Compositional Zero-Shot Artistic Font Synthesis

Xiang Li¹, Lei Wu¹, Changshuo Wang¹, Lei Meng¹,²*, Xiangxu Meng¹
¹School of Software, Shandong University
²Shandong Research Institute of Industrial Technology
202035260@mail.sdu.edu.cn, i_lily@sdu.edu.cn,
202115242@mail.sdu.edu.cn, lmeng@sdu.edu.cn, mxx@sdu.edu.cn

Abstract

Recently, many researchers have made remarkable achievements in the field of artistic font synthesis, with impressive glyph style and effect style in the results. However, due to less exploration in style disentanglement, it is difficult for existing methods to envision a kind of unseen style (glyph-effect) compositions of artistic font, and they can only learn the seen style compositions. To solve this problem, we propose a novel compositional zero-shot artistic font synthesis gan (CAFS-GAN), which allows the synthesis of unseen style compositions by exploring the visual independence and joint compatibility of encoding semantics between glyph and effect. Specifically, we propose two contrast-based style encoders to achieve style disentanglement due to glyph and effect intertwining in the image. Meanwhile, to preserve more glyph and effect detail, we propose a generator based on hierarchical dual styles AdaIN to reorganize content-styles representations from structure to texture gradually. Extensive experiments demonstrate the superiority of our model in generating high-quality artistic font images with unseen style compositions against other state-of-the-art methods. The source code and data is available at moonlight03.github.io/CAFS-GAN/.

1 Introduction

Artistic fonts are frequently employed in signboards, posters, magazines, and web pages, playing an integral role in captivating and sustaining the audience’s attention. The compelling nature of these fonts lies in the fact that designers meticulously craft visually appealing and harmonious glyph and effect styles that suit the occasion and theme. In the course of design, the designers draw upon design theory and aesthetic factors to conceive various style elements, often requiring only a momentary mental picture. It is worth noting that if we can provide a deep learning model with enough glyph styles and effect styles as prior knowledge, whether the model can also design a kind of artistic font with unseen integrated style like humans.

In order to achieve the automatic synthesis of artistic font based on deep learning, some conventional methods [Azadi et al., 2018; Gao et al., 2019; Li et al., 2020a] focus on the integrated style (glyph-effect) transfer and generate an artistic font library with the existing style. These works treat the style of artistic fonts as a whole and generalize the learned integrated style to any character content. However, they ignore the independence and decoupling of styles, making these methods ineffective in scenarios where glyph and effect styles must be controlled separately. Therefore, the conventional methods cannot synthesize artistic fonts with unseen style (glyph-effect) compositions. There are also some recent works [Ge et al., 2021; Li et al., 2022b] that propose learning disentangled style representations and synthesizing content-glyph-effect controllable artistic font images. Unfortunately, these methods focus on the seen style compositions and must require a large amount of data paired with the three attributes
of content, glyph, and effect. Due to pixel-level supervision information, these methods inevitably focus on pixel-level relationship instead of creating the new style compositions, resulting in the generated images with a messy structure and unclear texture.

In this paper, we propose a novel and practical task, called compositional zero-shot artistic font synthesis (CAFS), which focuses on unseen style composition synthesis, see Figure 1. It aims to learn the compositionality of glyphs and effects from the training set and is tasked with generalizing to unseen style (glyph-effect) compositions on any character. To realize this task, we propose a new model, CAFS-GAN, from the perspective of style disentanglement and content-styles representations reorganization.

For the style disentanglement, we propose two contrast-based style encoders, glyph encoder and effect encoder, which implement glyph and effect disentanglement and precise style feature extraction. The key idea is that we introduce glyph style contrastive loss and effect style contrastive loss to learn the style commonalities and differences. For the content-styles representations reorganization, we propose an artistic font generator based on hierarchical dual styles AdaIN, which progressively feeds glyph and effect information to preserve more image details. The key idea is that the hierarchical dual styles AdaIN completes the composition of glyph and content in the high-dimensional AdaIN layer, and the composition of effect and content in the low-dimensional AdaIN layer. Moreover, to enable the model to synthesize artistic font images with controllable style attributes, we adopt the well-known GAN [Goodfellow et al., 2014; Li et al., 2021] framework and introduce two multi-task discriminators, glyph discriminator and effect discriminator that constrain the style of the generated glyphs and effects, respectively. Finally, to comprehensively evaluate the generated results, we propose two evaluation metrics: glyph outline misalignment (GOLM) and effect perception error (EPE).

In summary, our contributions are as follows:

- We propose a novel compositional zero-shot artistic font synthesis gan (CAFS-GAN) to synthesize unseen style compositions for artistic font images. Meanwhile, our model supports the control of artistic font synthesis from three aspects (i.e., glyph, effect, and content).
- We propose two new evaluation metrics, called glyph outline misalignment (GOLM) and effect perception error (EPE), which enrich the evaluation methods from the unique attribute of the artistic font.
- Extensive experiments demonstrate the effectiveness and superiority of our model in synthesizing unseen style compositions in Chinese standard, creative, handwriting, calligraphy artistic fonts and English artistic fonts.

2 Related Work

2.1 Artistic Font Generation

Early artistic font generation approaches are based on the high regularity of the spatial distribution for effects. T-Effects [Yang et al., 2016] and DynTypo [Men et al., 2019] focus on texture and special effects for synthesizing complex and realistic artistic font images. TET-GAN [Yang et al., 2019a] and ShapeMatching-GAN [Yang et al., 2019b] establish the mapping between the original shape and the effect, using the CNN (Convolutional Neural Network) to realize the text effect transfer. Then, AGIS-Net [Gao et al., 2019] and FET-GAN [Li et al., 2020a] attempt the synchronous style transfer of glyphs and effects of arbitrary characters or symbols. Recently, DSE-Net [Li et al., 2022b] and GZS-Net [Ge et al., 2021] have conducted separate studies on the glyph structure [Chen et al., 2021] and effects of artistic fonts. Although these methods separately encode artistic font glyph and effects, they still have a significant data dependency on paired data. These models learn to synthesize artistic fonts by training on paired seen style combinations. Therefore, the optimization process for the model parameters is based on the pixel-level error between the generated and real images, which causes the model to focus excessively on pixel-level mapping relationships. This makes it difficult for the models to create new style combinations.

2.2 Disentangled Representation Learning

Disentangled representation learning aims to infer latent factors for a given object in the real world, where each latent factor is responsible for generating a semantic feature [Han et al., 2021; Yang et al., 2021; Saini et al., 2022; Dong et al., 2022a]. Following VAE, [Higgins et al., 2017] introduces β-VAE to discover interpretable latent factor representations in a completely unsupervised manner. [Chen et al., 2018] improved β-VAE, and further proposed a principled classifier-free measure of disentanglement. Recently, a large amount of works [Zhang et al., 2018; Li et al., 2020b; Luo et al., 2022] have made great contributions to disentangled shape and texture, unfortunately, they are unable to generate novel combinations not witnessed during training.

2.3 Compositional Zero-Shot Learning

Compositional zero-shot learning stands at the intersection of compositionality and zero-shot learning and focuses on state and object relations. Compositionality [Naeem et al., 2021] can loosely be defined as the ability to decompose an observation into its primitives. Zero-shot learning [Gao et al., 2018; Hong et al., 2022; Feng et al., 2022; Lin et al., 2022] aims at recognizing or generating novel classes that are not observed during training. Recently, [Yang et al., 2022] present a novel decomposable causal view that characterizes how compositional concepts are formed. [Karthik et al., 2022; Mancini et al., 2021] propose to address the problem of open-world compositional zero-shot learning. [Li et al., 2022c] propose a novel siamese contrastive embedding network to excavate discriminative prototypes of state and object.

In this paper, we propose a compositional zero-shot artistic font synthesis, and use the artistic font’s glyph and effect style as attribute primitives. More importantly, our method is the first to estimate the unseen style compositions, and uses the joint compatibility and differences between the two styles to synthesize and optimize the detailed characteristics of the image styles.
3 Method

3.1 Problem Define

Compositional zero-shot artistic font synthesis (CAFS) aims to predict an unseen style composition, namely to synthesize glyph-effect compositions that do not exist in the training set. Let us denote with $X = \{x_i\}_{i=1}^N$ the set of effect attributes, with $Y = \{y_j\}_{j=1}^N$ the set of glyph attributes, with $Z = \{z_k\}_{k=1}^N$ the set of style compositions, and with $C = X \times Y$ the set of all their possible compositions. $T = \{Z_t, C_t\}$ is a training set where $Z_t$ is a character set seen during training (for all characters $C_t \subseteq C$). When the glyph and effect elements in $C_t$ covers all elements in $X$ and $Y$, $T$ can be used to train the model $f : \{Z_t, C_t\} \rightarrow \{Z_t, C_u\}$ synthesizing the artistic font images with unseen style combinations where $C_u \subseteq C$ denote the unseen style compositions and $C_t \cup C_u = C$.

The difficulty of the CAFS task varies depending on the proportion of the $C_t$. If the style compositions in $C_t$ covers all compositions and $C_u \equiv \emptyset$, the task definition is the same as the conventional artistic font generation task, where the model only needs to predict the seen style combination on arbitrary character content. In the case of $C_t \subseteq C$, since the model only learns jointly compatibility of encoding semantics between glyph and effect in seen style compositions, it is very challenging to predict unseen style combinations. It is worth noting that as the $C_t$ shrinks, the training data can provide the model with fewer data on the joint compatibility relationship of glyph and effect. In this case, the shrink of composition information hinders the recognizability of glyph and effect, making it difficult for the model to predict unseen style combinations. Regarding this hypothesis, we verified it in Experiment 5.5.

3.2 Overview of CAFS-GAN

The CAFS-GAN consists of the following modules: two style encoders $E_x$ and $E_y$, two style similarity attention modules, a content encoder $E_z$, an artistic font generator $G$, and two style discriminators $D_x$ and $D_y$ [Lin et al., 2020; Dong et al., 2022b], as shown in Figure 2. First, $E_x$ and $E_y$ represent style encoder and glyph style encoder, respectively, which are used to disentangle and extract glyph and effect style features. At the end of the two style encoders, we add a style similarity attention (SSA) module, which uses the similarity of style attributes to enhance the model’s perception of various glyphs or effects. The structure details of $E_x$ and $E_y$ are similar to VGG11 [Simonyan and Zisserman, 2014]. Unlike $E_x$ and $E_y$, our $E_z$ adds several padding layers to increase the sampling times for the font strokes at the image’s edge. This operation protects the integrity of the character structure. In addition, since the content information of characters belongs to high-dimensional semantic information, we add resblocks at the end of the con-
tent encoder to retain more content information. Lastly, our $D_{s}$ and $D_{y}$ are two multi-task discriminators consisting of FRN (Filter Response Normalization) [Singh and Krishnan, 2020] and convolutional layer, which consists of multiple output branches. Each branch learns a binary classification determining whether an artistic font has real glyph style or real effect style.

In the next sections, we will look at the two aspects: style (glyph-effect) disentanglement (in sections 3.3 and 3.4) and content-styles representations reorganization (in section 3.5).

### 3.3 Contrast-Based Style Encoders

In the process of achieving the CAFS task, the style encoders need to provide the generator with disentangled glyph features and effect features. However, the actual situation is that the visual elements of effect, glyph, and content are entangled, and the commonly used data enhancement methods cannot eliminate or highlight a certain visual element. Therefore, we introduce a contrastive learning [He et al., 2020; Li et al., 2022a] strategy to encourage encoders to identify deep similarities and differences between the two style attributes, as shown in Figure 4. Taking the pipeline of effect extraction as an example, we define $s_{x}^{+}$ and $s_{x}^{-} = \{s_{1,x}, s_{2,x}, ..., s_{N_{e}−1,x}\}$ as the positive sample and negative sample set of the original input $s_{x}$ respectively. $N_{e}$ denotes the number of all kinds of effect styles, one of which is the effect of positive samples, and $N_{e} − 1$ is the number of all kinds of negative effects. The positive pair $(s_{x}, s_{x}^{+})$ only shares the same effect, and the negative pair $(s_{x}, s_{r,x}^{-})$ have different effects ($1 ≤ r ≤ N_{e} − 1$). At this time, we utilize the effect style contrastive loss to enhance the effect similarity between positive pairs and the dissimilarity between negative pairs:

$$L_{styg}^{E_{y}} = −\log\frac{\exp(f_{x} \cdot f_{x}^{+}/\tau)}{\sum_{i=1}^{N_{e}−1} \exp(f_{x} \cdot f_{r,x}^{+}/\tau)},$$

where $f_{x}$, $f_{x}^{+}$, $f_{r,x}^{+}$ are effect features obtained by $s_{x}$, $s_{x}^{+}$, $s_{r,x}^{-}$ through $E_{x}$. Similarly, we also impose the glyph style contrastive loss $L_{styg}^{E_{y}}$ to improve the glyph encoder. Furthermore, the total style contrastive loss can be defined as:

$$L_{styg} = L_{styg}^{E_{x}} + L_{styg}^{E_{y}},$$

### 3.4 Style Similarity Attention

To make full use of the style similarity between positive samples and original samples as auxiliary information for synthesizing disentangled style features, we introduce a style similarity attention module at the end of the style encoders. Specifically, we use the style features of positive samples as $K$ and $V$, and use the style features of original images as $Q$. Style similarity attention can be expressed as:

$$SSA(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^{T}}{\sigma}\right)F^{+},$$

where $f_{x}$, $f^{+}$ are style features from the original image and positive sample, and $\sigma$ factor follows Attention Mechanism [Vaswani et al., 2017; Guo et al., 2020] to prevent the magnitude of the dot product from growing extreme.

Overall, our proposed contrast-based style encoders encourage the encoders to have more robust style disentanglement ability. The SSA enhances the prominent glyph-effect characteristics by amplifying the specific style signal strength to obtain a pure glyph or effect representation.

### 3.5 Hierarchical Dual Styles AdaIN

Since neural networks are easier to retain abstract information in high-dimensional layers and easier to retain color information in low-dimensional layers [Gatys et al., 2016], we propose an artistic font generator based on hierarchical dual styles AdaIN. Specifically, we pass the disentangled glyph features and effect features through a fully connected layer (FC) to obtain high- and low-dimensional glyph style signals, respectively. Here, we input the glyph signal into the AdaIN layer [Huang and Belongie, 2017] of the generator and fuse the content information through high-dimensional connections, so that the generator can determine the overall outline and structural pattern [Wu et al., 2020] in the early stage of generation. Furthermore, the effect signal is input to the generator through low-dimensional connections to render the color and texture details of the artistic font based on the established glyph. Formally, we use the style encoders and SSA to extract the effect feature $f_{x}$ and glyph features $f_{y}$, and input them to the fully connected layer. The fully connected layer aims to align $f_{x}$ and $f_{y}$ with the channel means and variances of the content inputs $f_{z}$, and to use $f_{x}$ and $f_{y}$ as the adaptive affine parameters of the AdaIN layer (i.e., $w$ and $b$). Ultimately, we achieve a progressive reorganization of the content with glyph and effect using hierarchical dual styles AdaIN:

$$f_{l+1} = \begin{cases} w_{y} \left(\frac{f_{l} - \mu}{\sigma}\right) + b_{y}, & l \leq h \\ w_{x} \left(\frac{f_{l} - \mu}{\sigma}\right) + b_{x}, & l > h \end{cases},$$

where $l$ denotes the current layer number and $h$ denotes the threshold for dividing the high-dimensional AdaIN layers and the low-dimensional AdaIN layers.
and then obtain the texture gram matrix \[ G_{atys} \]
calculate the feature maps of the image in the deep layers, use the VGG19 [Simonyan and Zisserman, 2014] network to tic font images. EPE is used to evaluate whether the tex-

Effect perception error (EPE). The visual communication of effect is often presented in the form of texture in artistic font images. EPE is used to evaluate whether the texture information of the generated image is accurate. First, we use the VGG19 [Simonyan and Zisserman, 2014] network to calculate the feature maps of the image in the deep layers, and then obtain the texture gram matrix [ Gatys et al., 2016] through the inner product operation to represent the texture features. Then, EPE can be formulated as follows:

\[ EPE = \frac{1}{n} \sum_{i=1}^{n} (G_i - G'_i)^2, \]

where \( n \) denotes the number of network layers involved in the calculation of feature maps, \( G_i \) and \( G'_i \) denote the gram matrices calculated in the \( i \)-layer network of the real image and the generated image.

5 Experiments
5.1 Datasets
SSAF Dataset. SSAF [Li et al., 2022b] contains a large number of high-quality Chinese and English artistic images, with annotations for their glyphs, effects, and content.

Fonts Dataset. Fonts [Ge et al., 2021] is a computer generated RGB font image dataset. It consists of 52 English letters with 5 independent attributes: letter identity, font size, font color, background color, and glyph.

5.2 Implementation Details
In our experiments, all images are resized to 128×128 pixels. The hyperparameters are set as: \( \lambda_{adv} = 1.0 \) and \( \lambda_{sty} = 0.1 \). In training, we set the batch size as 8 and train \( 10^5 \) iterations for Chinese artistic font generation and \( 2 \times 10^5 \) iterations for English. The learning rate is set to 0.0001, using Adam optimizer. Regarding the division of all possible style compositions, we set the proportion of the number of style compositions in \( C_u \) to \( C_t \) to be 1: 8. In each category of artistic font, 775 Chinese characters and 22 uppercase English letters are used for training. 197 Chinese characters and 4 uppercase English letters are used for testing.

5.3 Comparison with SOTA Methods
Quantitative comparison. We compare three non-zero-shot methods, such as AGIS-Net [Gao et al., 2019], FET-GAN [Li et al., 2020a], and StarGANv2 [Choi et al., 2020]. The style (glyph-effect) compositions of the target artistic fonts synthesized by them are seen in the training. Meanwhile, we also compare two zero-shot methods, such as GZS-Net [Ge et al., 2021] and DSE-Net [Li et al., 2022b]. The style compositions they synthesized are unseen during training.

In Table 1, the CAFS-GAN proposed by us has achieved through the inner product operation to represent the texture features. Then, EPE can be formulated as follows:

\[ EPE = \frac{1}{n} \sum_{i=1}^{n} (G_i - G'_i)^2, \]

where \( n \) denotes the number of network layers involved in the calculation of feature maps, \( G_i \) and \( G'_i \) denote the gram matrices calculated in the \( i \)-layer network of the real image and the generated image.

5 Experiments
5.1 Datasets
SSAF Dataset. SSAF [Li et al., 2022b] contains a large number of high-quality Chinese and English artistic images, with annotations for their glyphs, effects, and content.

Fonts Dataset. Fonts [Ge et al., 2021] is a computer generated RGB font image dataset. It consists of 52 English letters with 5 independent attributes: letter identity, font size, font color, background color, and glyph.

5.2 Implementation Details
In our experiments, all images are resized to 128×128 pixels. The hyperparameters are set as: \( \lambda_{adv} = 1.0 \) and \( \lambda_{sty} = 0.1 \). In training, we set the batch size as 8 and train \( 10^5 \) iterations for Chinese artistic font generation and \( 2 \times 10^5 \) iterations for English. The learning rate is set to 0.0001, using Adam optimizer. Regarding the division of all possible style compositions, we set the proportion of the number of style compositions in \( C_u \) to \( C_t \) to be 1: 8. In each category of artistic font, 775 Chinese characters and 22 uppercase English letters are used for training. 197 Chinese characters and 4 uppercase English letters are used for testing.

5.3 Comparison with SOTA Methods
Quantitative comparison. We compare three non-zero-shot methods, such as AGIS-Net [Gao et al., 2019], FET-GAN [Li et al., 2020a], and StarGANv2 [Choi et al., 2020]. The style (glyph-effect) compositions of the target artistic fonts synthesized by them are seen in the training. Meanwhile, we also compare two zero-shot methods, such as GZS-Net [Ge et al., 2021] and DSE-Net [Li et al., 2022b]. The style compositions they synthesized are unseen during training.

Table 1: Quantitative comparison of the CAFS-GAN and the existing state-of-the-art methods.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Disentangled Style</th>
<th>Training</th>
<th>( L_1 ) loss ( \downarrow )</th>
<th>FID ( \downarrow )</th>
<th>SSIM ( \uparrow )</th>
<th>GOLM ( \downarrow )</th>
<th>EPE ( \downarrow )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGIS-Net [Gao et al., 2019]</td>
<td>×</td>
<td>paired</td>
<td>0.2277</td>
<td>107.01</td>
<td>0.4313</td>
<td>81.025</td>
<td>4.3981</td>
</tr>
<tr>
<td>FET-GAN [Li et al., 2020a]</td>
<td>×</td>
<td>paired</td>
<td>0.2050</td>
<td>100.56</td>
<td>0.4474</td>
<td>68.820</td>
<td>7.5113</td>
</tr>
<tr>
<td>StarGANv2 [Choi et al., 2020]</td>
<td>×</td>
<td>unpaired</td>
<td>0.2997</td>
<td>72.24</td>
<td>0.3647</td>
<td>82.934</td>
<td>3.7708</td>
</tr>
<tr>
<td>GZS-Net [Ge et al., 2021]</td>
<td>✓</td>
<td>paired</td>
<td>0.2460</td>
<td>140.35</td>
<td>0.3648</td>
<td>87.328</td>
<td>7.2335</td>
</tr>
<tr>
<td>DSE-Net [Li et al., 2022b]</td>
<td>✓</td>
<td>unpaired</td>
<td>0.1754</td>
<td>72.19</td>
<td>0.4428</td>
<td>83.345</td>
<td>3.7332</td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>unpaired</td>
<td>0.1271</td>
<td>64.79</td>
<td>0.5883</td>
<td>73.225</td>
<td>3.0734</td>
</tr>
</tbody>
</table>
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline
Content & Glyph & Effect & AGIS-Net & FET-GAN & StarGANv2 & GZS-Net & DSE-Net & Ours \human \hline
瑞 & 鹤 & 墙 & 瑞 & 瑞 & 瑞 & 瑞 & 瑞 & 瑞 \hline
矛 & 航 & 额 & 忙 & 忙 & 忙 & 忙 & 忙 & 忙 \hline
明 & 微 & 慧 & 明 & 明 & 明 & 明 & 明 & 明 \hline
勤 & 辣 & 使 & 勤 & 勤 & 勤 & 勤 & 勤 & 勤 \hline
B & R & X & b & B & B & D & b & B \hline
\end{tabular}
\caption{Synthesizing seen style compositions}
\end{table}

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Content & Glyph & Effect & Baseline (A) & (B) & (C) & (D) \hline
滨 & 梁 & 高 & 滨 & 滨 & 滨 & 滨 & \hline
\end{tabular}
\caption{Synthesizing unseen style compositions}
\end{table}

Figure 4: Comparison with state-of-the-art methods. Manual results by human are shown in the last column as ground truth. Six rows of experimental results correspond to (1) Chinese artistic font with normal glyph, (2) Creative glyph, (3) Handwriting glyph, (4) Calligraphy glyph, (5) English artistic font with simple effect, (6) English artistic font with delicate effect.

Figure 5: Ablation study of CAFS-GAN. The Baseline includes three encoders and a generator without two style contrastive losses and SSA, and it receives two style vectors that have been spliced in the basic AdaIN layer. The setup of 5 groups of experiments: (A) adding $L_{\text{sty}}^{E_x}$ to the Baseline, (B) incrementally adding $L_{\text{sty}}^{E_y}$, (C) incrementally adding SSA, (D) replacing AdaIN with a reverse version of hierarchical dual styles AdaIN based on (C), (E) incrementally adding hierarchical dual styles AdaIN based on (C). The setup of experiment (E) denotes the full of CAFS-GAN.

Qualitative comparison. In Figure 4, our method has generated photo-realistic glyph and effect style and is superior to other methods. We can easily observe that some methods work well in the normal glyph, but their performance in creative, handwriting, and calligraphy drops sharply. For English, some methods are difficult to generate the correct glyph and effect (e.g., DSE-Net), and the others are difficult to generate the correct character content (e.g., GZS-Net).

5.4 Ablation Study

We conducted ablation study to validate the effectiveness of the components and loss functions of the model. The experimental results are depicted in Figure 5 and Table 2.

Style contrastive losses. The purpose of style contrastive losses is to disentangle the glyph and effect and improve the encoder’s ability to extract pure glyph and effect features. In Figure 5(A), after we add $L_{\text{sty}}^{E_x}$, the dark red effect disappears obviously and the correct metal texture effect appears. After we simultaneously add $L_{\text{sty}}^{E_x}$ and $L_{\text{sty}}^{E_y}$, the glyph structure of (B) becomes more accurate than (A).

Style similarity attention. The SSA makes use of the style similarity between the positive and original samples to enhance the feature signal of the glyph and effect. We add SSA to the setup of experiment (B). In Figure 5(C), the stroke on the left side of this character has been significantly improved.

Hierarchical dual styles AdaIN. This structure helps the model to synthesize artistic fonts from structure to texture through hierarchically input to improve image details. The reverse version of this structure treats the glyph as low-dimensional information and the effect as high-dimensional information. We add the reverse version of hierarchical dual styles AdaIN to the setup of experiment (C). Figure 5 (C)(D) shows that the reverse version will lose a lot of effects and glyph details. Then, we add the right version of hierarchical...
Table 2: Quantitative evaluation of ablation study.

<table>
<thead>
<tr>
<th></th>
<th>$L_1$ loss ↓</th>
<th>FID ↓</th>
<th>SSIM ↑</th>
<th>GOLM ↓</th>
<th>EPE ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.2750</td>
<td>261.08</td>
<td>0.3039</td>
<td>189.29</td>
<td>2.6751</td>
</tr>
<tr>
<td>(A)</td>
<td>0.2852</td>
<td>257.73</td>
<td>0.2653</td>
<td>187.51</td>
<td>3.9582</td>
</tr>
<tr>
<td>(B)</td>
<td>0.2336</td>
<td>201.66</td>
<td>0.3345</td>
<td>183.83</td>
<td>2.0330</td>
</tr>
<tr>
<td>(C)</td>
<td>0.2290</td>
<td>178.07</td>
<td>0.3333</td>
<td>185.10</td>
<td>1.6954</td>
</tr>
<tr>
<td>(D)</td>
<td>0.2452</td>
<td>179.61</td>
<td>0.3099</td>
<td>185.10</td>
<td>1.2076</td>
</tr>
<tr>
<td>(E)</td>
<td><strong>0.2251</strong></td>
<td>179.61</td>
<td>0.3520</td>
<td>179.25</td>
<td><strong>1.0767</strong></td>
</tr>
</tbody>
</table>

Figure 6: Influence of the proportion of seen style compositions. The x-coordinate represents the proportion of $C_t$ to $C$, and the y-coordinate represents the value of each metric.

5.5 Proportion of the Seen Style Compositions

We also discussed the influence of the proportion of seen style composition $C_t$ to all possible style compositions $C$ on the experimental results. We use six different training sets to train CAFS-GAN, each containing the same three effects and three glyphs, but their number of compositions is different. The ratios of style combinations of $C_t$ to $C$ are set to 4/9, 5/9, 6/9, 7/9, 8/9, and 9/9. As shown in Figure 6, with the proportion increase, the model’s performance presents an overall improved state. Therefore, we concluded that sufficient glyph-effect joint compatibility relationship will improve the model’s ability to understand the artistic font’s attributes and help the model synthesize unseen style compositions.

5.6 Visualization

In order to further demonstrate the style disentanglement capability of the $E_x$ and $E_y$ and the ability to recombine content and styles of the generator, we visualize the attention maps generated by style encoders and feature maps generated by the generator. In Figure 7(a)(b), we feed three different effects of the artistic font images to $E_x$ and $E_y$. The texture part of these images got a lot of attention from the effect encoder. The glyph encoder tends to focus on local areas of artistic fonts, which are the unique characteristics of the glyphs, such as curves and corners. In Figure 7(c), the structure of feature maps of fonts are changed firstly (e.g., the lines become clear, and the corners become apparent). Then, there is more pixel filling inside the feature maps of the font. After that, the texture is rendered.

5.7 Style Interpolation

We further demonstrate the flexibility of CAFS-GAN through glyph style interpolation and effect style interpolation. In CAFS-GAN, we can explicitly control the weighting between different glyph or effect representations and decode the integrated representation back to the image space, obtaining the new mixed attributes, see Figure 8. This is meaningful to the diversification of artistic fonts.

5.8 Conclusion

In this paper, we propose a new task called compositional zero-shot artistic font synthesis (CAFS), which allows synthesizing arbitrary character’s artistic font image with unseen style compositions. To achieve this task, we propose the CAFS-GAN model, focusing on style disentanglement of glyph and effect, and hierarchical reorganization of content and styles representations. We also propose two evaluation metrics for a more comprehensive evaluation of artistic font images: glyph outline misalignment and effect perception error. Extensive experiments demonstrate the effectiveness of our model’s multi-attributes control and the superiority of generation quality over existing methods.
Acknowledgments

This work is supported in part by the Excellent Youth Scholars Program of Shandong Province (Grant no. 2022HWYQ-048), the TaiShan Scholars Program (Grant no. tsqn202211289), the National Key R&D Program of China (Grant no. 2021YFC3300203), the “20 Regulations for New Universities” funding program of Jinan (Grant no. 2021GXRC073).

References


[Han et al., 2021] Yuxuan Han, Jiaolong Yang, and Ying Fu. Disentangled face attribute editing via instance-aware latent space search. In IJCAI, pages 715–721, 2021.


