centre on Data and Knowledge for Engineering Sciences and Technology and the Chinese Knowledge Center for Engineering Sciences and Technology (CKCEST).

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 214–223.

Azadi, S.; Fisher, M.; Kim, V. G.; Wang, Z.; Shechtman, E.; and Darrell, T. 2018. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7564–7573.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2018. Everybody dance now. *Rippleffect Studio Limited*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8789–8797.

Du, X.; Wu, J.; and Xia, Y. 2016. Bayesian relevance feedback based chinese calligraphy character synthesis. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Gao, Y., and Wu, J. 2019. *CalliGAN: Unpaired Mutli-chirography Chinese Calligraphy Image Translation*. 334–348.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Bing, X.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 5767–5777.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jian-ping, W.; Zi-tuo, Q.; Jin-ling, W.; and Guo-jun, L. 2005. Chinese characters stroke thinning and extraction based on mathematical morphology [j]. *Journal of Hefei University of Technology (Natural Science)* 11:017.

Jiang, Y.; Lian, Z.; Tang, Y.; and Xiao, J. 2019. Scfont: Structure-guided chinese font generation via deep stacked networks.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

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

Lyu, P.; Bai, X.; Yao, C.; Zhu, Z.; Huang, T.; and Liu, W. 2017. Auto-encoder guided gan for chinese calligraphy synthesis. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, 1095–1100. IEEE.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; and Webb, R. 2017. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, 6.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Tian, Y. 2016. Rewrite: Neural style transfer for chinese fonts. Website. http://github.com/kaonashi-tyc/Rewrite.

Tian, Y. 2017. zi2zi: Master chinese calligraphy with conditional adversarial networks. Website. http://github.com/kaonashi-tyc/zi2zi.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wong, H. T., and Ip, H. H. 2000. Virtual brush: a model-based synthesis of chinese calligraphy. *Computers & Graphics* 24(1):99–113.

Wu, Y.; Zhuang, Y.; Pan, Y.; and Wu, J. 2006. Web based chinese calligraphy learning with 3-d visualization method. In *2006 IEEE International Conference on Multimedia and Expo*, 2073–2076. IEEE.

Wu, W.; Zhang, Y.; Li, C.; Qian, C.; and Loy, C. C. 2018. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*.

Xu, S.; Lau, F. C.; Cheung, W. K.; and Pan, Y. 2005. Automatic generation of artistic chinese calligraphy. *IEEE Intelligent Systems* 20(3):32–39.

Yu, J., and Peng, Q. 2005. Realistic synthesis of cao shu of chinese calligraphy. *Computers & Graphics* 29(1):145–153.

Zhang, Z.; Wu, J.; and Yu, K. 2010. Chinese calligraphy specific style rendering system. In *Proceedings of the 10th annual joint conference on Digital libraries*, 99–108. ACM.

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*, 2223–2232.