Learning to Compose Stylistic Calligraphy Artwork with Emotions

Shaozu Yuan[*]
yuanshaozu@jd.com
JD AI, Beijing, China

Ruixue Liu[*]
liuruixue@jd.com
JD AI, Beijing, China

Meng Chen
chenmeng20@jd.com
JD AI, Beijing, China

Baoyang Chen
chenbaoyang@cafa.edu.cn
CAFA, Beijing, China

Zhijie Qiu
qiuzhijie@cafa.edu.cn
CAFA, Beijing, China

Xiaodong He
xiaodong.he@jd.com
JD AI, Beijing, China

ABSTRACT

Emotion plays a critical role in calligraphy composition, which makes the calligraphy artwork impressive and have a soul. However, previous research on calligraphy generation all neglected the emotion as a major contributor to the artistry of calligraphy. Such defects prevent them from generating **aesthetic**, **stylistic**, and **diverse** calligraphy artworks, but only static handwriting font library instead. To address this problem, we propose a novel cross-modal approach to generate stylistic and diverse Chinese calligraphy artwork driven by different emotions automatically. We firstly detect the emotions in the text by a classifier, then generate the emotional Chinese character images via a novel modified Generative Adversarial Network (GAN) structure, finally we predict the layout for all character images with a recurrent neural network. We also collect a large-scale stylistic Chinese calligraphy image dataset with rich emotions. Experimental results demonstrate that our model outperforms all baseline image translation models significantly for different emotional styles in terms of content accuracy and style discrepancy. Besides, our layout algorithm can also learn the patterns and habits of calligrapher, and makes the generated calligraphy more artistic. To the best of our knowledge, we are the first to work on emotion-driven discourse-level Chinese calligraphy artwork composition.

CCS CONCEPTS

• **Applied computing** → **Fine arts**; • **Computing methodologies** → **Image representations**.

KEYWORDS

Calligraphy Generation, AI and Art, Generative Adversarial Network, Calligraphy Dataset

ACM Reference Format:
Shaozu Yuan, Ruixue Liu, Meng Chen, Baoyang Chen, Zhijie Qiu, and Xiaodong He. 2021. Learning to Compose Stylistic Calligraphy Artwork with Emotions. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475711

1 INTRODUCTION

Calligraphy (the art of writing) plays a significant role in Asian culture and art. It is regarded as one of the four basic skills of ancient Chinese literati, along with playing stringed musical instruments, the board game "Go", and painting. Calligraphy is celebrated not only as an artistic expression of writing, but also an outward expression of the artist's inner psychology and self-cultivation. Emotions and the various styles of calligraphy are closely related[1]. The psychological impulse of calligrapher affects not only the changing stroke movement of character but also the layout of whole artwork, and finally contributes to the artistry of the calligraphy [25]. It's crucial to take emotion status into consideration when exploring the automatic synthesis of calligraphy.

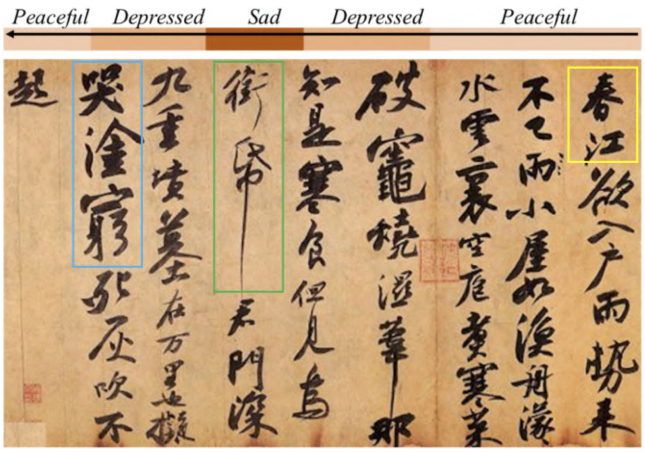

Figure 1: A human-created Chinese calligraphy artwork. The darkness color bar on top indicates the intensity of sadness emotion when writing this artwork.

As shown in Figure 1[2], this poem describes the melancholy and loneliness of the author when he was demoted. The calligraphy

[*] Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '21, October 20–24, 2021, Virtual Event, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10...$15.00
https://doi.org/10.1145/3474085.3475711

[1] The research on emotion and Chinese calligraphy style can date back to Yuan Dynasty (14th century). Yizeng Cheng, a calligrapher in Yuan Dynasty, wrote a famous book named *Essential Precepts of the Hanlin Academy*. In this book, he analyzed and concluded different emotions can result in stylistic variations for calligraphy creation.
[2] https://www.comuseum.com/calligraphy/masters/su-shi/cold-food-observance/

composition was driven by the transition of emotions. At the beginning of the poem, the artist was in peaceful mood when he was describing the windy weather and river view. Characters show clarity and are in a steady manner such as the characters in the yellow box. Gradually, the depressed emotion brews up when he was memorizing the dead people. The outburst of sadness is reflected by the content in the green box, where characters flow together into a continuous movement and stand out against the paper. Later depressed state came back when he described the desperate reality and his demotion. And strokes of characters in the blue box become thick and plump. Moreover, the layout features of all the characters vary rhythmically according to the emotional status. The character size gets larger when the sadness emotion becomes stronger. And the line spacing and character spacing also range considerably.

Previous research on Chinese calligraphy generation all neglected the importance of emotion. The early works focus on the reconstruction of strokes extracted from the character [34, 35]. With the popularity of Generative Adversarial Networks (GANs) [9], there are some attempts made to generate calligraphy images with pix2pix [30], variational auto-encoder (VAE) GANs [18] and CycleGAN [1]. Recently, some researchers propose to integrate the character structure or skeleton information into Chinese font generation [8, 13]. Although they have achieved significant progress in the quality of generated images, those researches only treat the Chinese calligraphy generation as a general-purpose style transfer task and overlook the artistry and completeness of it as a piece of artwork. None of them consider generating emotional calligraphy, and they only generate single character images and have no attempts to arrange characters into a complete piece of calligraphic artwork. Such defects cause they can not produce **aesthetic**, **stylistic**, and **diverse** calligraphy artworks, but only generate static handwriting font library instead.

In this paper, we are trying to challenge the problem of stylistic Chinese calligraphy artwork composition driven by emotions. Firstly, we train a classifier to detect the emotion for each sentence in the input text. Secondly, we propose a novel Emotional Calligraphy Generative Adversarial Network (ECA-GAN), equipped with specific auxiliary classifiers and discriminators to enhance both the **style control** and **content preservation** of generation. To facilitate one-to-many style transfer, a novel conditional layer instance normalization is devised. Besides, to encourage diverse generation, we design a new diversity loss to create diverse stochastic variations of images. Thirdly, to construct a discourse-level calligraphy artwork, we define several layout features for characters involved in calligraphy. Then we combine both linguistic and visual information of character images with Recurrent Neural Network (RNN) to predict the layout features. Due to the lack of off-the-shelf emotional Chinese calligraphy image dataset, we collect a large-scale Chinese calligraphy image dataset with rich emotions. To the best of our knowledge, this is the first work to focus on emotion-driven discourse-level artistic Chinese calligraphy composition.

Our main contributions can be summarized as follows: 1) We are the first to endow the calligraphy composition with the ability to express fickle emotions, which improves the artistry and diversity of the generated calligraphy images significantly; 2) We propose a novel multi-domain GAN structure (ECA-GAN), which is specifically designed for calligraphy synthesis task with rich emotional styles. The evaluation results on both content accuracy and style discrepancy show its effectiveness; 3) We propose a novel Layout Net and make the first attempt to realize the discourse-level calligraphy artwork composition. The qualitative analysis indicates the artistic advantage of our proposed method; 4) We construct a large-scale Chinese calligraphy image dataset and will release it to the research community soon.

2 RELATED WORK

Image-to-image translation (I2I) is a recent hot topic which aims to translate images from one domain either to another (one-to-one) [11, 14, 41] or to others (one-to-many) [4, 10, 16, 42]. Based on the generative adversarial networks (GANs) [9, 21], Isola et al. [11] propose a generic image-to-image translation algorithm (pix2pix) to handle I2I. BicycleGAN [42] combines both cVAE-GAN and cLR-GAN to encourage a bijective mapping between the latent and output spaces. To get rid of the constraint of paired data in pix2pix, CycleGAN [41] and DualGAN [36] utilize the cycle consistency to improve the training stability. UNIT [17] assumes a shared latent space for two image domains. It achieves unsupervised translation by learning the bijection between latent and image spaces using two generators. U-GAT-IT [14] applies a new attention module and a new learnable normalization function to flexibly control the amount of change in shape and texture. MUNIT [10], DRIT [16], and DMIT [38] make one-to-many mapping possible by decomposing the image into content part that is domain-invariant and style part that captures domain-specific properties. StarGAN [4] introduces a domain classifier and classification loss to achieve multi-domain translation in a single model.

Chinese character generation is a long studied problem [23, 32, 37]. Most previous works represent Chinese characters based on hierarchical encoding of simple strokes [22, 34, 35, 43]. With the boom of deep learning and neural networks, some researchers model this task as style transfer problem [2, 12, 27]. Zi2zi [29, 30] considers each Chinese character as a whole and learns to transform between fonts with paired training data. As it is difficult to obtain a large set of paired training examples, unsupervised character translation methods become popular [1, 7]. EMD [39] and SA-VAE [26] separate content and style as two irrelevant domains and use two independent encoders to model them. Recent works of [8, 13] divide the task into several subtasks including skeleton transformation and stroke rendering, which require multi-stage model architecture. CalliGAN [33] proposes to integrate the character structure information by aggregating the stroke and radical embeddings with LSTM, which is difficult to catch the high-level structure of character.

The differences between above works and ours are: 1) Existing Chinese calligraphy generation studies focus on the changes among different fonts, while we consider the emotion factor, which is the major contributor to the artistry of Chinese calligraphy; 2) We try to analyze the layout of an entire artwork and generate discourse-level pieces instead of single character images; 3) We adopt a new attention module to integrate the structure information of character as prior domain knowledge via auxiliary classifiers while previous works simply aggregate the stroke and radical embeddings by order. Compared with them, our method can learn better generalized content representation of character images.

3 APPROACH

As shown in Figure 3, the input of the task is plain text, which contains several sentences from a classic Chinese article, and the output is a complete calligraphy image. Our framework consists of three modules: 1) an emotion classifier based on BERT [5] to detect emotions in text; 2) an ECA-GAN network to transfer an existing standard font image to a specific emotional calligraphy image; 3) a RNN-based layout prediction network, to synthesize the layout with both textual and visual features.

3.1 Emotion Detection

To recognize the emotion from input, we adopt the state-of-the-art pre-trained model BERT [5] for emotion classification. Each line is analyzed and the emotion transition pattern is obtained by the classifier. As the input text of Chinese calligraphy is usually from classical Chinese articles, we fine-tune the model with a manually-labelled fine-grained sentiment poetry corpus [3], which covers 3 emotion classes, namely peaceful, happy, and sad. The classification accuracy is 65.5% in our experiment. (Note that [3] reports a three-class classification accuracy of 63.7% on the same dataset.)

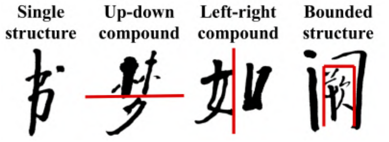

Figure 2: Basic structures of Chinese characters.

3.2 Image Generation

3.2.1 Model. As it is difficult and expensive to collect adequate training pairs for all emotional styles, we formulate the emotional Chinese calligraphy generation as an unsupervised image-to-image translation task. The objective is to learn a mapping function from the source domain to the target domain with huge unpaired samples drawn from each domain. Here, we propose a novel GAN-based model (denoted as ECA-GAN) to learn various emotional calligraphy styles. The encoder of the generator is composed of two convolution layers with the stride size of two for down-sampling and four residual blocks. The decoder of the generator consists of four residual blocks and two up-sampling convolution layers with the stride size of one. Two different scales of PatchGAN [11] are employed for the discriminator network, which classifies whether local and global image patches are real or fake. Different from previous style transfer tasks (e.g., photo2vangogh and photo2portrait), the target fonts usually have large shape change compared with source fonts. Therefore, both the **style control** and **content preservation** are crucial for our task. Style control makes the generated images stylized, while content preservation guarantees the generated images recognisable. To enhance content preservation, we propose to integrate the character knowledge into the model by considering the structure of Chinese characters. As Figure 2 shows, there are four basic structures including single structure, up-down compound, left-right compound, and bounded structure, which are shared among all the Chinese characters [28]. Inspired by Class Activation Map (CAM) [40], we design an auxiliary classifier η_c to classify the input image into four structure categories in the generator to learn the **content-aware representation** $a_c(x)$ of image. The auxiliary classifier is trained to learn the importance weights of all feature maps, then we calculate a set of domain specific attention feature maps and guide the generator to focus more on the structure region of characters. Meanwhile, we also design a structure discriminator D_c to help the model learn the content of characters better. D_c learns to classify the real image and generated image into the four structure categories. Similarly, to enhance style control, we design another auxiliary classifier η_s to learn the **style-aware representation** $a_s(x)$ of image. The task for η_s is to classify the input image into either source or target style. We also devise an extra style discriminator D_s to distinguish the difference between multiple emotional styles. We put the source domain images and three target domain images together to form a mixed-style data. Then we force the style discriminator D_s to classify the different emotional styles between the generated images and images from mixed-style data. Finally, we fuse the content-aware and style-aware representations by element-wise multiplication (addition and concatenation are also tried but less effective), and feed it into decoder for generation.

In order to equip ECA-GAN the ability to integrate multi-styles in a single model, we proposed a novel CLIN (Conditional Layer Instance Normalization) to guide the model to generate specific stylized character images based on the predicted emotion category from BERT. Inspired by CIN (conditional instance normalization) [6], scaling and shifting parameters are sufficiently specialized for an affine transformation after normalization for each style, while all convolutional weights of a style transfer model can be shared across different styles. Meanwhile, the AdaLIN [14] function, by adaptively selecting a proper ratio between Instance Normalization (IN) and Layer Normalization (LN), helps the attention-guided model to flexibly control the amount of change in shape and style. Therefore, CLIN combines the advantages of CIN and AdaLIN. The equation is as follows:

$$CLIN(x, \mu_e, \sigma_e) = \mu_e(\rho LN(x) + (1 - \rho)IN(x)) + \sigma_e \tag{1}$$

where μ_e, σ_e are emotion-dependent embedding vectors, and e indicates the emotion label. And IN represents channel-wise normalization and LN represents layer-wise normalization respectively, they are balanced by a parameter ρ, which is learned from datasets during training time.

Besides, considering diversity is a basic requirement of real calligrapher's artwork composition. Inspired by [31], we propose to introduce a noise tenor z in the input of the generator and design a new diversity loss that encourages to create stochastic variations for the same character image.

3.2.2 Training Objective. Our objective contains five parts: adversarial loss, diversity loss, cycle consistency loss, structure loss and the CAM loss. Let's set x and y as samples from source images and target images, z is a noise vector for the input, e is emotion label, and G_F, G_B, D_{bi}, D_c, D_s represent forward generator, backward generator, binary discriminator, structure discriminator and style discriminator respectively. G_F and G_B share the same network.

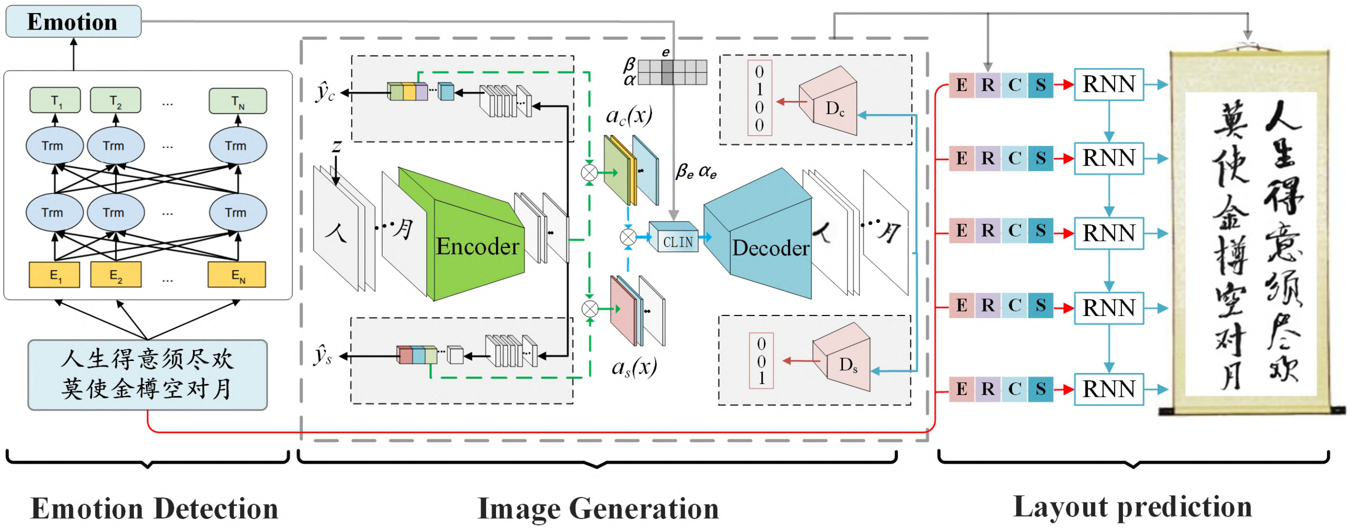

Figure 3: Overview of our emotional calligraphy generation framework.

The adversarial loss [9], making the generated images $G_F(x,z,e)$ indistinguishable from the real images y, can be written as:

$$\mathcal{L}_{adv}(G_F, D_{bi}) = \mathbb{E}_y[logD_{bi}(y)] + \mathbb{E}_x[log(1 - D_{bi}(G_F(x,z,e)))] \tag{2}$$

To encourage the diversity of generation, we design a diversity loss to measure the mean of the distance among output images and use reciprocal to maximize it.

$$\mathcal{L}_{div}(G_F) = \frac{1}{K}\sum_{i=1}^{K}\frac{1}{\mathbb{E}_x||G_F(x,z_i,e) - G_F(x,z_j,e)||_1 + \varepsilon} \tag{3}$$

where z_i and z_j are different noise vectors during training ($i \neq j$), K is the number of input noises. ε is a hyper-parameter to prevent division by zero.

To ensure that the cycle transformation is able to bring the image back to the original state, the cycle consistency loss [41] is defined as:

$$\mathcal{L}_{cycle}(G_F, G_B) = \mathbb{E}_x[||G_B(G_F(x,z,e),z',e') - x||_1] \tag{4}$$

To improve the generation quality, a structure loss $\mathcal{L}_c$ is proposed to facilitate training. We design a structure discriminator D_c to distinguish whether the generated sample $G_F(x,z,e)$ has the correct character structure y_c.

$$\mathcal{L}_c(G_F, D_c) = -\left(\mathbb{E}_x[log(D_c(y_c|G_F(x,z,e))]]\right) \tag{5}$$

where $D_c(y_c|G_F(x,z,e))$ represents a probability distribution over its corresponding four basic structures.

We also bring CAM loss to promote the style and the content transformation:

$$\mathcal{L}_{cam}(\eta_c, \eta_s) = -(\mathbb{E}_x[log(\eta_c(y_c|x'))] + \mathbb{E}_x[log(\eta_s(y_s|x'))]]) \tag{6}$$

where η_c, η_s are two auxiliary classifiers and x' denotes the encoded feature maps of image. They are forced to predict the correct category of the style y_s and the structure y_c, for exploiting distinct regions of images during image translation.

Apart from above loss, a style loss $\mathcal{L}_s$ is used to facilitate training and improve generation quality. We construct a style discriminative network D_s to distinguish the style of generated sample $G_F(x,z,e)$.

$$\mathcal{L}_s(G_F, D_s) = -\mathbb{E}_{x,c}[log(D_s(c|G_F(x,z,e))]) \tag{7}$$

where $D_s(c|G_F(x,z,e))$ represents the probability distribution of generated image $G_F(x,z,e)$ over its target domain label c.

Finally, we jointly train the generators, discriminators, and classifiers by using the full objective $\mathcal{L}$ as follows:

$$\mathcal{L} = \lambda_1\mathcal{L}_{adv} + \lambda_2\mathcal{L}_{cycle} + \lambda_3\mathcal{L}_{div} + \lambda_4\mathcal{L}_{cam} + \lambda_5\mathcal{L}_c + \lambda_6\mathcal{L}_s \tag{8}$$

where λ_i controls the scale of loss weight to balance different losses. Here, $\mathcal{L}_{adv}$ consists of forward and backward losses: $\mathcal{L}_{adv} = \mathcal{L}^F_{adv} + \mathcal{L}^B_{adv}$, and $\mathcal{L}_{cycle}$, $\mathcal{L}_c$, $\mathcal{L}_s$, $\mathcal{L}_{cam}$, $\mathcal{L}_{div}$ are defined in the same way.

3.3 Layout Prediction

The goal of Layout Net is to predict the overall spatial arrangement for the whole piece of calligraphy artwork, which includes the space between characters (character spacing), the distance between the lines (line spacing), and the allocation of characters (size of character image). For character and line spacing, we follow the rules which are summarized by professional calligraphers directly. The character spacing is between [0.3, 1], and the line spacing is between [0.25, 2].

For the character size prediction, we formulate it as a sequential modeling problem: given the sequence of character images in context, we need to predict the size of the following character image. Previous research [20] only takes shape feature of character as input and uses first-order Markov chain for modeling. Here, we propose to model it with a recurrent neural network and design four input features: 1) ***E***, emotion embedding, 2) ***R***, ink density, which is the ratio of black pixels to the number of pixels in a standard square box. 3) $\boldsymbol{C}$, the number of components of Chinese character, 4) $\boldsymbol{S}$, the number of strokes of Chinese character. As we only generate Chinese calligraphy artwork written in vertical direction, the column width is usually fixed, which is set by the largest character in the column. We first predict the normalized height of character, then scale the width by the same ratio. Thus, it won't damage the shape of the character. In order to predict the size of character image, we set the height H of character as our learning objective. To handle the size difference of various calligraphy works during training, we normalize the target height H_t as a ratio γ_t by the mean height of all character images in the vertical line.

The Layout Net can be formulated as follows:

$$h_t = RNN([E_t; S_t; C_t; R_t], h_{t-1}) \tag{9}$$

$$\gamma'_t = W_o h_t \tag{10}$$

where h_t is the hidden state of RNN, and W_o is used to map h_t to γ'_t. Mean-square loss L is used to optimize Layout Net:

$$L = \frac{1}{N}\sum_{t=1}^{N}(\gamma_t - \gamma'_t)^2 \tag{11}$$

where N is the number of characters in each vertical line.

4 CHINESE CALLIGRAPHY DATASET

To the best of our knowledge, there is no off-the-shelf Chinese calligraphy image dataset with emotion labels, thus we collect a large-sale Chinese calligraphy dataset, which includes a char-level calligraphy image dataset and a discourse-level calligraphy image dataset. As to emotion, we focus on the most common three emotions, which include happy, sad, and peaceful.

Table 1: Statistics of char-level calligraphy dataset

Style	Train	Valid	Test
Happy	2610	396	396
Sad	3451	507	507
Peaceful	2011	345	345

4.1 Char-level Calligraphy Dataset

To facilitate the research on emotion-driven calligraphy generation, we collect the data from a famous Chinese calligrapher[3] in China. We invite the artist to write three styles of calligraphy characters in the mood of *peaceful, happy and sad.* As shown in Figure 4 (a), characters of three emotions look quite different from each other, and they are also quite different from the standard Liukai font[4]. In general, we can characterize the three emotional styles as follows: *peaceful* characters are clear to read and strokes are separated from each other. *Happy* characters are usually written in a high speed and cursive way. *Sad* characters contain more ink and have higher color contrast. It's observed that, emotional stylistic calligraphy generation is more challenging than ordinary image translation tasks (e.g., cat2dog and picture2painting) as both the shape and style change greatly. Then we scan all characters into images, de-noise and reformat them into 1250 × 1250 pixels. We also annotate each image to get the text content. To perform image translation experiments, we use 4000 images from the standard Liukai font as the source images. Then form the dataset with above three emotional font images as target data correspondingly. We split each style of images into train, validation and test set, and the data statistics is shown in Table 1.

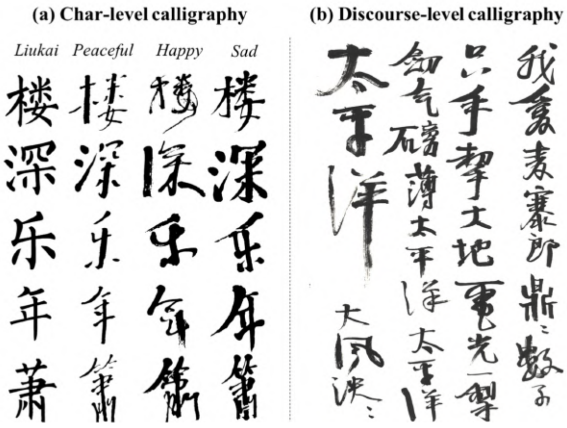

Figure 4: Examples of Chinese calligraphy dataset.

4.2 Discourse-level Calligraphy Dataset

As shown in Figure 4 (b), we also invite the same calligrapher to compose discourse-level Chinese calligraphy artworks. The artist is encouraged to create whole piece of calligraphy with emotions based on the text of the calligraphy. The article images are first scanned and split into lines by column. Then each line is segmented into characters, and the text labels are annotated manually. Totally 10,000 lines of calligraphy are finally obtained, which cover 1300 classical Chinese articles and 65,032 characters. The sentence-level emotion distributions are 42.4% for peaceful emotion, 27.5% for happy emotion, and 30.1% for sad emotion. The percentage is obtained by running the emotion detection in Section 3.1.

5 EXPERIMENTS

5.1 Character Image Generation

5.1.1 Experimental Setup. Following [1], we adopt content accuracy and style discrepancy as evaluation metrics. **1) Content accuracy.** We perform handwriting Chinese character classification experiments on the generated images. We use the pre-trained CNN-HCCR [19] network to classify images generated from different models and compute the classification accuracy. The intuition is that, if the generated characters are realistic, the model will be able to classify them correctly. **2) Style discrepancy.** To measure the discrepancy of style between target images and synthesised images, we calculate the mean-square difference ϕ between style feature representations of target image v^t and generated image v^g. The lower score of style discrepancy indicates a higher similarity between two images. In this experiment, the input of Conv-Block4

[3] https://en.wikipedia.org/wiki/Qiu_Zhijie
[4] https://www.foundertype.com/index.php/FontInfo/index/id/4709

Table 2: Evaluation on character image generation. *Top1-acc* represent content accuracy and *Sty-dis* means style discrepancy. *Pref-sc* means preference score. 'w/o η_s', 'w/o η_c', 'w/o D_c' and 'w/o D_s' represent corresponding ablated models.

Emotions	Peaceful			Happy			Sad		
Models	Top1-acc	Sty-dis	Pref-sc	Top1-acc	Sty-dis	Pref-sc	Top1-acc	Sty-dis	Pref-sc
zi2zi	0.294	0.030	4.5	0.121	0.040	1.7	0.336	0.037	2.3
CycleGAN	0.335	0.035	14.7	0.141	0.037	7.9	0.432	0.039	5.8
U-GAT-IT	0.487	0.033	33.3	0.348	0.033	39.2	0.733	0.032	42.1
ECA-GAN	**0.571**	**0.021**	**47.5**	**0.401**	**0.023**	**51.2**	**0.768**	**0.018**	**49.8**
ECA-GAN w/o η_s	0.498	0.034	–	0.351	0.033	–	0.709	0.035	–
ECA-GAN w/o η_c	0.457	0.028	–	0.312	0.027	–	0.663	0.025	–
ECA-GAN w/o D_s	0.534	0.029	–	0.374	0.031	–	0.722	0.028	–
ECA-GAN w/o D_c	0.502	0.023	–	0.369	0.024	–	0.712	0.020	–
Human	0.645	–	–	0.465	–	–	0.812	–	–

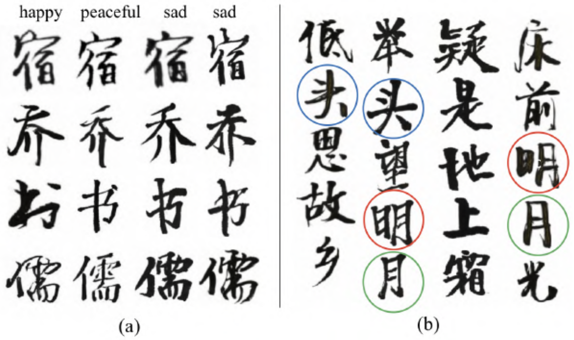

Figure 5: Diversity of generated calligraphy images.

of CNN-HCCR is used as the latent style representation. This can be formulated as follows:

$$v_{ij} = \sum_{k}^{n^2} F_{ik} F_{jk} / n^2 \tag{12}$$

$$\phi = \frac{1}{M^2} (v^g - v^t)^2 \tag{13}$$

where M is the number of filters in one layer, n is the size of feature map F and v_{ij} is the inner product between two feature map F_i, F_j. The final representation of an image, which is composed by v_{ij}, is given by the matrix $v \in \mathbb{R}^{\mathbb{N}\times\mathbb{N}}$.

We choose three models for comparison, including zi2zi [30], CycleGAN [41] and U-GAT-IT [14]. 1) **zi2zi** is used to transfer source images into multi-domain target images based on the pix2pix [11] framework with an additional category embedding for various styles. As zi2zi requires paired data for training, we align the three styles images to standard LiuKai font images respectively. 2) **CycleGAN** uses cycle-consistent adversarial network to learn mapping between two domains. It's a typical baseline of unpaired image-to-image translation model. 3) **U-GAT-IT** is a strong baseline for unsupervised two-domain image translation, which designs an attention module to focus on important regions distinguishing between two domains of images with large shape change. We have also tried StarGAN [4] in our task. However, as character images from different emotional styles are complex and diverse, a single StarGAN can not have enough capacity to model. We use Adam optimizer [15] with batch size of 1 and train 20 epochs. In Equation 8 we set λ_1=5, λ_2=10, λ_3=0.01, λ_4=1000, λ_5=10, and λ_6=10. For learning rate, we initial it with 0.0001 for the first 8 epoch and then linearly decay it every 5 epoch. All the models are trained on NVIDIA Tesla P40 (single card) for two days.

5.1.2 *Experimental Results.*

Both qualitative and quantitative analysis are conducted to evaluate the quality of generated images. For fair comparison, we do not add the diversity loss when comparing with zi2zi, CycleGAN, and U-GAT-IT. We only apply the diversity enhancement technique when generating discourse-level calligraphy artwork.

Quantitative analysis As shown in Table 2, we can get following interesting results: 1) zi2zi performs the worst in our task and is incapable of generating recognizable character images. 2) U-GAT-IT outperforms the CycleGAN, which proves the effectiveness of CAM loss and AdaLIN function on catching the important regions in large geometric variations. 3) ECA-GAN exceeds the U-GAT-IT in both content accuracy and style discrepancy significantly, which indicates the advantages of our proposed model on both content preservation and style control. 4) The content accuracy of human-written images in the testset is also given, which shows the difficulty of the tasks for different emotional styles. In general, our model achieves the highest content accuracy rate and the lowest style discrepancy for all three emotion styles. Besides, we also conduct a user study. 147 educated people of different ages are shown the generated images from different methods, and required to select the most visually pleasing image to the ground-truth. For each emotional style, 100 groups of character images are picked randomly each time for evaluation. Only images are present while the model name is hidden. The final preference score was obtained by averaging the total scores for each model over 147 people. Table 2 shows that our proposed method wins much higher preference score for all three emotional styles in the user study compared to other methods.

Qualitative analysis For qualitative analysis, we demonstrate some generated images from all models in Figure 6. It's obvious that, ECA-GAN can generate more recognizable and style-distinguished images than the baselines. Zi2zi produces very poor results in all three styles. The images generated by zi2zi are either incorrect (green box) or hard to read (blue box). CycleGAN can capture some of stylistic features of the target images such as the heavy ink for

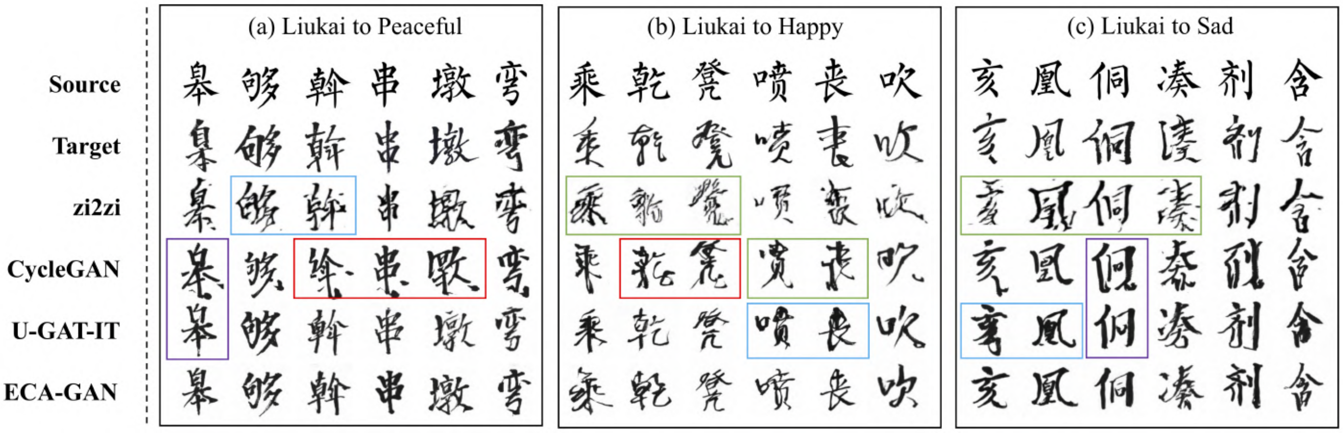

Figure 6: Comparison of the baselines and our approach. The green box shows error strokes, the red box shows the noise information, the purple box indicates missing strokes, and the blue box indicates blurred results.

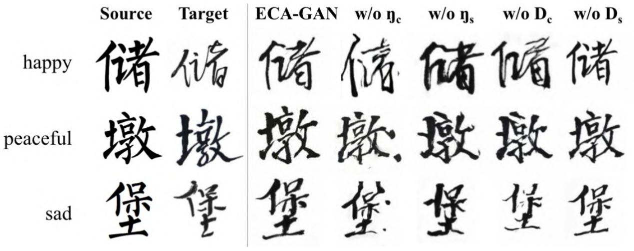

Figure 7: Generated images of ablated models

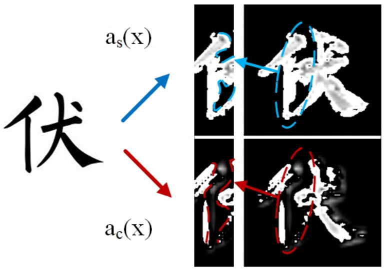

Figure 8: Examples of content CAM and style CAM.

sad mood, but the generated results have many unreasonable noise strokes (shown in red box). As for U-GAT-IT, the generated images are readable in general, but the detailed information reveals its weakness for style transformation. When imitating the cursive strokes of happy mood, U-GAT-IT inclines to produce blurs, just as the characters shown in the blue box. Moreover, it also produces some images with missing strokes as the purple box shows. The comparison illustrates the images generated by our model are more precise in content. What's more, ECA-GAN also learns the emotional styles better. For peaceful style, the strokes are tidy and separated from each other. For happy style, the generated characters are written in a cursive way. For sad style, more ink is presented and color is heavy. These features are consistent with the ground-truth images in the original dataset. Figure 5 (a) illustrates that our model can generate diverse character images. For the same character with different emotions, the generated images are in different styles. Even for the same character with the same emotion (sad), by applying our diversity enhancement technique, we can still generate different variants. Figure 5 (b) shows the diversity in a complete piece of generated calligraphy artwork.

5.1.3 Ablation Study and Visualization. To study the impact of different components on overall performance, we further do an ablation study for ECA-GAN. As Table 2 shows, when we remove the auxiliary classifiers η_c, η_s separately, both the performance drops significantly, which proves the efficacy of content-aware and style-aware representations. When we remove structure discriminator D_c, the *Top1-acc* drops more obvious than style discrepancy, which indicates D_c can guide the model on learning the content of characters. Figure 7 illustrates some generated examples from each ablated model. Besides, to show the effectiveness of CAM images presentation encoding, we also visualize the heated attention maps in Figure 8. We observe that, by adding auxiliary structure classifiers, the content-aware feature map $a_c(x)$ makes the structure region of character distinguishable (left and right components are separate), and the style-aware feature map $a_s(x)$ focuses more on the distinct outline of the character (left and right components are connected).

5.1.4 Universality of our approach. Considering ECA-GAN was invented for emotional stylistic character generation, it's interesting and necessary to explore if the model can be applied in other non-emotional stylized font generation tasks. Here, we conducted further experiments on two public calligraphy generation datasets. One is Luxun-style calligraphy dataset [5], which contains 324 character images. The other is Wangxizhi-style calligraphy dataset [6], which includes 600 character images. For space limitation, we omit the detailed comparison with our baselines and only demonstrate the generated examples in Figure 9. It's observed that, although

[5] https://www.foundertype.com/index.php/FontInfo/index/id/253

[6] https://github.com/changebo/HCCGCycleGAN/blob/master/lanting.zip

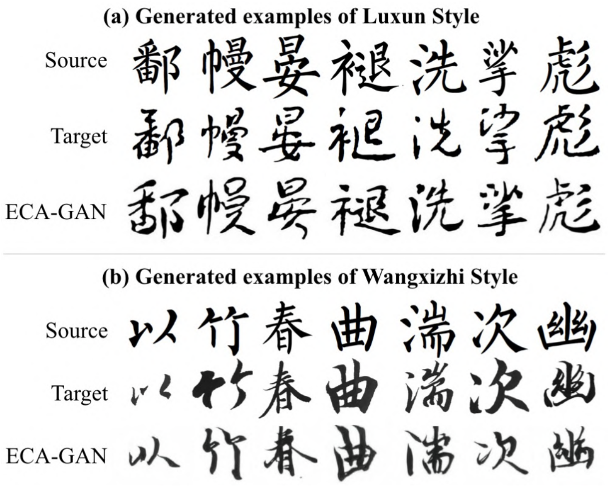
(a) Generated examples of Luxun Style
Source
Target
ECA-GAN
(b) Generated examples of Wangxizhi Style
Source
Target
ECA-GAN

Figure 9: Generated images for Luxun and Wangxizhi style.

the two styles are quite different, ECA-GAN can not only precisely inherit the overall and stroke style distinction from target images but also keep high-level content accuracy of characters. This indicates our model can be a universal framework for stylistic font generation task.

5.2 Layout Prediction

5.2.1 Experimental Setup. The vanilla Recurrent Neural Network is used for modeling (LSTM and GRU are also tried, however, the model is very easy to overfit with limited training data). We set the dimension of embedding to 5 for input features E, C, and S respectively. For the later two, we firstly calculate the distributions of stroke number and component number over all Chinese characters in the vocabulary, then do feature discretization by putting the value into 5 buckets. For ink ratio feature R, we use it directly. To prevent from over-fitting, we set the hidden size of recurrent neural network to 32 and train the model for 10 epochs. The RNN model is trained with the discourse-level Chinese calligraphy dataset mentioned before.

To evaluate the forecast errors, we choose **Mean Absolute Error** (MAE) and **Mean Absolute Percentage Error** (MAPE) [24] as metrics. MAE is a scale-dependant error measure and it reflects the average absolute error. MAPE is a percentage error measure and it's unit-free. The lower values of MAE and MAPE indicate the model behaves closer to human calligrapher in spatial arrangement of calligraphy artwork.

For comparison, we choose three baselines. 1) **Fixed-size layout**, which assigns the same size for every character image; 2) **Random layout**, which predicts the size of character by randomly sampling from a fixed range; 3) **Markov chain** [20], which considers the shape features as input features and predicts the size of character with first-order Markov chain.

5.2.2 Experimental Results. Table 3 shows that, with the help of the RNN, our Layout Net outperforms all baselines significantly on MAE and MAPE metrics. As R, C, and S all belong to shape features, here we do an ablation study by removing the emotion embedding E from input. We observe that: 1) With only shape features, our Layout Net still outperforms Markov chain, we argue that RNN model can capture the dependency for much larger context; 2) After removing emotion embedding, the accuracy of Layout Net drops significantly, which proves the importance of emotion on the layout of calligraphy artworks.

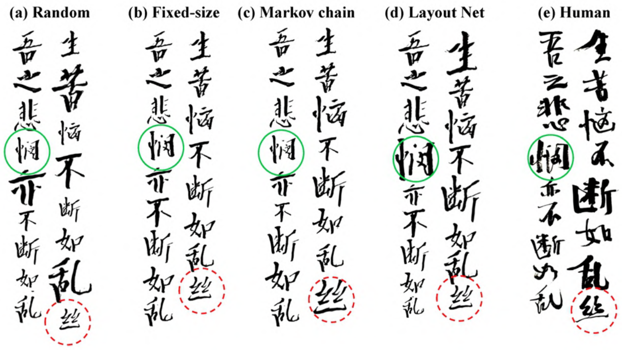

Figure 10: A case study for generated calligraphy with different layout prediction strategies.

Table 3: Evaluation on layout prediction.

Method	MAE	MAPE
Random layout	0.516	69.5%
Fixed-size layout	0.453	40.2%
Markov chain	0.327	35.7%
Layout Net - Emotion	0.281	33.6%
Layout Net	**0.244**	**28.4%**

5.2.3 Case Study. To demonstrate the effect of Layout Net, we do a case study in Figure 10. We compare the calligraphy artworks generated by different layout prediction methods. The human-created image is also given for showing the ground-truth. Figure 10 shows that our proposed model predicts the size of characters more consistent with the ground-truth (e.g., the character in green circle should be large and character in red circle should be small).

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a challenging task, which is to compose discourse-level Chinese calligraphy artwork driven by emotions. We firstly utilize an emotion classifier to detect the emotions from the given text. Then we propose a novel unsupervised multi-domain GAN structure (ECA-GAN) to learn various emotional styles accurately. Thirdly, we define the dominant features for calligraphy layout and propose a layout network to sequentially model these features with textual and visual information from character images. We also collect a large-scale Chinese calligraphy image dataset with rich emotions. Both qualitative and quantitative analysis indicate our approach can generate more **aesthetic**, **stylistic**, and **diverse** calligraphy artworks than baselines. In the future, we will explore more emotional styles of the generated calligraphy such as lonely, cheerful and fearful.

REFERENCES

[1] Bo Chang, Qiong Zhang, Shenyi Pan, and Lili Meng. 2018. Generating handwritten chinese characters using cyclegan. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 199–207.

[2] Jie Chang, Yujun Gu, and Ya Zhang. 2017. Chinese typeface transformation with hierarchical adversarial network. *arXiv preprint arXiv:1711.06448* (2017).

[3] Huimin Chen, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, and Zhipeng Guo. 2019. Sentiment-controllable Chinese poetry generation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 4925–4931.

[4] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 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.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[6] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629* (2016).

[7] Yiming Gao and Jiangqin Wu. 2018. CalliGAN: Unpaired Mutli-chirography Chinese Calligraphy Image Translation. In *Asian Conference on Computer Vision*. Springer, 334–348.

[8] Yiming Gao and Jiangqin Wu. 2020. GAN-Based Unpaired Chinese Character Image Translation via Skeleton Transformation and Stroke Rendering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[10] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 172–189.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

[12] Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. 2017. DCFont: an end-to-end deep Chinese font generation system. In *SIGGRAPH Asia 2017 Technical Briefs*. 1–4.

[13] Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. 2019. SCFont: Structure-Guided Chinese Font Generation via Deep Stacked Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4015–4022.

[14] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. 2020. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BJlZ5ySKPH

[15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)* (2015).

[16] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 35–51.

[17] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*. 700–708.

[18] Pengyuan Lyu, Xiang Bai, Cong Yao, Zhen Zhu, Tengteng Huang, and Wenyu Liu. 2017. Auto-encoder guided GAN for Chinese calligraphy synthesis. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. IEEE, 1095–1100.

[19] Pavlo Melnyk, Zhiqiang You, and Keqin Li. 2018. A high-performance CNN method for offline handwritten Chinese character recognition and visualization. *Soft Computing* (2018), 1–11.

[20] Yiping Meng, Fan Tang, Weiming Dong, and Xiaopeng Zhang. 2016. Optimal character composing for Chinese calligraphic artwork. In *SIGGRAPH ASIA 2016 Posters*. 1–2.

[21] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014).

[22] Wanqiong Pan, Zhouhui Lian, Rongju Sun, Yingmin Tang, and Jianguo Xiao. 2014. Flexifont: a flexible system to generate personal font libraries. In *Proceedings of the 2014 ACM symposium on Document engineering*. ACM, 17–20.

[23] Réjean Plamondon, Wacef Guerfali, and Xiaolin Li. 1998. The generation of oriental characters: new perspectives for automatic handwriting processing. *International journal of pattern recognition and artificial intelligence* 12, 01 (1998), 31–44.

[24] Maxim Shcherbakov, Adriaan Brebels, N.L. Shcherbakova, Anton Tyukov, T.A. Janovsky, and V.A. Kamaev. 2013. A survey of forecast error measures. *World Applied Sciences Journal* 24 (01 2013), 171–176. https://doi.org/10.5829/idosi.wasj.2013.24.itmies.80032

[25] Xiongbo Shi. 2017. The embodied art: an aesthetics of Chinese calligraphy. (2017).

[26] Danyang Sun, Tongzheng Ren, Chongxuan Li, Hang Su, and Jun Zhu. 2018. Learning to write stylized Chinese characters by reading a handful of examples. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 920–927.

[27] Donghui Sun, Qing Zhang, and Jun Yang. 2018. Pyramid Embedded Generative Adversarial Network for Automated Font Generation. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 976–981.

[28] Vincent Tam and KW Yeung. 2010. Learning to write Chinese characters with correct stroke sequences on mobile devices. In *2010 2nd International Conference on Education Technology and Computer*, Vol. 4. IEEE, V4–395.

[29] Yuchen Tian. 2016. Rewrite: Neural Style Transfer For Chinese Fonts. https://github.com/kaonashi-tyc/Rewrite.

[30] Yuchen Tian. 2017. Master chinese calligraphy with conditional adversarial networks. https://github.com/kaonashi-tyc/zi2zi.

[31] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6924–6932.

[32] Helena TF Wong and Horace HS Ip. 2000. Virtual brush: a model-based synthesis of Chinese calligraphy. *Computers & Graphics* 24, 1 (2000), 99–113.

[33] Shan-Jean Wu, Chih-Yuan Yang, and Jane Yung-jen Hsu. 2020. CalliGAN: Style and Structure-aware Chinese Calligraphy Character Generator. *arXiv preprint arXiv:2005.12500* (2020).

[34] Songhua Xu, Tao Jin, Hao Jiang, and Francis CM Lau. 2009. Automatic generation of personal chinese handwriting by capturing the characteristics of personal handwriting. In *Twenty-First IAAI Conference*.

[35] Songhua Xu, Francis CM Lau, William K Cheung, and Yunhe Pan. 2005. Automatic generation of artistic Chinese calligraphy. *IEEE Intelligent Systems* 20, 3 (2005), 32–39.

[36] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*. 2849–2857.

[37] Jinhui Yu and Qunsheng Peng. 2005. Realistic synthesis of cao shu of Chinese calligraphy. *Computers & Graphics* 29, 1 (2005), 145–153.

[38] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. 2019. Multi-mapping Image-to-Image Translation via Learning Disentanglement. In *Advances in Neural Information Processing Systems*. 2990–2999.

[39] Yexun Zhang, Ya Zhang, and Wenbin Cai. 2018. Separating style and content for generalized style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8447–8455.

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

[42] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. 2017. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*. 465–476.

[43] Alfred Zong and Yuke Zhu. 2014. Strokebank: Automating personalized chinese handwriting generation. In *Twenty-Sixth IAAI Conference*.