DG-Font: Deformable Generative Networks for Unsupervised Font Generation

Yangchen Xie Xinyuan Chen* Li Sun Yue Lu

Shanghai Key Laboratory of Multidimensional Information Processing,
East China Normal University, 200241 Shanghai, China

Abstract

Font generation is a challenging problem especially for some writing systems that consist of a large number of characters and has attracted a lot of attention in recent years. However, existing methods for font generation are often in supervised learning. They require a large number of paired data, which is labor-intensive and expensive to collect. Besides, common image-to-image translation models often define style as the set of textures and colors, which cannot be directly applied to font generation. To address these problems, we propose novel deformable generative networks for unsupervised font generation (DG-Font). We introduce a feature deformation skip connection (FDSC) which predicts pairs of displacement maps and employs the predicted maps to apply deformable convolution to the low-level feature maps from the content encoder. The outputs of FDSC are fed into a mixer to generate the final results. Taking advantage of FDSC, the mixer outputs a high-quality character with a complete structure. To further improve the quality of generated images, we use three deformable convolution layers in the content encoder to learn style-invariant feature representations. Experiments demonstrate that our model generates characters in higher quality than state-of-art methods. The source code is available at https://github.com/ecnuycxie/DG-Font.

1. Introduction

Every day, people consume a massive amount of texts for information transfer and storage. As the representation of texts, the font is closely related to our daily life. Font generation is critical in many applications, *e.g.*, font library creation, personalized handwriting, historical handwriting imitation, and data augmentation for optical character recognition and handwriting identification. Traditional font library creating methods heavily rely on expert designers by drawing each glyph individually, which is especially expensive and labor-intensive for logographic languages such as Chinese (more than 60,000 characters), Japanese (more than 50,000 characters), and Korean (11,172 characters).

Figure 1. **Unsupervised font generation results.** The reference calligraphy is a Tang poem written by a calligrapher, and imitation result is another famous Tang poem generated from our model which are with rich details, such as stroke tips, joined-up writing, and thickness of strokes.

Recently, the development of convolutional neural networks enables automatic font generation without human experts. There have been some attempts to explore font generation and achieve promising results. [49, 1, 18] utilize deep neural networks to generate entire sets of letters for certain alphabet languages. Two notable projects, “Rewrite" [40] and “zi2zi" [61], generate logographic language characters by learning a mapping from one style to another with thousands of paired characters. After that, EMD [58] and SA-VAE [44] design neural networks to separate the content and style representation, which can extend to generate character of new styles or contents. However, these methods are

*Corresponding author: xychen@cee.ecnu.edu.cn

in supervised learning and required a large amount of paired training samples.

Some other methods exploit auxiliary annotations (*e.g*., strokes, radicals) to facilitate high-quality font generation. For example, [30] utilizes labels for each stroke to generate glyphs by writing trajectories synthesis. [26] employ the radical decomposition (*e.g*., radicals or sub-glyphs) of characters to achieve font generation for certain logographic language. DM-Font [7] and its improved version LF-Font [39] propose disentanglement strategies to disentangle complex glyph structures, which help capture local details in rich text design. However, these methods rely on prior knowledge and can only apply to specific writing systems. Some labels such as the stroke skeleton can be estimated by algorithms, but the estimation error would decrease the generated quality. Also, these methods still require thousands of paired data and annotated labels for training. Recently, there are some attempts [19, 9] for unsupervised font generation. [9] introduces a novel module that transfers the features across sequential DenseNet blocks [23]. [19] proposes a fast skeleton extraction method to obtain the skeleton of characters, and then utilize the extracted skeleton to facilitate font generation.

For the problem of image-to-image translation, a series of works in unsupervised learning have been proposed by combining adversarial training [32, 54] with consistent constraints [59, 47, 3]. FUNIT [33] maps an image of a source class to an analogous image of a target class by leveraging a few target class images. They extract the style feature of the target class images and employ adaptive instance normalization (AdaIN) [25] to combine the content and the style features. However, these image-to-image translation methods cannot be directly applied to font generation tasks. Although consistent constraints preserve the structure of a content image, they still encounter some problems for font generation (*e.g*., blurry, missing some strokes). Also, they usually define the style as the set of textures and colors. The AdaIN-based methods transfer style by aligning feature statics, which tends to transform texture and color, which is not suitable to transform local style patterns (*e.g*., geometric deformation) for the font. Moreover, [9, 19] achieve unsupervised font generation by learning a mapping between two fonts directly, they also ignore the geometric deformation for the font. To learn the mapping across geometry variations, [20] introduces a discriminator with dilated convolutions as well as a multi-scale perceptual loss that is able to represent error in the underlying shape of objects. [52] disentangles image space into a Cartesian product of the appearance and the geometry latent spaces.

Compelled by the above observations, we propose a novel deformable generative model for unsupervised font generation (DG-Font). The proposed method is designed to deform and transform the character of one font to another by leveraging the provided images of the target font. The proposed DG-Font separates style and content respectively and then mix two domain representations to generate target characters. We introduce a feature deformation skip connection (FDSC) which predicts pairs of displacement maps and employs the predicted maps to apply deformable convolution to the low-level feature maps from the content encoder. The outputs of FDSC are fed into a mixer to generate the final results. To distinguish different styles, we train our model with a multi-task discriminator, which ensures that each style can be discriminated independently. In addition, another two reconstruction losses are adopted to constrain the domain-invariant characteristics between generated images and content images.

The feature deformation skip connection (FDSC) module is used to transform the low-level feature of content images, which preserves the pattern of character (*e.g*., strokes and radicals). Different from the image-to-image translation problem that defines style as a set of texture and color, the style of font is basically defined as geometric transformation, stroke thickness, tips, and joined-up writing pattern. For two fonts with the same content, they usually have correspondence for each stroke. Taking advantage of the spatial relationship of fonts, the feature deformation skip connection (FDSC) is used to conduct spatial deformation, which effectively ensures the generated image to have complete structures.

Extensive experiments demonstrate that our model achieves comparable results to the state-of-the-art font generation methods. Besides, results show that our model is able to extend to generate unseen style character.

2. Related works

2.1. Image-to-Image Translation

The purpose of image-to-image translation is to learn a mapping from an image in the source domain to the target domain. Image-to-image translation has been applied in many fields such as artistic style transfer [31, 56], semantic segmentation [43, 38], image animation [50, 53, 15], object transfiguration [12], and video frames generation [8, 13, 17] *et al*. Pix2pix [27] is the first model proposed for image-to-image translation based on conditional GAN [37]. To achieve unsupervised image-to-image translation, a lot of works [34, 59, 6, 42] have been proposed, where CycleGAN [59] introduces a cycle consistency between source and target domain to discover the relationship of samples between two domain. However, above-mentioned methods can only translate from one domain to another specific domain. To tackle this problem, recent works [11, 33, 2, 5] are proposed to simultaneously generate multiple style outputs given the same input. Gated-GAN [11] proposes a gated transformer to transfer multiple styles in a single model.

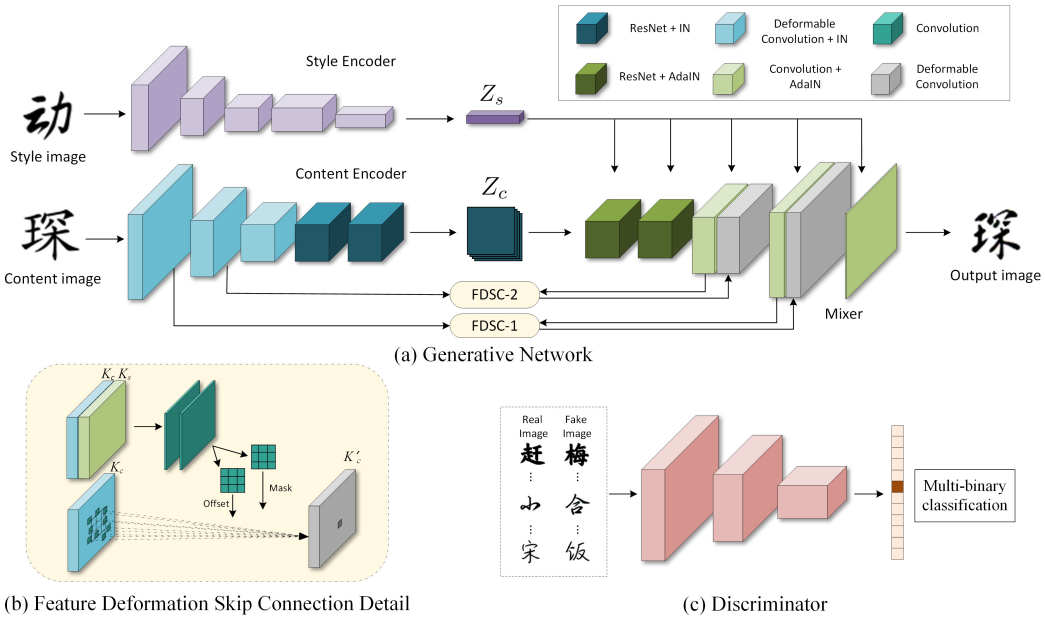

Figure 2. **Overview of the proposed method**. a) Overview of our generative network. The Style/content encoder maps style/content image to style/content representation Z_s/Z_c. FDSC-1 and FDSC-2 have the same architecture and apply transformation convolution to the low-level feature from the content encoder and inject the results into the mixer. The mixer generates the output image. b) A detailed illustration of the FDSC module. c) The discriminator output a binary vector, where each element indicates a binary classification to distinguish between generated and real images.

FUNIT [33] encodes content image and class image respectively, and combines them with AdaIN [25]. TUNIT [2] further introduce a guiding network as an unsupervised domain classifier to automatically produce a domain label of a given image. DUNIT [5] extract separate representations for the global image and for the instances to preserve the detailed content of object instances.

2.2. Font Generation

Font generation aims to automatically generate characters in a specific font and create a font library. Recent studies have employed image translation methods for font generation. “Zi2zi" [61] and “Rewrite" [40] implement font generation on the basis of GAN [21] with thousands of character pairs for strong supervision. After that, a series of models are proposed to improve the generated quality based on zi2zi [61]. PEGAN [45] sets up a multi-scale image pyramid to pass information through refinement connections. HAN [10] improves zi2zi by designing a hierarchical loss and skip connection. AEGG [36] adds an additional network to refine the training process. DC-Font [29] introduces a style classifier to get a better style representation. However, all the above methods are in supervised learning and require a large number of paired data.

In addition to the paired data, lots of methods employ auxiliary annotations (*e.g.*, stroke and radical decomposition) to further improve the generation quality. SA-VAE [44] disentangles the style and content as two irrelevant domains with encoding Chinese characters into high-frequency character structure configurations and radicals. CalliGAN [51] further decomposes characters into components and offers low-level structure information including the order of strokes to guide the generation process. RD-GAN [26] proposes a radical extraction module to extract rough radicals which can improve the performance of discriminator and achieves the few-shot Chinese font generation. Also, some other attempts have been made in Chinese character generation by adopting skeleton/stroke extraction algorithm [19, 30]. However, they need extra annotations or algorithms to guide font generation; while the estimation error would decrease the generation performance. In contrast, our proposed model, DG-Font, can generate images in an unsupervised way without other annotations.

2.3. Deformable Convolution

CNNs have inherent limitations in modeling geometric transformations due to the fixed kernel configuration. To enhance the transformation modeling capability of CNNs, [16] proposes the deformable convolutional layer. It augments the spatial sampling locations in the modules with additional offsets. The deformable convolution has been applied to address several high-level vision tasks, such as

object detection [4, 16, 60] video object detection [14] sampling, semantic segmentation [60], and human pose estimation [46]. Recently, some methods attempt to apply deformable convolution in the image generation tasks. TDAN [48] addresses video super-resolution task by using deformable convolution to align two continuous frames and output a high-resolution frame. [55] synthesizes novel view images by deformable convolution given the view condition vectors. In our proposed DG-Font, offsets are estimated by a learned latent style code.

3. Methods

3.1. overview

Given a content image I_c and a style image I_s, our model aims to generate the character of the content image with the font of the style image. As illustrated in Fig. 2, the proposed generative network consists of a style encoder, a content encoder, a mixer, and two feature deformation skip connection (FDSC) modules. The architecture of the style encoder and discriminator is simplified in Fig. 2. The detailed architecture is shown in Appendix A. **The style encoder** is designed to learn the style representation from input images. Specifically, the style encoder takes a style image as the input and maps it to a style latent vector Z_s. **The content encoder** is introduced to extract the structure feature of the content images. The content encoder maps the content image into a spatial feature map Z_c. The content encoder module is made of three deformable convolution layers followed by two residual blocks. The introduced deformable convolution layer enables the content encoder to produce style-invariant features for images with the same content. **The mixer** aims to output characters by mixing the content feature representations Z_c and style feature representations Z_s. AdaIN [25] is adopted to inject the style feature to the mixer. Besides, **the feature deformation skip connection modules** transfers the deformed low-level feature from the content encoder to the mixer. Details are described in Sec 3.2.

When character images are generated from the generative network, **a multi-task discriminator** is adopted to conduct discrimination for each style simultaneously. For each style, the output of the discriminator is a binary classification whether the input image is a real image or a generated image. As there are several different styles of fonts in the training set, the discriminator outputs a binary vector whose length is the number of styles.

3.2. Feature Deformation Skip Connection

As illustrated in Fig. 3, there lies in a geometric deformation of two fonts for a character and exists a correspondence for each stroke. Compelled this observation, we propose a feature deformation skip connection (FDSC) module to apply geometric deformation convolution to the content image in the feature space and directly transfer the deformation low-level feature to the mixer. Specifically, the module predicts offsets based on the guidance code to instruct the deformable convolution layer performing a geometric transformation on the low-level feature. As demonstrated in Fig. 2, the input of FDSC module is a concatenation of two feature maps: a feature map K_c extracted from the content image and a style guidance map K_s. K_s is extracted from the mixer after injecting the style code Z_s. The module estimates sampling parameters after applying convolution to the concatenation of K_s and K_c:

$$\Theta = f_\theta(K_s, K_c). \tag{1}$$

Here, f_θ refers to a convolution layer, and $\Theta = \{\Delta p_k, \Delta m_k \mid k = 1, \cdots, |\mathcal{R}|\}$ refers to the offsets and mask of the convolution kernel, where $\mathcal{R} = \{(-1, -1), (-1, 0), \cdots, (0, 1), (1, 1)\}$ indicates a regular grid of a 3×3 kernel. Under the guidance of sampling parameter Θ, a geometrically deformed feature map $K_c^{'}$ is obtained from Θ and K_c based on deformable convolution $f_{DC}(\cdot)$:

$$K_c^{'} = f_{DC}(K_c, \Theta). \tag{2}$$

Specifically, for each position p on the output $K_c^{'}$, the deformable convolution $f_{DC}(\cdot)$ is applied as follow:

$$K_c^{'}(p) = \sum_{k=1}^{\mathcal{R}} w(p_k) \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k, \tag{3}$$

where the $w(p_k)$ indicates the weight of the deformable convolution kernel at k-th location. The convolution is operated on the irregular positions ($p_k + \Delta\ p_k$) where $\Delta\ p_k$ may be fractional. Followed [16], the operation is implemented by using bilinear interpolation. At last, the output of feature deformation skip connection module is fed to the mixer and $K_c^{'}$ is then concatenated with K_s like a common used skip connection [41].

Deformable convolution introduces 2D offsets to the regular grid sampling locations in the standard convolution. It enables free form deformation of the sampling grid. There are lots of areas of the same color in character images, such as background color and character color. By using the deformable convolution, an area can be related to any other area with the same color. It is difficult to optimize the non-unique solution. To efficiently use our FDSC module, we impose a constrain on the offsets Δp. We introduce the constrain in detail in the next subsection. Section 4.4 demonstrates the visualization of the offsets Δp.

Our FDSC module aims to deform the spatial structure of the content image in the feature space. It is crucial to select which level of features to be transformed. As we know, low-level features contain more spatial information than high-level features. In our model, we employ the feature maps

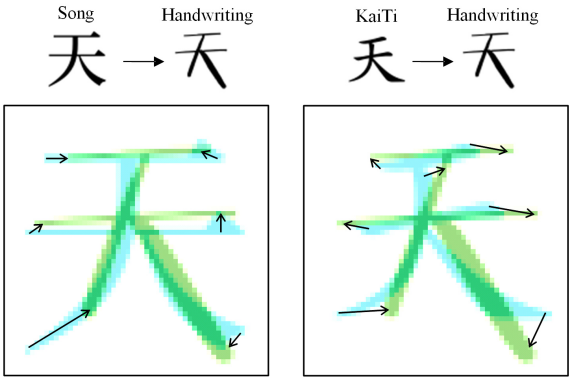

Figure 3. **The geometric deformation of two fonts for a character**. We employ the character "Tian" to compare a handwritten style with the fonts of Kaiti and Song. There is a correspondence for each stroke between two fonts.

after the first and second convolution layer as input to the FDSC module. Appendix B demonstrates the analysis of the performance of the model with different numbers of the FDSC module.

3.3. Loss Function

Our model aims to achieve automatic font generation via an unsupervised method. To this end, we adopt four losses: 1) adversarial loss is used to produce realistic images. 2) content consistent loss is introduced to encourage the content of the generated image to be consistent with the content image; 3) image reconstruction loss is used to maintain the domain-invariant features; 4) deformation offset normalization is designed to prevent excessive offsets of the FDSC module. We introduce the formula of each loss and the full objective in this section.

Adversarial loss: the proposed network aims to generate plausible images by solving a minimax optimization problem. The generative network G tries to fool discriminator D by generating fake images. The adversarial loss penalty the wrong judgement when real/generated images are input to discriminator.

$$\begin{aligned}\mathcal{L}_{adv} = \max_{D_s}\min_{G}\mathbb{E}_{I_s\in P_s, I_c\in P_c}[\log D_s(I_s) \\ + \log(1 - D_s(G(I_s, I_c)))],\end{aligned} \tag{4}$$

where $D_s(\cdot)$ denotes the logit from the corresponding style of discriminator's output.

Content consistent loss: adversarial loss is adopted to help the model to generate a realistic style while ignoring the correctness of the content. To prevent mode collapse and ensure that the features extracted from the same content can be content consistent after the content encoder f_c, we impose an content consistent loss here:

$$\mathcal{L}_{cnt} = \mathbb{E}_{I_s\in P_s, I_c\in P_c}\left\|Z_c - f_c(G(I_s, I_c))\right\|_1. \tag{5}$$

L_{cnt} ensures that given a source content image I_c and corresponding generated images, their feature maps are consistent after content encoder f_c.

Image Reconstruction loss: To ensure that the generator can reconstruct the source image I_c when given with its origin style, we impose an reconstruction loss:

$$\mathcal{L}_{img} = \mathbb{E}_{I_c\in P_c}\left\|I_c - G(I_c, I_c)\right\|_1. \tag{6}$$

The objective helps preserve domain-invariant characteristics (*e.g.*, content) of its input image I_c.

Deformation offset normalization: The deformable offsets enable free form deformation of the sampling grid.As there are lots of areas of the same color between input images and generated images (such as background color and character color), it leads to a non-unique solution which is difficult to optimize. Meanwhile, the font generation focus on the stroke relationship between content character image and target character image, such as the thickness and tips of stroke. However, given images with the same content but different style, the position of the same stroke in these two images are close. To efficiently use this deformable convolutional network, we impose a constrain on the offsets Δp:

$$\mathcal{L}_{offset} = \frac{1}{|\mathcal{R}|}\left\|\Delta p\right\|_1, \tag{7}$$

where Δp denotes offsets of the deformable convolution kernel, $|\mathcal{R}|$ denotes the number of the convolution kernel.

Overall Objective loss: Combining all the abovementioned loss, we have the overall loss function for training our proposed framework:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{img}\mathcal{L}_{img} + \lambda_{cnt}\mathcal{L}_{cnt} + \lambda_{offset}\mathcal{L}_{offset}, \tag{8}$$

where λ_{adv}, λ_{img}, λ_{cnt}, λ_{offset} are hyperparameters to control the weight of each loss function. In our model, the generative network aims to minimize the overall object loss, while the discriminator aims to maximize it.

4. Experiments

In this section, we evaluate our proposed model for the Chinese font generation task. We first introduce our dataset. After that, the results of our experiments are shown to verify the advantages of our model. More implementation details are shown in Appendix A.

4.1. Dataset

To evaluate our model for Chinese font generation, we collect a dataset that contains 410 fonts (styles) including handwritten fonts and printed fonts, each of which has 990 commonly used Chinese characters (content). The dataset is randomly partitioned into a training set and testing set. The

Methods	one-to-many	training	L1 loss	RMSE	SSIM	LPIPS	FID
Seen fonts							
EMD [58]	✓	paired	0.0538	0.1955	0.7676	0.1036	89.65
Zi2zi [61]	✓	paired	**0.0521**	0.1802	0.7789	0.1065	142.23
Cycle-GAN [59]	×	unpaired	0.0863	0.2555	0.6392	0.1825	175.24
GANimorph [20]	×	unpaired	0.0563	**0.1759**	**0.7808**	0.1403	72.89
FUNIT [33]	✓	unpaired	0.0807	0.2510	0.6669	0.1216	53.77
Ours	✓	unpaired	0.0562	0.1994	0.7580	**0.0814**	**46.15**
Unseen fonts							
EMD [58]	✓	paired	0.0430	0.1755	0.7849	0.1255	82.53
FUNIT [33]	✓	unpaired	0.0588	0.2089	0.7417	0.1125	59.98
Ours	✓	unpaired	**0.0414**	**0.1709**	**0.7982**	**0.0867**	**50.29**

Table 1. **Quantitative evaluation on the whole dataset**. We evaluate the methods on seen and unseen font sets. The bold number indicates the best.

(a) Easy cases (*i.e.*, non-cursive writing).

(b) Challenging cases (*i.e.*, cursive writing).

Figure 4. Comparisons to the stat-of-art methods for font generation.

training set contains 400 fonts, and each font contains 800 characters. The testing set consists of two parts. One part is the remaining 190 characters of the 400 fonts. Another part is the remaining 10 fonts which are used to test the generalization ability to unseen fonts.

4.2. Comparison with State-of-art Methods

In this subsection, we compare our model with the following methods for Chinese font generation: 1) Cycle-GAN [59]: Cycle-GAN consists of two generative net-

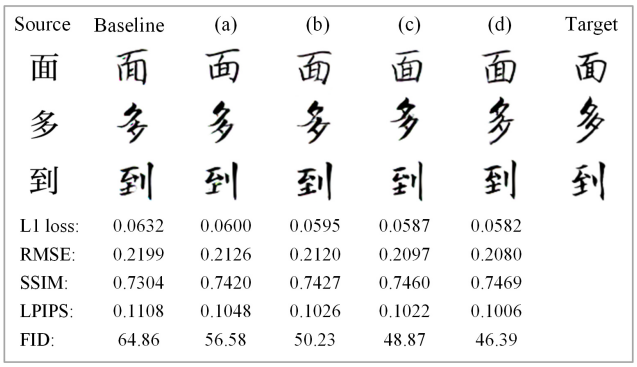

Figure 5. **Effect of different components in our method**. We add different parts into our baseline successively. (a) Replace the first three convolution layers of content encoder with deformable convolution layers; (b) add the FDSC-1 module (without normalization); (c) impose normalization on FDSC-1 module; (d) add the FDSC-2 module (full model).

works which can translate images from one domain to another using a cycle consistency loss. Cycle-GAN is also an unsupervised learning method; 2) EMD [58]: EMD employs an encoder-decoder architecture, and separates style/content representations. EMD is optimized by L1 distance loss between ground-truth and generated images; 3) Zi2zi [61]: Zi2zi is a modified version of pix2pix [27] model, it achieves font generation and uses Gaussian Noise as category embedding to achieve multi-style transfer. Zi2zi still requires paired data; 5) GANimorph [20]: GANimorph adopts the cyclic image translation framework like Cycle-GAN and introduce a discriminator with dilated convolutions to get a more context-aware generator; 6) FUNIT [33]: FUNIT is an unsupervised image-to-image translation model which separates content and style of natural animal images and combine them with adaptive instance normalization (AdaIN) layer.

For a fair comparison, we employ the font of Song as the source font which is commonly used in font generation task [58, 26]. Our model, EMD, TUNIT are trained with 400 fonts. For Cycle-GAN and GANimorph, they can only train one paired translation at once, hence we train 399 models of Cycle-GAN individually for each target style. In our experiments, we find that the model of Zi2zi trained with 400 fonts performs worse than trained with two fonts. As a result, we train 399 models for Zi2zi for each target style.

Quantitative comparison. The quantitative results are shown in Table 1. In the experiments, DG-Font is comparable to compared methods in pixel-level evaluation metrics, *e.g.*, L1 loss, RMSE, SSIM. It is noted that these metrics focus on pixel-wise between generated image and ground-truth and ignore the feature similarity which is closer to human perceptions. In perceptual-level metrics FID [22] and LPIPS [57], we can observe that DG-Font outperforms the compared methods and reaches the state-of-the-art performance for both seen fonts and unseen fonts.

Method	L1 loss	RMSE	SSIM	LPIPS	FID
SC	0.0641	0.2212	0.7252	0.1114	46.88
FDSC	**0.0582**	**0.2080**	**0.7469**	**0.1006**	**46.39**

Table 2. **Comparison with skip-connection (SC) proposed by U-Net** [41] . We replace two FDSC modules with skip-connections and then compare the new model with the full model of DG-Font.

Qualitative comparison. In order to verify the capability of deforming and transforming source character patterns (*e.g.*, stroke, skeleton), two kinds of visual comparisons are displayed in Fig. 4. First, we compare DG-Font to other methods with relative simple fonts that are close to printed fonts with no cursive writing. As demonstrated in Fig. 4(a), Cycle-GAN can only generate parts of characters or sometimes unreasonable structures. Characters generated by Zi2zi EMD, and GANimorph can maintain a complete structure, but they are usually vague. FUNIT can generate characters with a clear background but the generated characters lose their structure to some degree. DG-Font is able to generate character close to the target well. In contrast to fonts in Fig. 4(a), fonts in Fig. 4(b) are more challenging for the rich details and joined-up writing. We can observe that Cycle-GAN, EMD, Zi2zi and GANimorph can hardly generate characters under challenging cases. While FUNIT maintains the ability to generate characters with incomplete structure, but the skeleton of generated character is not well transformed. Our proposed DG-Font can not only generate characters with complete structure but also learn joined-up writing.

4.3. Ablation Study

In this part, we add different parts into the model successively and analyze the influence of each part, including deformable convolution, feature deformation skip connection and deformable offset normalization. We conduct the ablation study on the data set of 187 handwritten fonts. Our baseline is the models that replace deformable convolution with normal convolution and remove FDSC modules. Qualitative and quantitative comparisons are shown in Fig. 5.

1) **Effectiveness of deformable convolution in the content encoder.** Fig. 5(a) shows the results by replacing the first three convolution layers of the content encoder with deformable convolution layers. We can see that the quantitative results improve obviously in terms of L1 loss, RMSE, and SSIM. This indicates that deformable convolution layers in the content encoder effectively help improve the performance of our model.

2) **The influence of the FDSC module.** In this part, we add an FDSC module (without offset normalization in Eq. 7) that connects the features after the first layer and penulti-

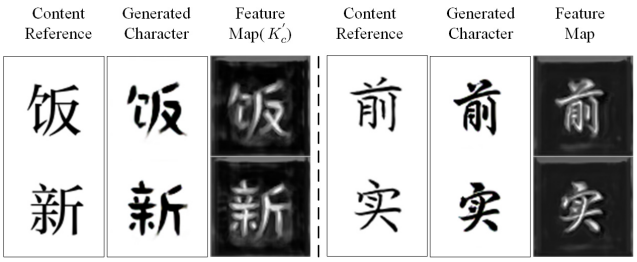

Figure 6. **Feature visualization**. We visualize the features $K_c^{'}$ generated from the FDSC-1 module. For each case, from left to right: content reference characters, the corresponding generated characters, the visualization of feature maps. For feature map images, the whiter the area, the larger the activation value.

mate layer. Results are shown in Fig. 5(b). Comparing with Fig. 5(a), we observe that the generated characters preserve more structure information and are able to reconstruct the complete structure of characters.

3) **Effectiveness of deformable offset constrain.** We investigate the impact of deformable offset normalization by comparing FDSC module without and with offset normalization. As shown in Fig. 5(b) and (c), adding offset normalization helps the model generate images whose style become more similar to the target.

4) **Effectiveness of two FDSC modules.** Fig. 5 (d) shows the results of our full model with two FDSC modules. It is noted that the generated images get more details, less noise, and achieves better quantitative results.

In addition, we compare our proposed FDSC module with common used skip-connection [41, 58, 30, 51] proposed by U-Net [41]. Skip-connection is often adopted to transfer feature maps with different resolution directly from encoder to decoder, which is effective in semantic segmentation [28, 35] and photo-to-art [24] tasks whose content of inputs and outputs share the same structure. However, the font generation requires a geometric deformation between content inputs and the corresponding generated images in structure. To compare FDSC module with skip-connection, We replace two FDSC modules with skip-connection in our proposed DG-Font network. The comparison results are shown in Table 2. We can observe that models with FDSC modules outperform models with skip-connection, which prove the advantage of FDSC.

4.4. Visualization

In order to show the effectiveness of FDSC, we visualize the feature maps generated by the FDSC-1 module. As shown in Fig. 6, the feature maps $K_c^{'}$ preserve the pattern of characters well, which helps generate a character with complete structure. On the other hand, we can observe that the FDSC module effectively transform features extracted from the content encoder.

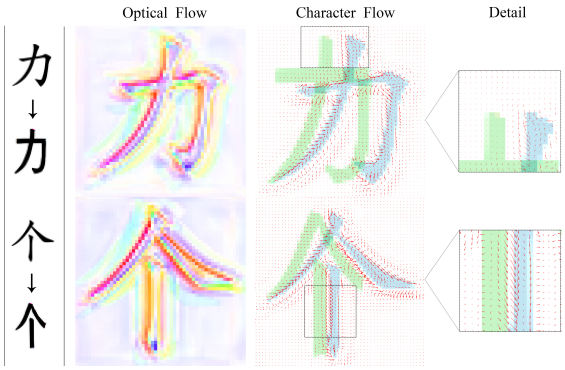

Figure 7. **The visualization of learned offsets**. First column: source image and generated image. Second column: the optical flow displays the estimated offsets Δp. Third column: character flow visualized the offsets Δp. Forth column: zoomed-in details. Source and generated images are in blue and green respectively.

In addition, we visualize the learned offsets from the FDSC-1 module using optical flow and character flow respectively. To visualize the offsets clearly, the kernel of deformable convolution in the FDSC module is set to 1×1.As demonstrated in Fig. 7, we observe that the learned offsets mainly affect the character region. The offsets value of the background tends to zero, which proves the usefulness of the proposed offset loss Eq. 7. In character flow, we can see that most of the offset vectors point from the stroke in target characters to the corresponding source stroke. The results show that in the convolution process, the sampling locations of target characters tend to shift to corresponding locations in source character by the learned offsets.

5. Conclusion

In this paper, we propose an effective unsupervised font generation model which is capable to generate realistic characters without paired images and can extend to unseen font well. To ensure the integration of generated characters, we propose a Feature Deformation Skip Connection (FDSC) module to transfer the deformable low-level spatial information to the mixer. Besides, we employ deformable convolution layers in content encoder to learn style-invariant feature representations. Extensive experiments on Chinese font generation verify the effectiveness of our proposed model.

Acknowledgements This work was partly supported by the National Key Research and Development Program of China under No. 2020AAA0107903, the China Postdoctoral Science Foundation under No. 2020M681237, and the Science and Technology Commission of Shanghai Municipality under No.19511120800, No.19ZR1415900, No.18DZ2270800.

References

[1] Samaneh Azadi, Matthew Fisher, Vladimir G. Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content GAN for few-shot font style transfer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7564–7573. IEEE Computer Society, 2018.

[2] Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. *CoRR*, abs/2006.06500, 2020.

[3] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 752–762, 2017.

[4] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pages 342–357. Springer, 2018.

[5] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. DUNIT: detection-based unsupervised image-to-image translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4786–4795. IEEE, 2020.

[6] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 95–104. IEEE Computer Society, 2017.

[7] Junbum Cha, Sanghyuk Chun, Gayoung Lee, Bado Lee, Seonghyeon Kim, and Hwalsuk Lee. Few-shot compositional font generation with dual memory. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIX*, volume 12364 of *Lecture Notes in Computer Science*, pages 735–751. Springer, 2020.

[8] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942, 2019.

[9] Bo Chang, Qiong Zhang, Shenyi Pan, and Lili Meng. Generating handwritten chinese characters using cyclegan. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 199–207. IEEE Computer Society, 2018.

[10] Jie Chang, Yujun Gu, Ya Zhang, and Yan-Feng Wang. Chinese handwriting imitation with hierarchical generative adversarial network. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*, page 290. BMVA Press, 2018.

[11] Xinyuan Chen, Chang Xu, Xiaokang Yang, Li Song, and Dacheng Tao. Gated-gan: Adversarial gated networks for multi-collection style transfer. *IEEE Trans. Image Process.*, 28(2):546–560, 2019.

[12] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. Attention-gan for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 164–180, 2018.

[13] X. Chen, C. Xu, X. Yang, and D. Tao. Long-term video prediction via criticization and retrospection. *IEEE Transactions on Image Processing*, 29:7090–7103, 2020.

[14] Zhu Chen, Weihai Li, Chi Fei, Bin Liu, and Nenghai Yu. Spatial-temporal feature aggregation network for video object detection. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 1858–1862. IEEE, 2020.

[15] Zhuo Chen, Chaoyue Wang, Bo Yuan, and Dacheng Tao. Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[16] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 764–773. IEEE Computer Society, 2017.

[17] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[18] Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. Scrabblegan: Semi-supervised varying length handwritten text generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4323–4332. IEEE, 2020.

[19] Yiming Gao and Jiangqin Wu. Gan-based unpaired chinese character image translation via skeleton transformation and stroke rendering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 646–653. AAAI Press, 2020.

[20] Aaron Gokaslan, Vivek Ramanujan, Daniel Ritchie, Kwang In Kim, and James Tompkin. Improving shape deformation in unsupervised image-to-image translation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, volume 11216 of *Lecture Notes in Computer Science*, pages 662–678. Springer, 2018.

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a

two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6626–6637, 2017.

[23] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[24] Siyu Huang, Haoyi Xiong, Tianyang Wang, Qingzhong Wang, Zeyu Chen, Jun Huan, and Dejing Dou. Parameter-free style projection for arbitrary style transfer. *CoRR*, abs/2003.07694, 2020.

[25] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1510–1519. IEEE Computer Society, 2017.

[26] Yaoxiong Huang, Mengchao He, Lianwen Jin, and Yongpan Wang. RD-GAN: few/zero-shot chinese character style transfer via radical decomposition and rendering. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 156–172. Springer, 2020.

[27] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5967–5976. IEEE Computer Society, 2017.

[28] Simon Jégou, Michal Drozdzal, David Vázquez, Adriana Romero, and Yoshua Bengio. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1175–1183. IEEE Computer Society, 2017.

[29] Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. Dcfont: an end-to-end deep chinese font generation system. In *SIGGRAPH Asia 2017 Technical Briefs, Bangkok, Thailand, November 27 - 30, 2017*, pages 22:1–22:4. ACM, 2017.

[30] Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. Scfont: Structure-guided chinese font generation via deep stacked networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 4015–4022. AAAI Press, 2019.

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pages 694–711. Springer, 2016.

[32] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1857–1865, 2017.

[33] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 10550–10559. IEEE, 2019.

[34] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 469–477, 2016.

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440. IEEE Computer Society, 2015.

[36] Pengyuan Lyu, Xiang Bai, Cong Yao, Zhen Zhu, Tengteng Huang, and Wenyu Liu. Auto-encoder guided GAN for chinese calligraphy synthesis. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 1095–1100. IEEE, 2017.

[37] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[38] Luigi Musto and Andrea Zinelli. Semantically adaptive image-to-image translation for domain adaptation of semantic segmentation. *BMVC*, 2020.

[39] Song Park, Sanghyuk Chun, Junbum Cha, Bado Lee, and Hyunjung Shim. Few-shot font generation with localized style representations and factorization. *CoRR*, abs/2009.11042, 2020.

[40] Rewrite. https://github.com/kaonashi-tyc/rewrite.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[42] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2242–2251. IEEE Computer Society, 2017.

[43] Samarth Shukla, Luc Van Gool, and Radu Timofte. Extremely weak supervised image-to-image translation for semantic segmentation. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 3368–3377. IEEE, 2019.

[44] Danyang Sun, Tongzheng Ren, Chongxuan Li, Hang Su, and Jun Zhu. Learning to write stylized chinese characters by reading a handful of examples. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 920–927. ijcai.org, 2018.

[45] Donghui Sun, Qing Zhang, and Jun Yang. Pyramid embedded generative adversarial network for automated font generation. In *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018*, pages 976–981. IEEE Computer Society, 2018.

[46] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 536–553. Springer, 2018.

[47] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *CoRR*, abs/1611.02200, 2016.

[48] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: temporally-deformable alignment network for video super-resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3357–3366. IEEE, 2020.

[49] Paul Upchurch, Noah Snavely, and Kavita Bala. From A to Z: supervised transfer of style and content using deep neural network generators. *CoRR*, abs/1603.02003, 2016.

[50] Chaoyue Wang, Chang Xu, and Dacheng Tao. Self-supervised pose adaptation for cross-domain image animation. *IEEE Transactions on Artificial Intelligence*, 1(1):34–46, 2020.

[51] Shan-Jean Wu, Chih-Yuan Yang, and J. Hsu. Calligan: Style and structure-aware chinese calligraphy character generator. *ArXiv*, abs/2005.12500, 2020.

[52] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8012–8021. Computer Vision Foundation / IEEE, 2019.

[53] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, pages 1618–1625, 2017.

[54] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[55] Mingyu Yin, Li Sun, and Qingli Li. Novel view synthesis on unpaired data by conditional deformable variational auto-encoder. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII*, volume 12373 of *Lecture Notes in Computer Science*, pages 87–103. Springer, 2020.

[56] Hang Zhang and Kristin J. Dana. Multi-style generative network for real-time transfer. In *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11132 of *Lecture Notes in Computer Science*, pages 349–365. Springer, 2018.

[57] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 586–595. IEEE Computer Society, 2018.

[58] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating style and content for generalized style transfer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8447–8455. IEEE Computer Society, 2018.

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251. IEEE Computer Society, 2017.

[60] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets V2: more deformable, better results. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9308–9316. Computer Vision Foundation / IEEE, 2019.

[61] Zi2zi. https://github.com/kaonashi-tyc/zi2zi.