


ORIGINAL RESEARCH PAPER

Disentangled representation learning GANs for generalized and stable font fusion network

Mengxi Qin¹  | Ziyang Zhang² | Xiaoxue Zhou¹¹ School of Automation, China University of Geosciences, Wuhan, China² School of Arts and Communication, China University of Geosciences, Wuhan, China**Correspondence**

Ziyang Zhang, School of Arts and Communication, China University of Geosciences, Wuhan, 430073, China.

Email: ziyang.zhang@foxmail.com

Funding information

National Natural Science Foundation of China, Grant/Award Number: 61873248

Abstract

Automatic generation of calligraphy fonts has attracted broad attention of researchers. However, previous font generation research mainly focused on the known font style imitation based on image to image translation. For poor interpretability, it is hard for deep learning to create new fonts with various font styles and features according to human understanding. To address this issue, the font fusion network based on generative adversarial networks (GANs) and disentangled representation learning is proposed in this paper to generate brand new fonts. It separates font into two understandable disentangled features: stroke style and skeleton shape. According to personal preferences, various new fonts with multiple styles can be generated by fusing the stroke style and skeleton shape of different fonts. First, this task improves the interpretability of deep learning, and is more challenging than simply imitating font styles. Second, considering the robustness of the network, a fuzzy supervised learning skill is proposed to enhance the stability of the fusion of two fonts with considerable discrepancy. Finally, instead of retraining, the authors' trained model can be quickly transferred to other font fusion samples. It improves the efficiency of the model. Qualitative and quantitative results demonstrate that the proposed method is capable of efficiently and stably generating the new font images with multiple styles. The source code and the implementation details of our model are available at <https://github.com/Qinmengxi/Fontfusion>.

1 | INTRODUCTION

Font design plays an essential role in visual communication and cultural transmission. Compared to handwriting recognition [1–3] creating a novel personalized font can bring a pleasing visual experience and reduce visual fatigue. Moreover, font design combines different eras, different regions and different cultural backgrounds; it efficiently promotes the integration and spread of cultures in the world.

The current font design methods are divided into manual design and artificial intelligence methods. Manual design requires a lot of prior knowledge or expert guidance, and then constantly adjust the stroke structure and parameters to design a set of satisfactory fonts. It is able to generate accuracy and clear font images. But the manual design methods are time-consuming and laborious; the general design cycle of a new font takes 1–2 years [4]. Compared with manual design, automatic

design using artificial intelligence methods has high efficiency and short design cycle. It is very popular among researchers and has made remarkable achievements. Artificial intelligence methods are mainly separated into disentanglement-based methods, translation-based methods and traditional machine vision algorithms.

Traditional machine vision algorithms [5, 6] generated new fonts by modeling and weighting the outline or strokes of different fonts. However, the results were prone to obvious distortion if the structure of calligraphy characters was complicated. And this method required plenty of material and manpower resources to establish mathematical expressions for font; it brought inconvenience to the font design, which cannot be widely used in industry.

The translation-based methods [7, 8] regarded the font images generation process as a translation process from printed font images to target style font images, and an end-to-end

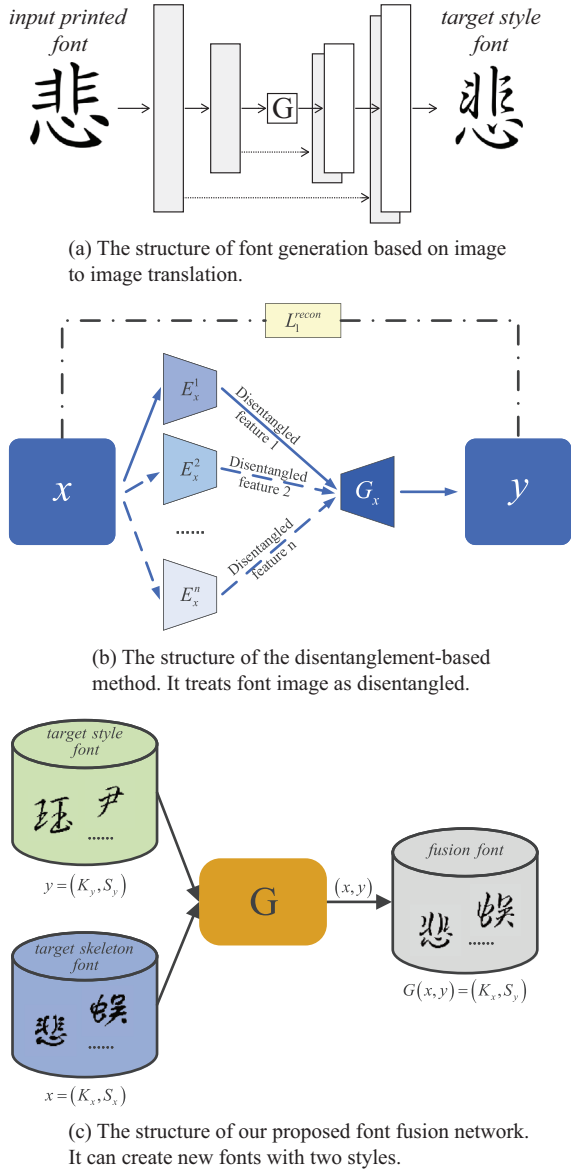


FIGURE 1 (a) represents the translation-based methods and (b) represents the disentanglement-based methods. Instead of imitating the style data distribution of the known font like (a) and (b), proposed method (c) can automatically generate a new font fusion images $G(x, y) = (K_x, S_y)$ with the stroke style K_x of x and the skeleton shape S_y of y

structure using convolutional neural network (CNN) was proposed to learn the mappings between different image domains. As shown in Figure 1a, the printed font is fed to the generator, and it can automatically learn and generate the target style font images under the constraints of L1 loss and adversarial loss. The translation-based methods avoid a lot of modeling and preprocessing time. However, it only focuses on imitating the known font style data distributions, which cannot create new style font by combining different font styles.

Instead of learning the data distribution of the known target style fonts, Figure 1b showed that the disentanglement-based methods [9, 10] decomposed the font images into multiple

interpretable disentangled feature representations. And then the designers can start new font design based on the combination of these interpretable disentangled feature representations. For example, the EMD [9] proposed content encoder and style encoder to separate the font images into content and style. Then the font images were generated by combining the content and style of different character. It provided a good explanation for the abstract feature representations while the content feature representations represents the semantics of the character. The essence of this method is still the image to image translation of font images instead of creating new font images. On the contrary, the FusionGAN [10] further disentangled style into identity and pose. By swapping one's identity and another's pose, a new fusion style of the person was generated. The above methods create new images, but it is unstable for font images fusion, especially for two font images with considerable discrepancy. Obviously, how to stably and efficiently create new font images instead of imitating font images is still the key of current research.

To address the above issue of stably and efficiently designing new fonts instead of imitating, this paper proposes a font fusion network based on disentangled representation learning. It can stably and efficiently create new font by fusing the style and skeleton of difference fonts. Specifically, the font images are regarded as disentangled, and separated into stroke style and skeleton shape using convolutional neural networks. By fusing target skeleton font images x and target style font images y , as described in Figure 1c, it is capable of automatically generating new font images with the guidance of the loss function. Extensive experiment results prove the robustness and effectiveness of our proposed method.

In conclusion, the main contributions of this work are as follows:

- 1) We propose a font fusion network to create new font images by fusing the disentangled skeleton shape and stroke style of different complex font images, while the existing approaches [11–14] focus on font image translation (imitation of existing font images).
- 2) Aiming at the problem that the existing method [10] cannot be used for the fusion of font images stably, we propose fuzzy supervised learning skill, which can stabilize the training process of GANs by designing fault tolerance factors.
- 3) To improve the efficiency of our proposed method, an universal style encoder with dynamically computed AdaIN [15, 16] module is introduced into the font fusion model. As a result, the trained model can directly fuse the trained target style font and another untrained target skeleton font instead of retraining.

The rest of this paper is arranged as follows. First, Section 2 discusses some related work on the font generation. And then the architecture of the proposed font fusion network will be described in detail in Section 3. Section 4 displays the experimental results and validates the effectiveness of our method. Lastly, Section 5 sums up the contributions of this paper.

2 | RELATED WORK

This section describes the research of existing font generation methods in detail. Considering that the traditional machine vision methods are far too limited, as shown in the third paragraph of Section 1, only the font generation based on disentangled representations and image to image translation will be analyzed in related work.

2.1 | Image to image translation

As can be seen in Figure 1a, most existing methods cast font generation problem as a task of image to image translation, such as from printed fonts to target style fonts. Tian et al. [11, 17] proposed to apply pix2pix [7] with end-to-end network structure to transfer KaiTi into HeiTi. But it needed paired data for training, which was difficult to collect in many applications. To address this issue, Chang et al. [18] proposed to employ cycle consistency loss [19] to transfer printed font to target style font images without paired samples for training. It improved efficiency in data collection, but also reduced the accuracy at the same time because there was no standard output for comparison.

These methods mentioned above open up new ideas for font generation, but the simple end-to-end network structure cannot handle complex fonts as well. Consequently, some researchers increased the complexity of network to deal with the above problem. Lyu et al. [12] employed autoencoder to guide GANs to learn the detailed stroke information from autoencoder's low level features. Finally, two subnets were trained together with adversarial loss and reconstruct loss to make the output look real. Yue et al. [13] proposed a structure-guided network to generate Chinese font by using deep stacked networks. They decomposed the font generation task into two separate procedures, and utilized a multi-stage strategy to progressively generate target font. Both of them efficiently improved the precision for complex font generation.

Instead of combining different styles to generate new fonts with multiple styles, these models can only imitate and learn the style data distribution of known target fonts through an end-to-end network structure. However, there are many important reference significance for font generation from the perspective of the end-to-end network structure.

2.2 | Disentangled representations

Recently, there has been significant progress in disentangled representations. This is helpful to understand and utilize abstract feature to create new font with multiple styles. The EMD [9] and MUNT [16] used the content encoder and style encoder to extract the content features and style features from the target style images and target content images, respectively. And then it generated images with target content and style [9, 16] when the above two features were integrated and fed into the decoder. In addition, the AdaIN module [15] was proposed in MUNT to improve the generalization of the network. It

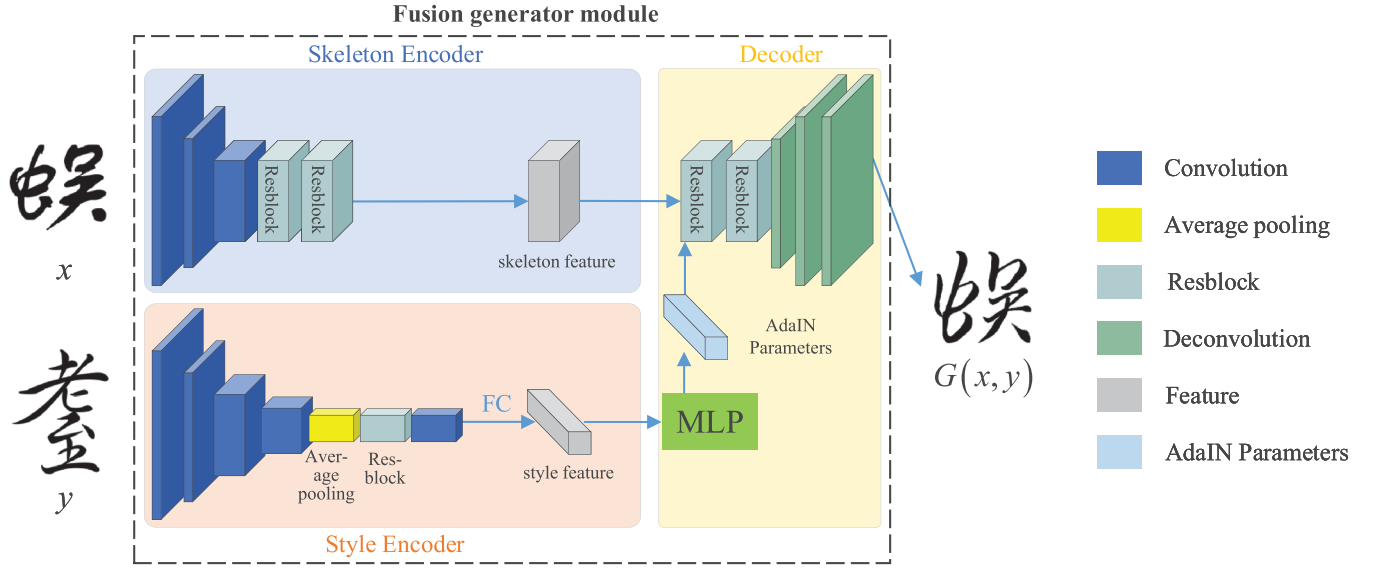
can directly transfer the learned style to a new font instead of relearning the mapping function from the new certain source style to the target style. SSNet [14] held the point that current methods do not take the Chinese semantics and structure into consideration. Instead of directly generating font images, they employed the structure module and semantic module to require the font feature and character semantics, and then combined this information to generate the final target typography. Although, these methods applied the disentangled representation learning, the extracted style was transferred to another font instead of fusing styles of different font. The essence of these methods were still the style transfer of fonts, and the style cannot be further decomposed into fine-grained sub-features for the deep fusion of fonts.

LSCGAN applied improved coherent point drift algorithm to extract disentangled feature representations such as stroke. By fusing the one-hot vector representation of different strokes as input, a new style fusion font was generated. Experiments showed the font style fusion of the standard HeiTi and the HeiTi generated by LSCGAN under different fusion coefficients. However, there was no experiment data to support that the model can fuse different fonts, especially for artistic fonts. At present, LSCGAN is still centered on font image translation research instead of font fusion.

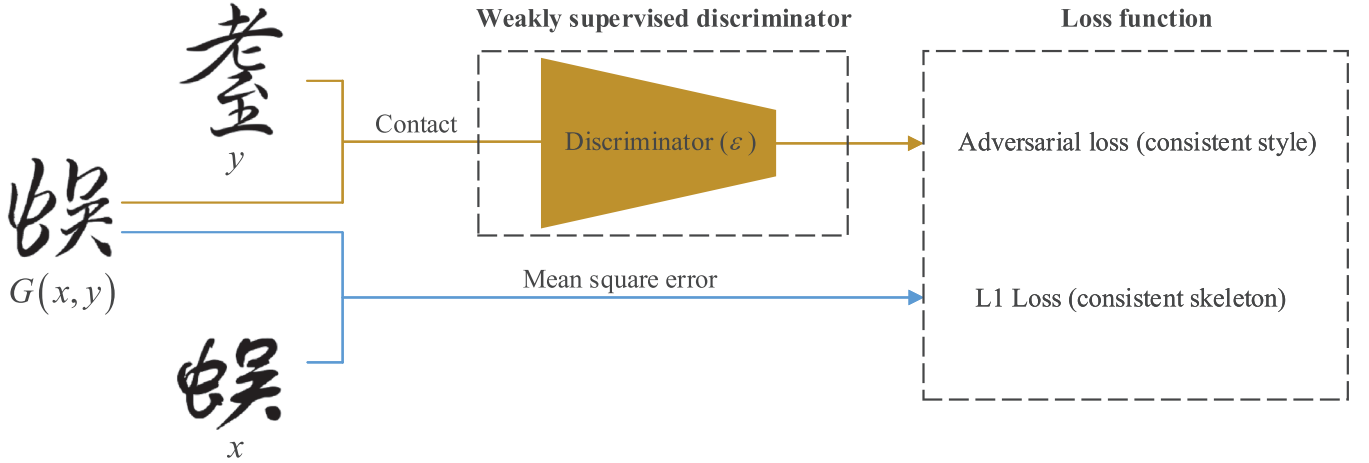
Different from these methods, GeneGAN [20] learned to disentangle the object features from other factors in feature space from weakly supervised 0/1 labeling of training data. Therefore, it is capable of extracting the object feature vectors from the images and transfer it to another images, allowing fine-grained control over the generated images, such as "putting eyeglasses of A onto noses of B". As the improvement, Elegant [21] exchanged the certain part encodings to transfer the attributes of same type from one kind of image to another, which enabled model to manipulate several attributes simultaneously, such as smiling, hair style and eyeglasses. GeneGAN and Elegant focused on the disentanglement and fusion of actual object features such as glasses, smiles and other intuitively understood features on face images. However, abstract object features such as stroke style and skeleton shape in font objects cannot be disentangled.

Aiming at this issue, FusionGAN [10] used identity loss L_I and shape loss L_S to disentangle abstract object feature: identity and shape. After fusing, the new font images with the shape of input images y and identity of input images x were generated. It was successfully applied to the fusion of faces and gestures of different people. Unfortunately, the application of this model is still unstable for grayscale font images.

Different from this method, [22] reckoned that disentangled representations was manifold. They applied CNN to get the representation features of font images, then the font manifold was built through the non-linear mapping. By interpolating and moving between those existing font images manifold, various new style font images were generated. It is obvious that not all interpolation can produce beautiful font images. Uncontrollability is still the limitation of current method. As a result, it always takes a lot of time to explore a suitable interpolation.



(a) Fusion generator module



(b) Fuzzy supervised discriminator and loss function

FIGURE 2 The structure of our proposed font fusion network, where discriminator (ε) uses fuzzy supervised learning. As ε increases, the recognition accuracy of the discriminator will gradually decrease. (a) Fusion generator module. (b) Fuzzy supervised discriminator and loss function

Lacking stable and efficient new font image design method is a common problem at present. In response to this issue, this paper proposes a font fusion network combining AdaIN module and fuzzy supervised learning to create new font image. Compared with [10, 22], our model is more stable, efficient and suitable for font image processing.

3 | METHOD DESCRIPTION

This section describes our proposed model in detail. As can be seen in Figure 2, the whole model is composed of fusion generation module, fuzzy supervised discriminator, loss function and evaluation metrics.

On the one hand, fusion generation module is responsible for fusion of target skeleton font images x and target style font images y . On the other hand, fuzzy supervised discriminator discriminates style similarity of $G(x, y)$ and y . The adversarial loss guides fusion generation module to fuse the stroke style of y . And L1 loss keeps skeleton shape consistent with x . Under the constraints of loss function (adversarial loss and L1 loss), the two networks make progress together after training and create realistic images with the fusion feature representation of two font images.

Eventually, an evaluation metrics and visualization tool from the perspective of quantitative and qualitative is presented to verify the effectiveness of our fusion model.

3.1 | Font generation module

Specifically, suppose that x and y are two input font images with different skeleton shapes and stroke styles, so we can express the font image x as $x = (S_x, K_x)$ and y as $y = (S_y, K_y)$, where S_x, S_y represent the stroke style of different font images and K_x, K_y represent the skeleton shape of different font images.

For font generation module, the style encoder and skeleton encoder is employed to extract the stroke style and skeleton shape from two font images, respectively, such as S_x, S_y, K_x and K_y mentioned above. Then MLP module [16] computes the affine parameters of AdaIN [15] to equip the residual block for better style learning. After recombining these feature representation, the decoder transform the combined feature representation (K_x, S_y) into font fusion image $G(x, y) = (K_x, S_y)$, which have the stroke style of font images x and the skeleton shape of another font images y at the same time.

AdaIN module comes from the arbitrary style transfer idea in [15]. It can directly transfer the trained style to another image instead of retraining. The application of this module also greatly reduces our training time. It is described as follows,

$$AdaIN(\gamma, x, \beta) = \gamma \left(\frac{x - \mu(x)}{\delta(x)} \right) + \beta, \quad (1)$$

where x are the features after convolution and activation layer, δ and μ represent standard deviation and channel-wise mean, β and γ are the affine parameters representing the style. It can be learned from MLP module during training process. Consequently, as long as two font images are successfully fused, our model can transfer the learned stroke style (affine parameters) of one font to another font. Then new font image can be directly fused and generated without wasting a lot of time to retrain. This strategy greatly improves the efficiency of the model.

In addition, to further verify the integrity of the stroke style and skeleton shape after image fusion, the image $G(x, y)$ and y is reconstructed. As shown in Figure 3, our model first disentangles the fusion image $G(x, y)$ into stroke style S_y and skeleton shape K_x . And then recombines it with the S_y and K_y extracted from images y . If the reconstructed images $G(x, y)'$ are equal to $G(x, y)$ and y' are equal to y ; it means that the stroke style and skeleton shape information is complete without missing, and our model successfully fuses the stroke style and skeleton shape from different font images.

3.2 | Fuzzy supervised discriminator

If two fonts are wildly different in skeleton shape and stroke style, it is intractable to stably fuse two font images and generate a new visual pleasing font image. As this time, it is necessary to adjust fusion ratio of skeleton shape and stroke style of different font images and find the balance of fusion. And this ratio is often a fuzzy boundary, which cannot be adjusted by a certain value. To address this issue, our developed fuzzy supervised discriminator by introducing a fault tolerance factor ε in

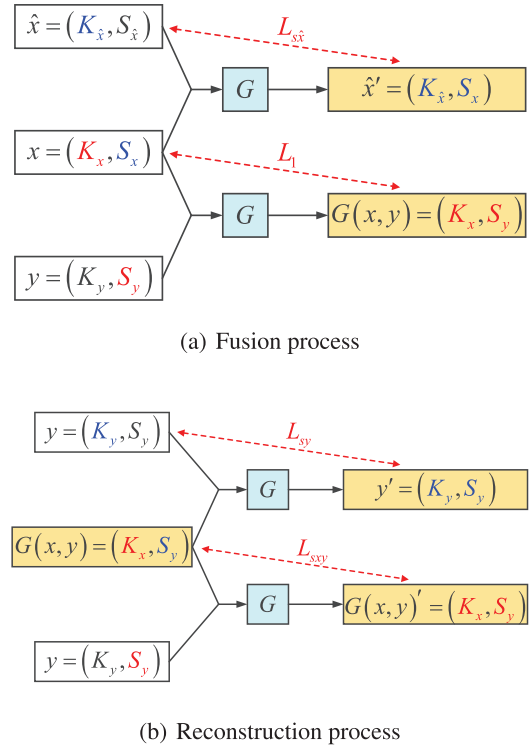


FIGURE 3 The illustration of fusion process and reconstruction process, where the red dotted line connects the reconstructed image and the real image, respectively. (a) Fusion process. (b) Reconstruction process

the adversarial loss function to adjust learning ability. It only needs to fuse part stroke style of one font image. In this way, our model can dynamically adjust ε according to the font difference, and then generate satisfactory results. Take the traditional GANs as example, the adversarial loss function [23] is described in formula 1:

$$\begin{aligned} L_D &= E_{y \sim p_{data}(y)} [\log(D(y) - \varepsilon)] \\ &+ E_{z \sim p_{data}(x), y \sim p_{data}(y)} [\log(1 - \varepsilon - D(G(z)))], \quad (2) \\ L_G &= E_{z \sim p_{data}(z)} [\log(D(z) - \varepsilon)], \end{aligned}$$

where z represents input image and y represents target image. Suppose that ε is equal to 0.1. For the generated images, it can be considered true when the output probability of discriminator is 0.9 (90% similarity in style). Similarly, it is false when the output probability of discriminator is 0.1 (90% dissimilarity) instead of requiring completely different. The experimental analysis finds that this partial style fusion method can efficiently promote the font fusion stability via adjusting ε , especially for two fonts with considerable discrepancy.

3.3 | Loss function

For better guiding the model to fuse the stroke style and skeleton shape of different fonts, multiple indicators such as style loss, skeleton loss and reconstruction loss are combined into

the ultimate objective loss function. It can, respectively, constrain the stroke style and skeleton shape of the generated images and the target images to be consistent, and generate fusion fonts.

3.3.1 | Style loss

In order to guarantee the fusion images having the same style with the target style font image y , it is necessary to train the discriminator to discriminate whether the $G(x, y)$ looks like y . So, the style loss is denoted as follow:

$$\begin{aligned} L_{SD} &= E_{x, \hat{x} \sim p_{data}(x)} [\log(D(x, \hat{x}) - \epsilon)] + \\ &E_{x \sim p_{data}(x), y \sim p_{data}(y)} [\log(1 - \epsilon - D(y, G(x, y)))], \quad (3) \\ L_{SG} &= E_{x \sim p_{data}(x), y \sim p_{data}(y)} [\log(D(y, G(x, y))) - \epsilon], \end{aligned}$$

where \hat{x} are other images with the same style as x . While G tries to generate fake images to confuse D , as shown in formula 2, on the contrary, D tries to distinguish the fake images and real images generated by G . Then, the two networks make progress together after adversarial training, and synthesize realistic images with same style as y .

3.3.2 | Skeleton loss

Considering that the skeleton shape of the fusion images $G(x, y)$ and the target skeleton shape images x should be consistent, L1 loss is utilized to constrain skeleton as follow:

$$L_1 = \|x - G(x, y)\|_1, \quad (4)$$

where the small L1 loss means that the skeleton shape of two font images are similar.

3.3.3 | Reconstruction loss

To further verify the feature information integrity of the fusion image, as shown in Figure 3, the reconstruction loss is designed to minimize the difference between the reconstructed images and real images. We formulate the reconstruction loss as follows:

$$\begin{aligned} L_{s\hat{x}} &= \|\hat{x} - \hat{x}'\|, \\ L_{s\hat{y}} &= \|y - y'\|, \\ L_{s\hat{y}} &= \|G(x, y) - G(x, y')\|, \\ L_{reconstruction} &= L_{s\hat{x}} + L_{s\hat{y}} + L_{s\hat{y}}, \end{aligned} \quad (5)$$

where $G(x, y)'$ and y' are the reconstructed images. The low reconstruction loss indicates that the stroke style and the skeleton shape information are complete.

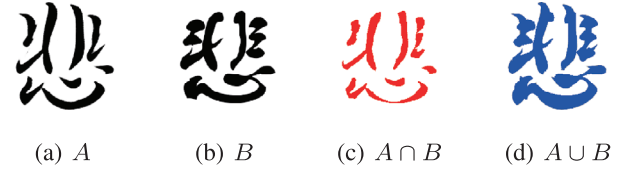


FIGURE 4 The architecture of the font style recognition network based on VGG19. For better intuitive analysis, the features of the last layer of the network are used for visualization



FIGURE 5 Schematism of the intersection and union between the fusion images $G(x, y)$ and the target skeleton images x . The high ratio of $A \cap B$ and $A \cup B$ indicates that highly similarity of two image. (a)-A. (b) B. (c) $A \cap B$. (d) $A \cup B$

Combining the above indicators, the final objective is formulated in the form of a weighted sum:

$$\begin{aligned} L_G &= L_{SG} + \alpha L_1 + \lambda L_{reconstruction}, \\ L_D &= L_{SD}, \end{aligned} \quad (6)$$

where λ and α are hyper-parameters employed to adjust the weight between the L1 loss, the reconstruction loss and the adversarial loss. Generally, the hyper-parameters are set by experience.

3.4 | Font performance metrics

For better quantitatively evaluating the performance of fusion, the style accuracy and skeleton accuracy are built to evaluate font fusion performance. Furthermore, a visualization tool is used to analyze the problem more intuitively.

3.4.1 | Style accuracy

As shown in Figure 4, a VGG19 [23] network is pre-trained to recognize font style. The higher the output probability, the greater the style similarity. Through the pre-trained VGG19 recognition network, it can be determined whether another font style is successfully fused.

Besides, the t-SNE [24], which is a nonlinear dimensionality reduction algorithm, is employed to map the feature of the last layer of VGG19 into 2D space for better visualization. If the

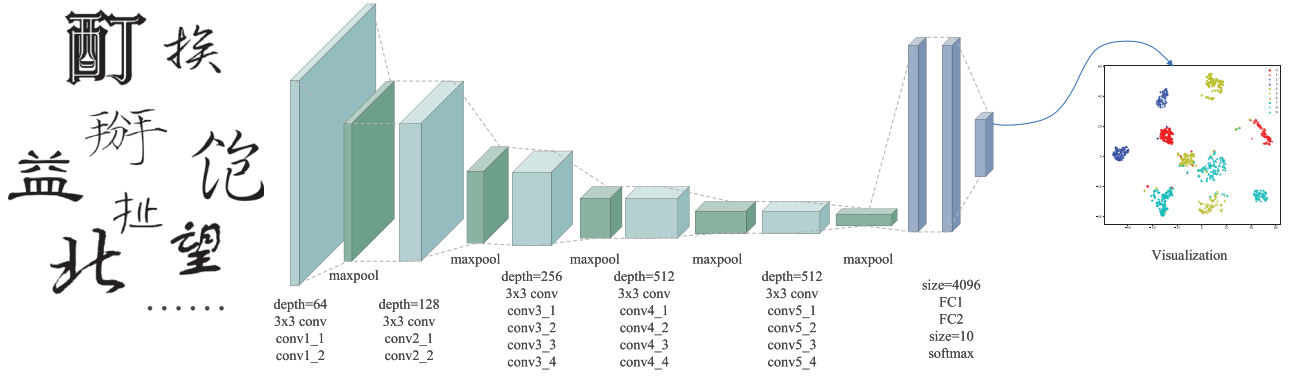


FIGURE 6 Some font image samples in the dataset, where F1 to F10 represent 10 different fonts. (a) F1. (b) F2. (c) F3. (d) F4. (e) F5. (f) F6. (g) F7. (h) F8. (i) F9. (j) F10

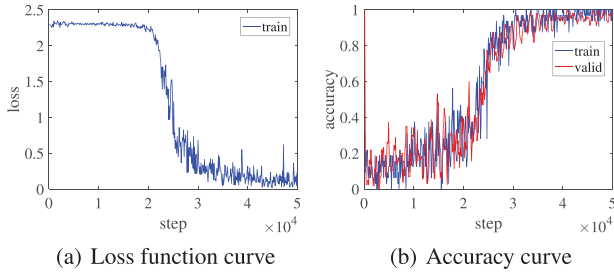


FIGURE 7 Demonstrate some samples of 5 fusion fonts. It is obvious that the fusion font G1 to G5 successfully inherit the skeleton shape of the input font image x and stroke style of input font image y

features of two fonts are very close or intersected on the 2D visualization plane, it means that they are similar in style.

3.4.2 | Skeleton accuracy

IOU is the overlap rate between the image A generated by the model and the original image B . It is formulated as follows:

$$IOU = \frac{A \cap B}{A \cup B} \times 100\%. \quad (7)$$

The high IOU indicates the large overlapping area, and the skeleton shapes of A and B are highly similar. In case of full overlap, the IOU is equal to 100% which represents A and B are 100% similar. So, the skeleton accuracy can be expressed by calculating the IOU of the target skeleton shape image x and the generated fusion font image $G(x, y)$.

Figure 5 shows the schematism of the intersection and union between the target skeleton image and the fusion image, where $A \cap B$ represents the intersection and $A \cup B$ represents the union. Generally speaking, it can be considered that when the IOU is around 0.5, most of the skeleton shape is retained after fusion. Eventually, the Boxplot is utilized to intuitively analyze accuracy distribution.

This section conducts a series of experiments to prove the superiority of our proposed font fusion network.

First, this section builds a dataset and describes the implementation details of our proposed method. Afterwards, the fusion font images are analyzed in terms of qualitative (the details of the fusion font images) and quantitative (the style and skeleton accuracy). For better intuitive analysis, the style accuracy and skeleton accuracy are visualized. In addition, a series of comparative experiments prove that fuzzy supervised learning can effectively promote the fusion of font images. Finally, we show the transfer of font fusion. The specific performance is that the trained model can be directly transferred to other font fusion without retraining.

The extensive experiments illustrate that our proposed model can stably and effectively fuse two different font images and generate novel font images. Moreover, our model also has fine transfer ability.

3.5 | Dataset

Since there is no existing calligraphy dataset, 5965 commonly used Chinese characters and 10 fonts are collected to build our dataset. And Chinese characters and fonts both come from authoritative data such as the Chinese government standard GB18030 and FounderType. The dataset consists of training set and testing set, which contains 4700 and 1265 images with 256×256 resolution, respectively. Figure 6 shows some samples. It can be found that these 10 fonts have obvious differences in the thickness, style and character structure of the strokes. Therefore, it is easy to see the difference between the input font images and the generated font images after fusion.

3.6 | Baseline method

In this section, we compare our method with the state-of-the-art FusionAN algorithms via the authors' publicly available code. It designs a new network that can generate the fusion images with the shape of images y and the identity of images x . In YouTube Pose dataset, it successful generated an image with the pose of one person and the identity of another person.



FIGURE 8 The accuracy curve and loss function curve of the font style recognition network during training. (a) Loss function curve. (b) Accuracy curve

TABLE 1 The classification accuracy on the test set

	F1	F2	F3	F4	F5
Accuracy(%)	98.8	99.5	94.0	99.8	100.0
	F6	F7	F8	F9	F10
Accuracy(%)	96.7	95.8	96.3	98.2	91.8

TABLE 2 The style accuracy and the skeleton accuracy of the font fusion image

	G1	G2	G3	G4	G5
ε	0.0	0.0	0.0	0.0	0.1
Style accuracy(%)	98.0	84.0	97.0	81.0	61.0
Skeleton accuracy(%)	42.2	47.7	40.0	62.3	53.6

The comparison between the generated results and the training process highlights the advantages of our proposed model from the perspective of efficiency and stability.

4 | EXPERIMENT

4.1 | Implementation details

The style encoder and skeleton encoder consist of several convolution layers with 4×4 kernels and residual blocks with

3×3 kernels. Symmetrical to the structure of the encoder, the decoder is composed of two residual blocks with 3×3 kernels and deconvolution layers with 4×4 kernels. Finally, it outputs a RGB font image after convolution with 7×7 kernels and 1×1 stride. Since our experiment only involves 10 fonts, the last fully connected layer of VGG19 is designed with 10 neuron number to train a font classifier. Additionally, the fault tolerance factor ε is more appropriate to be set from 0 to 0.3 when two fonts are wildly different. Otherwise, it is intractable to reach game balance and model is unstable.

It is worth noting that since there is no standard output as a reference, the visual experience of some results that meet the constraints of the loss function is very poor. In this case, it is significant to manually adjust training epochs according to the visual experience of the fusion font images, which are generated during the training process. Generally speaking, 3 to 5 epochs of training are appropriate.

In this experiment, our method is implemented in TensorFlow. And the computer that performed the experiment was equipped with a NVIDIA GTX1080Ti GPU, 32GB RAM and 3.3GHz Intel Xeon E5-2600 CPU.

4.2 | Qualitative and quantitative experiment results

This part of the experimental results will prove the superiority of our proposed model from a qualitative and quantitative perspective.

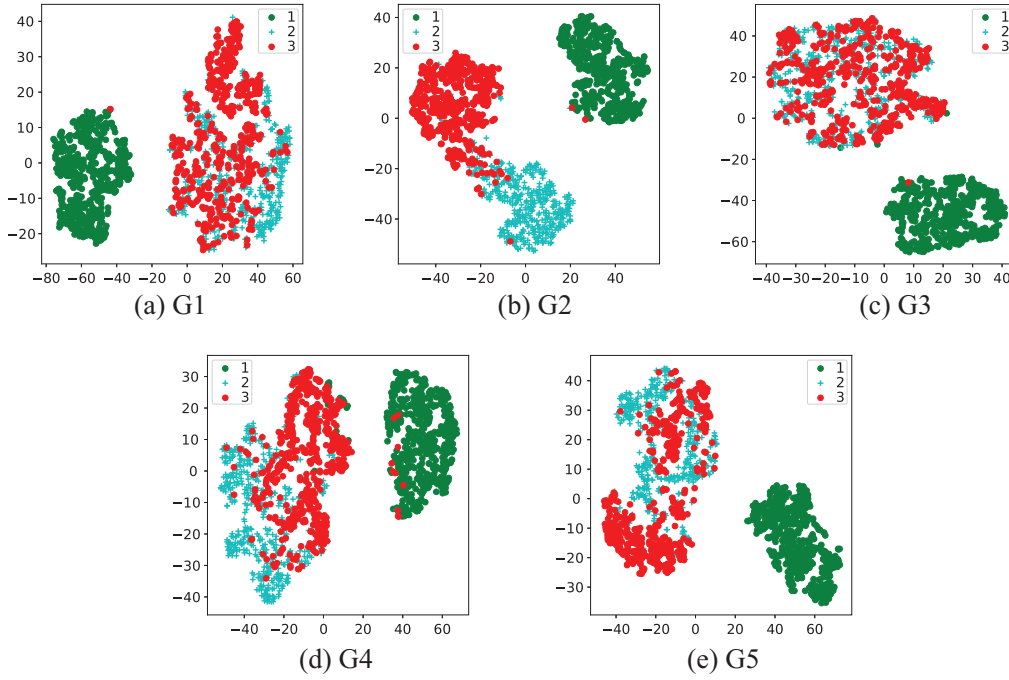


FIGURE 9 The visualization of the output style feature of the last layer of VGG19. G1 to G5 are the visualization of 5 different fusion font mentioned above. And labels 1 to 3 represent the target skeleton shape font image x , target stroke style font image y and fusion font image $G(x, y)$. (a) G1. (b) G2. (c) G3. (d) G4. (e) G5

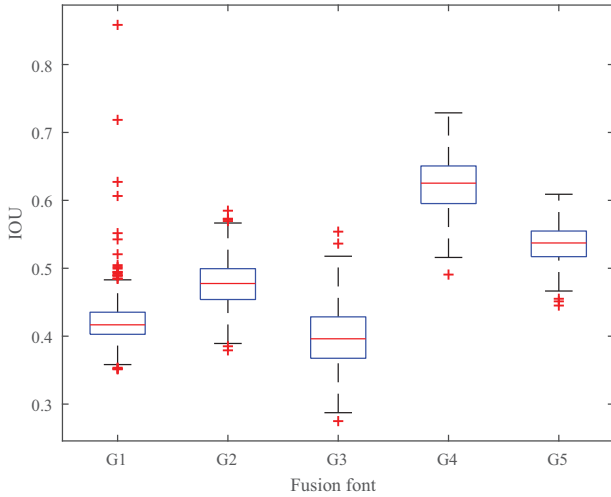


FIGURE 10 The Boxplot of the IOU of the generated fusion font image, where G1 to G5 represents the 5 different fusion font images

4.2.1 | Qualitative analysis

Qualitative analysis is to compare the details of the new fusion font images $G(x, y)$ and the input font images x and y . Due to the space limitation, only 5 representative font combinations are chosen for fusion. Part results are presented in Figure 7, the input font image x and input font image y , respectively, represent two font with different skeleton shape and stroke style, and $G(x, y)$ indicates new generated font after fusion.

It can be seen that all generated fonts have successfully inherited the skeleton shape of input font images x and the stroke style of input font images y . And it has a delicate appearance without strokes missing and noise. For instance, the fusion font G1 not only has elegant script similar to the skeleton of the font F9, but also has slender figures and distinct vigorous strokes like the style of font F8. Structurally, each character of font F3 is wider and flatter, while the style of font F4 is more square and sharper. After fusion, the new generated fusion font G2 incorporates a sharp stroke style while retaining a flat structure. Similarly, the fusion font G3, G4 and G5 also successfully combines the skeleton shape of the input font image x and stroke style of input font image y , and makes it look more visual pleasing.

4.2.2 | Quantitative analysis

For better quantifying the accuracy of the fusion font image $G(x, y)$, we first analyze the training process and results of the proposed font classification network to prove its effectiveness (the IOU directly calculates the overlap ratio of the two images, there is no necessary to pre-train to verify its effectiveness). Then, the validated IOU and font style recognition network are employed to quantitatively evaluate the generated fusion font images.

Effectiveness of the evaluation method

As shown in the Figure 8, the loss of the font style recognition network converges quickly during the training process. And the accuracy of train set and validation set gradually increased and

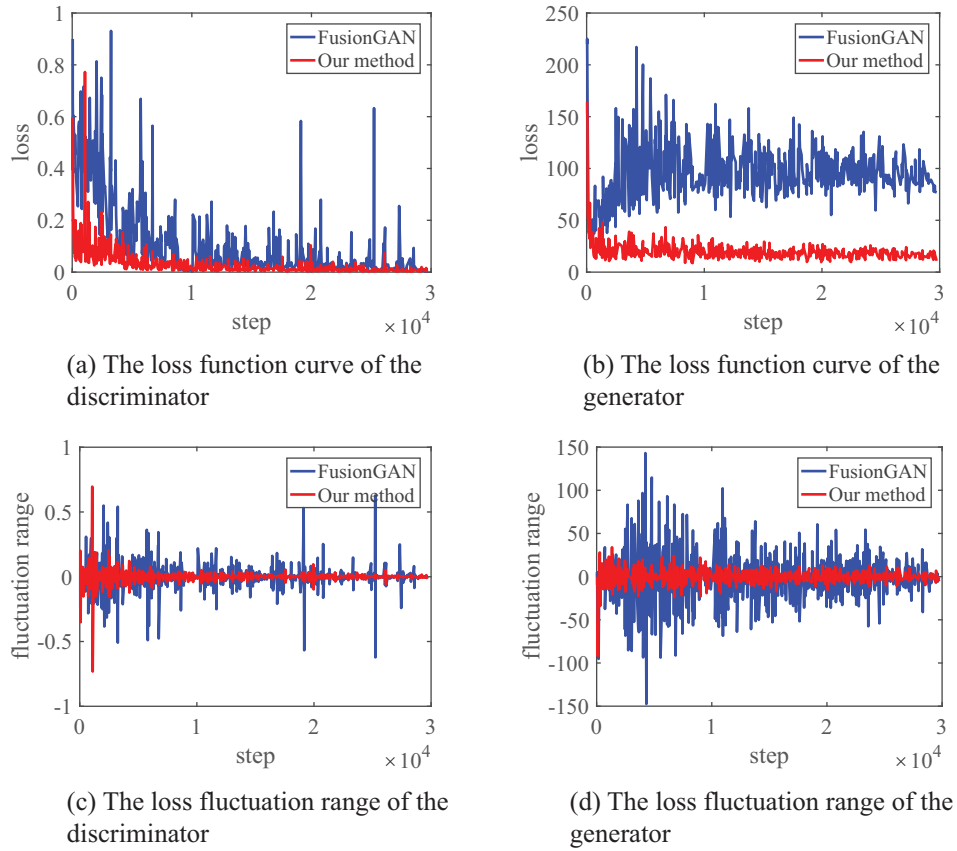


FIGURE 11 The loss function curve and its fluctuation range of the discriminator and generator during the training process. (a) The loss function curve of the discriminator. (b) The loss function curve of the generator. (c) The loss fluctuation range of the discriminator. (d) The loss fluctuation range of the generator

tended to 100%. It indicates that the training process is stable and there is no overfitting phenomenon. Furthermore, as shown in Table 1, the font classification network can achieve an average accuracy of 97.1% on the test set. The experimental results demonstrate that it can efficiently evaluate font style accuracy.

Accuracy analysis

As described in Table 2, the style accuracy and the skeleton accuracy of the font fusion images are analysed. For the skeleton accuracy, it is in the range of 40–65%, which means that about 50% of the area is overlapped with the target skeleton shape font image x . The experiment demonstrates that our method can fuse the skeleton shape of the target skeleton shape font image x .

For the style accuracy, it is related to the fault tolerance factor ε in fuzzy supervised discriminator. When ε is equal to 0, the goal of the network is to completely fuse the stroke style of the target style image y . So, the style accuracy of the G1 to G4 reaches 98.0%, 84.0%, 97.0% and 81.0% respectively. It means that the target style font images y and the style of the fusion font images $G(x, y)$ are consistent. On the contrary, if ε is not equal to 0, it can only fuse the part stroke style of the target stroke style images y due to the fuzzy supervised discriminator. Therefore, the style accuracy varies with the ε and the stroke style discrepancy of the two font used for fusion.

For example, the ε of G5 is equal to 0.1. Consequently, G5 is only 61% similar to the target stroke style font y . It also can be seen from Figure 7 that only part stroke style of the fusion font image G5 is similar to target stroke style font image y . It proves that our model can adjust the degree of style fusion by setting ε .

4.2.3 | Visualization analysis

For more intuitive evaluation, the output features of the FC layer of the font style recognition network are visualized. If the visualized feature points of the two fonts are very close to each other, it indicates that they have similar style and style accuracy is high. Figure 9 shows the visualization of the output style feature of G1 to G5. Obviously, all font fusion images $G(x, y)$ both intersect the target style fonts image y and are far away from the style of the target skeleton shape font image x on the 2D visualization plane. It demonstrates that our proposed method is capable of fusing the stroke style of another font.

Moreover, the accuracy of the skeleton shape is visualized using boxplot, which shows the distribution of the values in qualitative and quantitative data. As described in Figure 10, “+” represents the abnormal point and blue rectangular box represents the reasonable data distribution. Besides, the red

	FusionGAN			Our method		
	Input x	Input y	$G(x, y)$	Input x	Input y	$G(x, y)$
Step 2000	幾	恫	幾	瓶	洮	洮
Step 4000	娛	決	娛	數	勸	勸
Step 6000	皮	賄	皮	彼	床	床
Step 8000	蠟	嫂	蠟	情	仆	仆
Step 10000	籃	吳		孟	市	市
Step 12000	籟	雳	籟	視	陂	陂
Step 14000	王	叨		葩	威	威
Step 16000	孖	坝		桅	凹	凹
Step 18000	泥	洹		个	能	能
Step 20000	旺	客		嘲	萝	萝
Step 22000	輸	矜		沟	皋	皋
Step 24000	属	政		酣	冽	冽
Step 26000	認	覷		莽	跛	跛
Step 28000	史	幅		扭	末	末
Step 30000	謙	紗		緬	薪	薪

FIGURE 12 Demonstrate the generated font fusion image every 2000 steps, where input x and input y respectively represent the target skeleton shape font and target style font y

horizontal line is the median of the data. From the analysis of the boxplot of the skeleton shape accuracy of 5 fusion font, the following rules can be found:

- 1) The medians of the skeleton shape accuracy of 5 fusion font are both in the middle of the rectangular box, which are about 0.5.
- 2) The distribution intervals of the skeleton shape accuracy are all less than 0.07.
- 3) The abnormal point is very few, and the most skeleton shape accuracy are concentrated in the reasonable data distribution area.

Obviously, the rules mentioned above indicate that the distribution of the generated fusion font accuracy is quite uniform and concentrated, and the fusion font $G(x, y)$ retains the skeleton shape of the input font image x . It means that the diversity of the new generated fusion font and the robustness of network are excellent, which proves the effectiveness of the proposed method.

Extensive experiments prove that our proposed model successfully fuses two fonts and generates a new font with well visual experience from qualitative aspect and quantitative aspect.

4.3 | Compared experiment results

4.3.1 | Stability comparative analysis

We compared the new fusion fonts and loss function curve generated by FusionGAN and our method to verify the stability of our method, especially for two font images with considerable discrepancy.

Take the fusion of font F1 and F6 as example, two fonts are quite different whether they are from skeleton shape or stroke style. In term of skeleton shape, font F1 has a bigger width and finer strokes than font F6, and it is composed of dynamic arcs without any straight lines. In term of stroke style, the horizontal strokes of font F1 are inclined upward to the upper right corner, and there is a hook at the start of the pen. On the contrary, F6 has a elegant style and looks more square. So, fusing these two fonts is a very challenging task. Furthermore, it can also effectively illustrate the stability of our proposed method.

The loss function curve and fluctuation range of the FusionGAN and our method are shown in Figure 11. The experiment finds that the loss function curve of the FusionGAN fluctuates greatly. As shown in Figures 9c and 9d, the loss fluctuation of the discriminator reaches plus or minus 0.6, and the loss fluctuation of the generator reaches plus or minus 100. And there usually are large burrs, especially at 3245, 5724, 19120, 25230 steps, the highest loss even reaches 0.9 in discriminator. In contrast, the loss of our method begins to converge at 2000 steps with small fluctuations, and the training process is stable and quickly converges.

In addition, the generated results are output every 2000 steps. Figure 12 shows that the structure of the font images $G(x, y)$ generated by FusionGAN is very poor. For example, most areas of the font image $G(x, y)$ generated at steps 9000–10,000 is blank (limited by space, the figure only shows the results generated at step 10,000), the font image generated at steps 14,000–22,000 lacks a lot of stroke information, it does not look like a font image at all. And the font image generated in steps 24,000–28,000 only retains the information of the input x without fusing the skeleton information of the input y . Oppositely, the result generated by our method combines the stroke style of the input x and the skeleton structure of the input y , and no stroke information is lost. Besides, the visual effect of the generated font fusion image gets better with the increase of training steps.

The experiment demonstrates that the discriminator using fuzzy supervised learning with appropriate fault tolerance

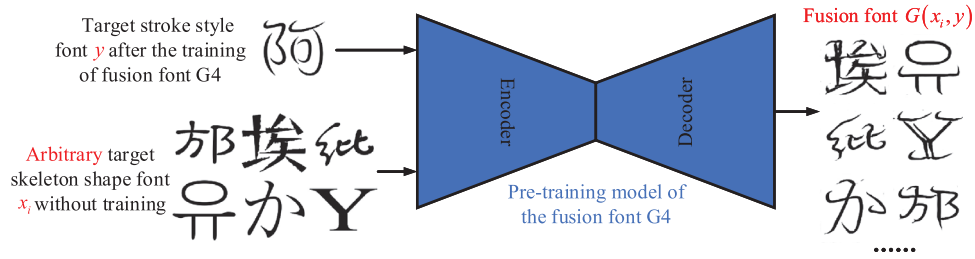


FIGURE 13 An overview of our font fusion transfer



(a) Some results of the Chinese font fusion transfer.



(b) Some results of transferring Chinese font stroke style to other language fonts.

FIGURE 14 More results of the fusion font transfer instead of retraining. (a) Some results of the Chinese font fusion transfer. (b) Some results of transferring Chinese font stroke style to other language fonts

factor can promote the fusion of two font, especially for two font images with considerable discrepancy.

4.3.2 | Efficiency analysis

In addition, our model additionally introduces the universal style encoder module to improve the extensiveness of neural networks. Once the training of the fusion of the target stroke style font y and target skeleton shape font \hat{x} is completed, our model can directly fuse this target stroke style font y and arbitrary untrained font x_i to generate a new fusion font $G(x_i, y)$ instead of retraining.

As shown in Figure 13, it takes the fusion font G4 as an example. The pre-training model of fusion font G4 directly fuses the trained target stroke style font y and the untrained target skeleton shape font x_i (F3, F4 and F9). Figure 14a shows some generated fusion font $G(x_i, y)$, it can be seen that even if two font images are fused using font fusion transfer instead of retraining, the $G(x_i, y)$ (framed with a red dashed box) are still satisfactory. Besides, the font fusion transfer can also cross language. As shown in Figure 14b, the disentangled stroke style extracted from Chinese font is transferred to other language, such as English, Japanese and Korean font. And various novel cross-language style fonts are generated. Consequently, for the same target style font, it is unnecessary to waste a lot of time to retrain. Once the target style font is trained, font fusion transfer can directly fuse it and other target skeleton fonts, even if they are in different languages. It greatly improves the efficiency of our model.

5 | CONCLUSION

This paper first proposes a font fusion network to create new font images, while the existing approaches mainly focus on font style imitation. Our proposed model separates font image into stroke style and skeleton shape using convolutional neural networks. According to personal preference, various interesting fonts can be generated by fusing the stroke style and skeleton shape of different fonts. Our model greatly reduces the time for font design. It is creative and innovative, and has great applications prospect in culture industry. Additionally, we propose fuzzy supervised learning and introduce AdaIN into model. Compared to [8], the fuzzy supervised learning makes the training of our model more stable. Even if two font images with considerable discrepancy in stroke style and skeleton shape, our model can also converge quickly and produce satisfactory results. With the introduction of AdaIN, the trained target style font y can be directly integrated with other target skeleton shape font x_i instead of retraining. This greatly improves the efficiency of our model. The extensive experiments show the priority of our proposed model.

In the future, several directions can be explored on this basis. First, considering that some fonts lack sufficient samples, the training samples can be reduced to improve the applicability of the model. Last but not least, the proposed fuzzy supervised learning skill can be used for other types of GANs.

ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China under Grant 61873248.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

DATA AVAILABILITY STATEMENT

The raw/processed data required to reproduce these findings can be shared at this time as the data.

ORCID

Mengci Qin  <https://orcid.org/0000-0001-7552-4341>

REFERENCES

1. Drobac, S., Lindén, K.: Optical character recognition with neural networks and post-correction with finite state methods. *Int. J. Document Anal. Recognit. (IJ DAR)* 23(4), 279–295 (2020)
2. Zhang, J., Guo, M., Fan, J.: A novel CNN structure for fine-grained classification of Chinese calligraphy styles. *Int. J. Document Anal. Recognit. (IJ DAR)* 22(2), 177–188 (2020)
3. Gupta, J. D., Samanta, S., Chanda, B.: Ensemble classifier-based off-line handwritten word recognition system in holistic approach. *IET Image Process.* 12(8), 1467–1474 (2018)
4. Lei, Y., Zhou, L., Pan, T., Qian, H., Sun, Z.: Learning and Generation of Personal Handwriting Style Chinese Font. In: 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1909–1914. IEEE, Piscataway (2018)
5. Suveeranont, R., Igarashi, T.: Example-based automatic font generation. In: *Proceedings of the International Symposium on Smart Graphics*, pp. 127–138. Springer, Berlin (2010)
6. Saito, J., Nakamura, S.: Fontender: Interactive Japanese text design with dynamic font fusion method for comics. In: *Proceedings of the International Conference on Multimedia Modeling*, pp. 554–559. Springer, Cham (2019)
7. Isola, P., Zhu, J. Y., Zhou, T., Efros, A. A.: Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134. IEEE, Piscataway (2017)
8. Gao, Y., Guo, Y., Lian, Z., Tang, Y., Xiao, J.: Artistic glyph image synthesis via one-stage few-shot learning. *ACM Trans. Graph. (TOG)* 38(6), 1–12 (2019)
9. Zhang, Y., Zhang, Y., Cai, W.: Separating style and content for generalized style transfer. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8447–8455. IEEE, Piscataway (2018)
10. Joo, D., Kim, D., Kim, J.: Generating a fusion image: One's identity and another's shape. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1635–1643. IEEE, Piscataway (2018)
11. Tian, Y.: 'zi2zi: Master chinese calligraphy with conditional adversarial networks. (2017) URL <https://github.com/kaonashi-tyc/zi2zi>
12. Lyu, P., Bai, X., Yao, C., Zhu, Z., Huang, T., Liu, W.: Auto-encoder guided GAN for Chinese calligraphy synthesis. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), pp. 1095–1100. IEEE Computer Society, Los Alamitos (2017)
13. Jiang, Y., Lian, Z., Tang, Y., Xiao, J.: Sefont: Structure-guided chinese font generation via deep stacked networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4015–4022. AAAI Press, Palo Alto (2019)
14. Zhang, J., Chen, D., Han, G., Li, G., He, J., Liu, Z., Ruan, Z.: SSNet: Structure-semantic net for Chinese typography generation based on image translation. *Neurocomputing* 371, 15–26 (2020)
15. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510. IEEE, Piscataway (2017)

16. Huang, X., Liu, M. Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189. Springer, Berlin (2018)
17. Tian, Y.: Rewrite: Neural style transfer for Chinese fonts (2016)
18. Chang, B., Zhang, Q., Pan, S., Meng, L.: Generating handwritten chinese characters using cyclegan. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 199–207. IEEE, Piscataway (2018)
19. Zhu, J. Y., Park, T., Isola, P., Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232. IEEE, Piscataway (2017)
20. Zhou, S., Xiao, T., Yang, Y., Feng, D., He, Q., He, W.: Genegan: Learning object transfiguration and attribute subspace from unpaired data. *arXiv preprint, arXiv:1705.04932* (2017)
21. Xiao, T., Hong, J., Ma, J.: Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 168–184. Springer, Berlin (2018)
22. Guo, Y., Lian, Z., Tang, Y., Xiao, J.: Creating new Chinese fonts based on manifold learning and adversarial networks. In: *Eurographics (Short Papers)*, pp. 61–64. ACM, New York (2018)
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* pp. 2672–2680. ACM, New York (2014)
24. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9(11), 2579–2605 (2008)
25. Lin, X., Li, J., Zeng, H., Ji, R.: Font generation based on least squares conditional generative adversarial nets. *Multimedia Tools Appl.* 78(1), 783–797 (2019)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint, arXiv:1409.1556* (2014)

How to cite this article: Qin, M., Zhang, Z., Zhou, X.: Disentangled representation learning GANs for generalized and stable font fusion network. *IET Image Process.* 16, 393–406 (2022).
<https://doi.org/10.1049/ipr2.12355>