SCIENCE CHINA
Information Sciences

June 2022, Vol. 65 160109:1–160109:2
https://doi.org/10.1007/s11432-021-3435-8

• LETTER •
Special Focus on Deep Learning for Computer Vision

Few-shot font style transfer with multiple style encoders

Kejun ZHANG[1†], Rui ZHANG[1†], Yonglin WU[1], Yifei LI[2], Yonggen LING[3], Bolin WANG[1], Lingyun SUN[1] & Yingming LI[4*]

[1]*State Key Lab of CAD&CG, Zhejiang University, Hangzhou* 310027*, China;*
[2]*School of Software Technology, Zhejiang University, Hangzhou* 310027*, China;*
[3]*Robotics X, Tencent, Shenzhen* 518054*, China;*
[4]*College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou* 310027*, China*

Received 31 August 2021/Revised 2 December 2021/Accepted 15 January 2022/Published online 22 April 2022

Citation Zhang K J, Zhang R, Wu Y L, et al. Few-shot font style transfer with multiple style encoders. Sci China Inf Sci, 2022, 65(6): 160109, https://doi.org/10.1007/s11432-021-3435-8

Dear editor,

From the sketch to an encapsulated font, text font design is a labor-intensive and time-consuming process that relies heavily on the expertise of designers and usually takes months, even years, for a professional institution. This scenario is particularly prominent for glyph-rich scripts, such as Chinese, where each character is composed of varying numbers of highly complex structured components.

Recently, few-shot font style transfer, as a more practical scheme, has attracted considerable attention as it uses only a few reference font images to generate all of the characters of a new font and can produce more acceptable glyphs. A typical strategy is to separate the representations of style and content from the given reference font images, where encoder-mixer-decoder (EMD) [1] is one of the most representative studies and consists of a content encoder, a style encoder, a mixer, and a decoder. Furthermore, this structure has been widely used in subsequent studies, such as AGIS-Net [2] and lffont [3]. However, most of the previous few-shot font style transfer methods learn different style representations with a universal style encoder, which limits their representations of diverse styles. This effect is obvious when generating fonts for a style-rich script. Moreover, the structure of a single style encoder cannot easily deal with a special style transfer task, that is, the fusion of different fonts, where different characteristics from multiple character styles/sets are fused to form a new style/set. For example, a typeface usually has several fonts with different weights (such as extra bold, bold, regular, and light). Because the font with different weights cannot be easily generated by a simple method, such as graphic interpolation, the designers need to redesign many characters for every new weight to fulfill the aesthetic requirements.

In this study, we propose a novel flexible framework, that is, multi-style EMD (MS-EMD), to perform multiple style transfer through a many-to-many correspondence, which can be considered a direct multitask learning scenario. Specifically, we introduce multiple style encoders to enhance style representation learning and attempt to learn different representations for the corresponding styles. Compared with single encoder methods, MS-EMD enables fast font fusion by directly encoding different style/weight fonts with the setup of multiple style encoders. Thus, MS-EMD has considerable potential in real-world applications. Meanwhile, the use of multiple style encoders can be regarded as ensemble learning, where the average outputs of several encoders help improve the generalization performance of font generation.

Network. As shown in Figure 1(a), the proposed few-shot font style transfer network, denoted as MS-EMD thereafter, consists of multiple style encoders, a content encoder, a mixer, a decoder, and a discriminator. Given several sets of style reference images and a set of content reference images, the style encoders and the content encoder leverage the conditional dependence of styles and content to learn style/content representations. Then, the mixer combines the corresponding style and content representations using a bilinear model. Finally, the decoder generates the target images based on the combined representations. Furthermore, the discriminator is introduced to classify fake and real images. The details are presented in Appendix A.

Multi-style and content encoders. Generally, we use K style encoders E_{S_k} $(k = 1, 2, \ldots, K)$ and one content encoder E_C to extract the corresponding features. The input to each style encoder E_{S_k} is the style reference set $R_{S_{i_k}}$, and the input to the content encoder E_C is the content reference set R_{C_j}: $R_{S_{i_k}} = \{I_{i_k,j_1}, I_{i_k,j_2}, \ldots, I_{i_k,j_r}\}$ $(k = 1, 2, \ldots, K)$, $R_{C_j} = \{I_{i_1,j}, I_{i_2,j}, \ldots, I_{i_r,j}\}$, where I_{ij} is the image with style S_i and content C_j. Particularly, the target images $\{I_{i_k,j}\}$ are excluded from $\{R_{S_{i_k}}\}$ and R_{C_j}.

The style encoders E_{S_k} and content encoder E_C ex-

* Corresponding author (email: yingming@zju.edu.cn)
† Zhang K J and Zhang R have the same contribution to this work.

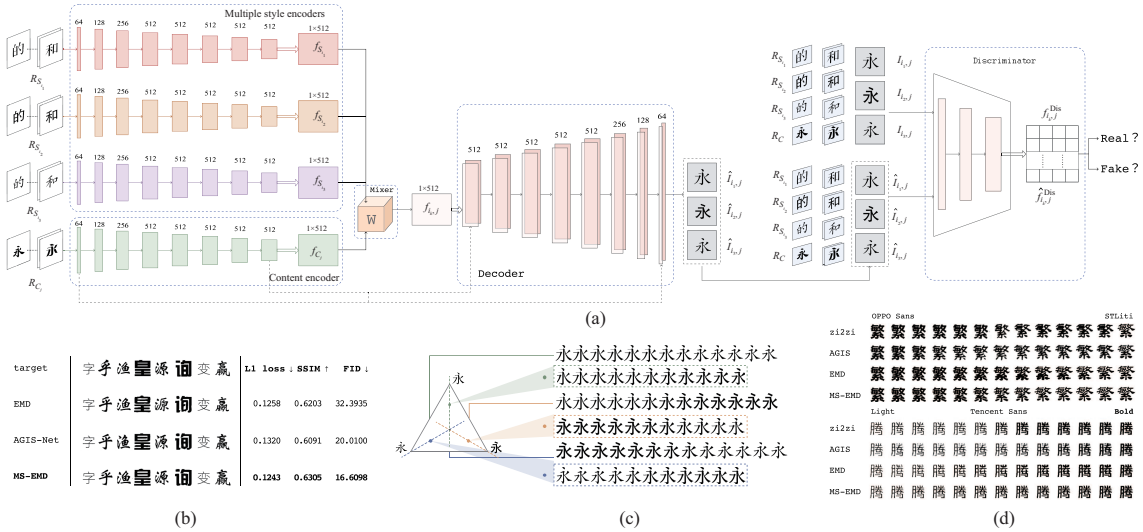

Figure 1 (Color online) (a) Architecture of the MS-EMD network; (b) few-shot font generation for a novel style; (c) font fusion for three styles: Heiti, Kaiti, and Songti; (d) font fusion for two styles (OPPO Sans and STLiti, and two weights of Tencent Sans).

tract features $f_{S_{i_k}}$ and f_{C_j} from $R_{S_{i_k}}$ and R_{C_j}, where $f_{S_{i_k}} = E_{S_k}(R_{S_{i_k}})$, $f_{C_j} = E_C(R_{C_j})$.

Mixer network. After extracting the features $\{f_{S_{i_k}}\}$ and f_{C_j}, they are combined by the mixer network. Because $\{f_{S_{i_k}}\}$ and f_{C_j} have no relations between each other, the features of content C and style S_{i_k} can be calculated in parallel as $f_{i_k,j} = f_{S_{i_k}} \times W \times f_{C_j}$, where W is a tensor with the size $F \times M \times F$, the size of $f_{S_{i_k}}$ and f_{C_j} is F, and $f_{i_k,j}$ is a feature vector with the size M.

Furthermore, when performing font fusion, the style feature $\{f_{S_{i_k}}\}$ is fused to obtain a new style feature vector $f_{\hat{S}_i} = \sum_{k=1}^{K} a_k \cdot f_{S_{i_k}}$, where a_k is the coefficient of $f_{S_{i_k}}$ and $\sum_{k=1}^{K} a_k = 1$. Then, we can obtain a new mixed representation $f_{\hat{i},j} = f_{\hat{S}_i} \times W \times f_{C_j}$, where $f_{\hat{i},j}$ is a new image with content C_j and fused style $\hat{S}_i$ from style $\{S_{i_k}\}$. By fusing the latent vectors using the aforementioned process, we achieve font fusion in real time.

Decoder network. The decoder network generates K images $\{\hat{I}_{i_k,j}\}$ from the features $f_{i_k,j}$, $\hat{I}_{i_k,j} = \mathrm{De}(f_{i_k,j})$ $(k = 1, 2, \ldots, K)$. For the fused feature vector $f_{\hat{i},j}$, a new image with fused style $\hat{I}_{\hat{i},j} = \mathrm{De}(f_{\hat{i},j})$ is produced.

Discriminator network. The discriminator network is designed to conduct the generation process of the decoder De. The discriminator network distinguishes real images $\{I_{i_k,j}\}$ and fake images $\{\hat{I}_{i_k,j}\}$ from the decoder to make the images conform to the conditions $R_{S_{i_k}}$ and R_{C_j}, $f^{\mathrm{Dis}}_{i_k,j} = \mathrm{Dis}(R_{C_j}, R_{S_{i_k}}, \hat{I}_{i_k,j})$, $f^{\mathrm{Dis}}_{i_k,j} = \mathrm{Dis}(R_{C_j}, R_{S_{i_k}}, I_{i_k,j})$, where the mean values of f^{Dis} and $\hat{f}^{\mathrm{Dis}}$ are used to judge TRUE or FALSE.

Loss. Our loss function has two parts: the L_1 loss and the adversarial loss, $\mathrm{Loss} = w_{L_1} \cdot L_1 + w_{\mathrm{adv}} \cdot L_{\mathrm{adv}}$.

Datasets. To evaluate the capability of MS-EMD to generate Chinese characters, we collected 243 fonts (each font has 1125 characters) from a publicly available database. The images are in 80×80 resolution. We randomly divided styles into a training set (75%, 182 fonts) and a validation set (25%, 61 fonts). Then, the entire dataset has two subsets: images with known styles and images with novel styles.

Experiments. We make qualitative and quantitative comparisons between several existing methods (i.e., zi2zi [4], EMD [1], and AGIS-Net [2]) and our method, that is, MS-EMD. Figure 1(b) shows the results of image generation for the novel style with ten reference images. MS-EMD obtains the best qualitative and quantitative results for L_1 loss, SSIM, and FID. As shown in Figures 1(c) and (d), MS-EMD generates characters clearly and correctly, while zi2zi, AGIS, and EMD have many flaws. More experimental details are presented in Appendix B.

Conclusion. In this study, we propose a novel font style transfer network, that is, MS-EMD, which has multiple style encoders to perfectly perform font generation and font fusion tasks simultaneously. Extensive experiments demonstrate the exceptional performance of MS-EMD on novel style font generation and font fusion.

Acknowledgements This work was supported by Key R&D Program of Zhejiang Province (Grant No. 2022C03126), Key Project of Natural Science Foundation of Zhejiang Province (Grant No. LZ19F020002), National Science and Technology Innovation 2030 Major Project of the Ministry of Science (Grant No. 2018AAA0100703), National Key R&D Program of China (Grant No. 2018YFB1403600), and Tencent Robotics X Lab Rhino-Bird Joint Research Program (Grant No. JR2020001 TEG&ZJU).

Supporting information Appendixes A and B. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhang Y X, Zhang Y, Cai W B. Separating style and content for generalized style transfer. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018. 8447–8455
2 Gao Y, Guo Y, Lian Z H, et al. Artistic glyph image synthesis via one-stage few-shot learning. ACM Trans Graph, 2019, 38: 185
3 Park S, Chun S, Cha J, et al. Few-shot font generation with localized style representations and factorization. In: Proceedings of AAAI Conference on Artificial Intelligence, 2021. 2393–2402
4 Tian Y. zi2zi: master Chinese calligraphy with conditional adversarial networks. 2017. https://github.com/kaonashi-tyc/zi2zi