Journal of
Computer Science & Technology

Indexed in: SCIE, EI, INSPEC, CBST, DBLP, etc.

Sponsored by: ICT, CAS & CCF

A Novel Three-Staged Generative Model for Skeletonizing Chinese Characters with Versatile Styles

Tian Ye–Chuan, Xu Song–Hua, Sylla Cheickna

View online: http://doi.org/10.1007/s11390–023–1337–8

JCST Homepage: https://jcst.ict.ac.cn
SPRINGER Homepage: https://www.springer.com/journal/11390
E–mail: jcst@ict.ac.cn
Online Submission: https://mc03.manuscriptcentral.com/jcst
Twitter: JCST_Journal
LinkedIn: Journal of Computer Science and Technology

Tian YC, Xu SH, Sylla C. A novel three-staged generative model for skeletonizing Chinese characters with versatile styles. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 38(6): 1250-1271 Nov. 2023. DOI: 10.1007/s11390-023-1337-8

A Novel Three-Staged Generative Model for Skeletonizing Chinese Characters with Versatile Styles

Ye-Chuan Tian[1] (田业川), Song-Hua Xu[1,*] (徐颂华), and Cheickna Sylla[2]

[1] *School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710000, China*

[2] *Martin Tuchman School of Management, New Jersey Institute of Technology, Newark 07101, U.S.A.*

E-mail: tyc19960606@stu.xjtu.edu.cn; songhuaxu@mail.xjtu.edu.cn; cheickna.sylla@njit.edu

Received January 29, 2021; accepted February 1, 2023.

Abstract Skeletons of characters provide vital information to support a variety of tasks, e.g., optical character recognition, image restoration, stroke segmentation and extraction, and style learning and transfer. However, automatically skeletonizing Chinese characters poses a steep computational challenge due to the large volume of Chinese characters and their versatile styles, for which traditional image analysis approaches are error-prone and fragile. Current deep learning based approach requires a heavy amount of manual labeling efforts, which imposes serious limitations on the precision, robustness, scalability and generalizability of an algorithm to solve a specific problem. To tackle the above challenge, this paper introduces a novel three-staged deep generative model developed as an image-to-image translation approach, which significantly reduces the model's demand for labeled training samples. The new model is built upon an improved G-net, an enhanced X-net, and a newly proposed F-net. As compellingly demonstrated by comprehensive experimental results, the new model is able to iteratively extract skeletons of Chinese characters in versatile styles with a high quality, which noticeably outperforms two state-of-the-art peer deep learning methods and a classical thinning algorithm in terms of F-measure, Hausdorff distance, and average Hausdorff distance.

Keywords skeletonization of characters, three-staged skeletonization, X-net

1 Introduction

The skeleton of a character richly reveals essential structural and shape features of the character, which benefit a wide array of character image processing tasks such as optical character recognition, image restoration, stroke segmentation and extraction, and style learning and transfer, as well as identity authentication and verification based on handwriting analysis. Skeletons of Chinese characters are no exception in furnishing these merits. For example, character skeletons have been utilized to significantly improve the performance of handwritten and calligraphic characters recognition[1–4]. Empowered by the valuable informational aid provided by character skeletons, researchers have successfully improved the quality of style transfers among characters[5–7], and enhanced the capability of conventional deep generative models[8–10] in skeletonizing Chinese characters with versatile styles and complex structures as well as conducting end-to-end deep style learning and generation[11].

Unfortunately, the aforesaid benefits of skeletons and skeleton analysis in processing Chinese character images are largely underexplored for the following reasons. First, skeletonizing Chinese characters on a manual basis incurs an inhibitive cost due to the large character set in use. For example, the standard character set used in Chinese mainland, GB 2312—80, includes 6 763 most frequently used characters while the

Regular Paper

This work was supported by the “2030 Major Project—New Generation Artificial Intelligence” funded by the Chinese Ministry of Science and Technology under Grant No. 2020AAA0106302, and the National Natural Science Foundation of China under Grant No. 6187615.

*Corresponding Author

total number of Chinese characters is above 50 000①. Moreover, skeleton analysis is severely hampered due to the wide assortment of writing styles deployed in everyday use. The form of a Chinese character can be heavily influenced when written in different styles. In Subsection 4.5, we show the appearance of a Chinese character, "鼎", an ancient vessel with two handles and three or four legs that symbolizes the throne of a kingdom, written in 28 well-recognized styles. This example vividly demonstrates the wide variation in the possible look of a character with versatile styles. Countless handwritten styles have been invented throughout centuries in Chinese history[12], which further raises the difficulty in conquering the Chinese character skeletonization task.

To computationally skeletonize a Chinese character image, thinning algorithms are often leveraged, especially during the pre-deep learning era. For example, the neighbor-based thinning algorithm introduced in [13] presents a classical solution, which iteratively deletes pixels on the foreground boundary of a character until reaching the middle axis of the character. Follow-up research efforts modified the particular thinning rules used in the original work, resulting in a large number of derived algorithms, e.g., [14, 15]. The benefit of these thinning algorithms is their data-free nature. That is, no training data is required to construct an algorithm thanks to the carefully designed and manually encoded thinning heuristics deployed in building such an algorithm. Unfortunately, no known heuristics are able to reliably cope with the diverse and complex stroke combinations and compositions exhibited in Chinese characters, and thus heuristics-based thinning algorithms tend to be error-prone and fragile. Fig.1 shows skeletons extracted from a few sample characters using the algorithm of [13] where the enlarged inserts show erroneous skeletonization results, which frequently occur in the areas of overlapping strokes.

Most recently, the research published by Wang and Liu[1] introduced a deep learning based solution to address the above limitation in skeletonizing Chinese characters. Their method relies upon a handwritten Chinese character recognition task for pre-training the new model. The derived features are subsequently repurposed for the skeletonization task. To successfully conduct the first recognition task, their method anticipates recognition labels for a large number of Chinese characters. Such a heavy demand for the training data severely limits the transferability and generalizability of their method in processing characters with versatile styles. To obtain skeletons with a single-pixel width, Wang and Liu[1] further applied binarization and thinning to post-process the probability map outputted by their model, leading to unnatural distortions shown in Subsection 4.5.

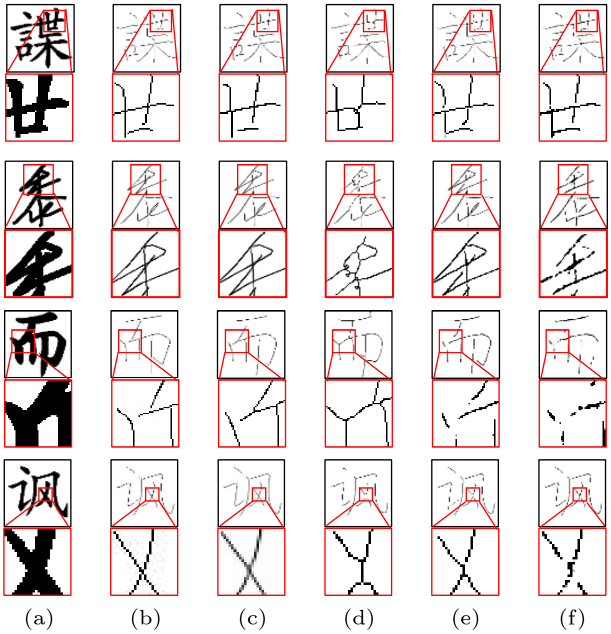


Fig.1. Ground truth (GT) skeletonization results in comparison with results by different algorithms, including results by the proposed model (Proposed), by a classical thinning algorithm (Thinning[13]), and by two state-of-the-art deep learning models (SegNet[16] and HED[17]). (a) Input. (b) GT. (c) Proposed. (d) Thinning. (e) SegNet. (f) HED.

On a separate line of research, holistically-nested edge detector (HED)[17], which is originally designed for edge detection, is found to have positive effect on detecting skeletons of objects and thus becomes a popular model for conquering the skeletonization task. Recent methods of object skeleton detection[18–21] are based on HED. However, these revised approaches still fail to reliably extract character skeletons. Another class of skeletonization models is developed based upon the classic SegNet algorithm[16], which is originally introduced as an image-to-image translation approach for semantic segmentation. Compared with peer methods[22, 23], SegNet has fewer layers and therefore attains a faster inference efficiency. SegNet is suitable for skeletonization because skeleton points are primarily determined by their surrounding pixels

①https://www.bbc.co.uk/languages/chinese/real_chinese/mini_guides/characters/characters_howmany.shtml, Oct. 2021.

within a local neighbourhood in a Chinese character image, whose extraction thus does not require imagery information supplied from distant elements in the image. However, skeletons extracted by the both HED and SegNet methods tend to carry many breakpoints and uneven skeleton trajectories as shown in Section 4.

To overcome shortcomings of the above state-of-the-art solutions, this paper introduces a novel deep generative model to extract high-quality skeletons of Chinese characters following an image-to-image translation framework, which significantly reduces the model's demand for training samples when applied to process characters in versatile styles. The new model employs a three-staged generative pipeline, which leverages a modified G-net, an enhanced X-net, and a newly proposed F-net. The newly introduced F-net comprises a deep multiresolution convolutional structure, coupled with a pair of attention modules, responsible for extracting channel and spatial attention from underlying feature maps. The new model is further equipped with a newly composed contextual loss term to enhance the visual quality of the end character skeletonization results.

The main contribution of the proposed Chinese character skeletonization model lies in the novel design of its underlying deep neural network. Enabled by the newly introduced three-staged iterative image-to-image generation model, the proposed solution is capable of attaining a noticeably superior quality in its skeletonization results in comparison with results generated by multiple state-of-the-art peer methods. To the best of our knowledge, the proposed model is the first one capable of effectively skeletonizing Chinese characters in versatile styles with a satisfactory quality.

Inspired by the prior work[24], we further introduce a contextual loss term that calculates the similarity of a pair of images by comparing their respectively derived feature maps while loosening certain types of spatial constraints. An auto-encoder trained with skeleton data is employed to obtain features for computing contextual loss, which evidently outperforms the VGG19[25] used in original work[24]. Empowered by the contextual loss and the specially designed feature derivation method[25], the proposed model achieves better performance under the measure of frechet inception distance (FID)[26], the metric of which is widely considered to resemble human perception and hence valuable for evaluating the capability of an image generation model.

It is noteworthy that the proposed model requires a much smaller size of training samples than its peer methods to attain a comparable visual quality. For example, the leading deep learning based peer method[1] needs 1.121 million pairs of Chinese character images and their associated skeletons for the training when applied to tackle the skeletonization task defined in [27]. In contrast, the proposed model only needs 0.14 million pairs of training samples, a reduction of training data by 87.5%, to obtain the results which perform better on subjective opinions scores. For two other smaller datasets, SkeletonMF[28] with 13 500 pairs of training samples and Kaiti9574② with 7 000 pairs of training samples, the peer method[1] cannot be adequately trained to obtain decent skeletonization results. In contrast, the proposed model is still able to produce visually satisfactory results as shown later in Section 4. It is noted that for the Kaiti9574 dataset, the smallest one among the three experimental sets, the new model is able to learn using only 40 training samples to produce visually acceptable results (see Subsection 4.4). These results consistently demonstrate the capability of the new model in learning using a much smaller sample size than all its peer deep learning solutions.

The remaining sections of this paper are organized as follows. Section 2 briefly overviews existing work most closely related to this study. Section 3 presents the key design of the proposed method in detail. Section 4 elaborates on the experimental results by the new algorithm, which are compared with the results of multiple state-of-the-art peer methods. Section 5 concludes this study and points out directions for future research.

2 Related Work

2.1 Image-to-Image Translation

Image-to-image translation aims to learn specific mapping between two image domains while preserving their shared characteristics. Once the mapping is acquired, the style of one domain can be transferred to that of the other domain without distorting the underlying image content. For example, the PIX2PIX work[29] applies conditional adversarial networks as a

② https://github.com/skishore/makemeahanzi, Oct. 2021.

general-purpose solution to tackle image-to-image translation tasks. It can effectively solve tasks such as synthesizing photos from label maps and reconstructing objects from edge maps. In addition, image-to-image translation algorithms have also been successfully applied to tackle a variety of image generation tasks, such as neural style transfer[30], cross-view image translation[31, 32], and font generation[5–7]. In this study, we consider the image of a character and its corresponding skeleton as an object manifested in two image domains, under which perspective the character skeletonization task is converted into an image translation task.

When conquering generative tasks with complex or visually challenging goals, a single image-to-image translation network may not be able to deliver satisfactory results. In this circumstance a cascade of networks are sometimes leveraged to progressively synthesize a desirable result, e.g., [11, 31, 33–36]. In the character skeletonization task, our model considers the width of a skeleton as a single pixel width. In this way our model recognizes the difficulty for a single generative network to accurately produce a skeleton in one shot. Hence, our model recognizes the need to utilize a cascade of generative networks to provide a more adept solution. This idea inspires the three-staged generative deep network approach proposed in this paper. We propose a set of comprehensive experimental results to demonstrate the effectiveness of such a design approach.

For the font generation, many studies in recent years have tried to integrate the domain knowledge into the model[8–10, 37–41] to further improve performance. This is actually a return to the traditional Chinese character generation method in the pre-deep learning era[42–45]. As a key form of expression of Chinese characters, skeletons play an important role in this trend, which highlights the importance of automatic skeleton extraction.

2.2 Skeleton Detection

Skeleton extraction and detection are intensively investigated in computer vision and image processing, under tasks such as action detection[46] and natural object skeletonization[21, 47], to name just a couple. Traditional algorithms are architectured around some thinning process, which derive a target skeleton either by iteratively deleting points on the boundary of an object or directly through a single hop. Modern skeleton detection algorithms are mostly empowered by deep learning methods.

End-to-end skeleton detection through CNNs is a popular class of approaches, of which many are based on an end-to-end edge detection algorithm, including HED[17] and many of its variants and improvements[18–21, 47, 48]. Among the follow-up studies of HED, [18] leverages a bidirectional residual learning scheme; [21] adopts a hierarchical fusion procedure; [47] employs a geometry-based loss function. Very limited attention has been paid however to extracting skeletons of Chinese characters. To the best of our knowledge, [1] is the only deep learning based approach solely developed for skeletonizing Chinese characters. As discussed earlier, the deep learning based model requires a much larger size of training samples than the proposed approach. It is therefore difficult to apply the peer method to skeletonize characters in versatile styles, a drawback we seek to address in this study.

3 Proposed Model

Fig.2 presents the main architecture of the three-staged proposed model for skeletonizing Chinese characters in versatile styles. G, X and F represent a pre-generation network, a refined X-net, and a multiresolution feature fusion net (see details in Figs.3–5), respectively. These three networks, which are referred to as the G-Net, X-net, and F-net from now on, respectively, sequentially power each of the three key generation stages of the new model. The initial motivation of using the multi-stage model is an observation that the single-pixel width characteristics of the skeleton make it quite difficult to directly generate a satisfactory skeleton. In particular, this task is not

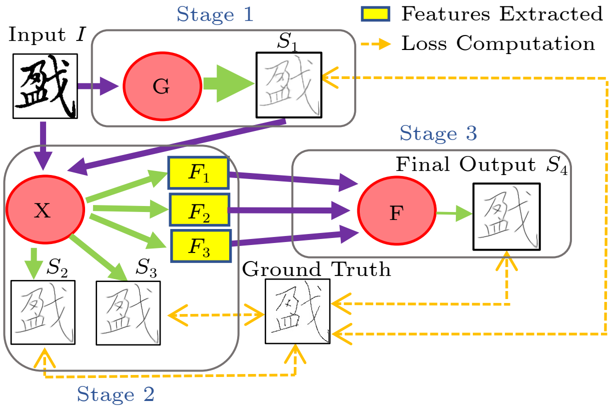


Fig.2. Architecture of the proposed model. Details of three sub-networks G, X, and F are displayed in Figs.3–5, respectively.

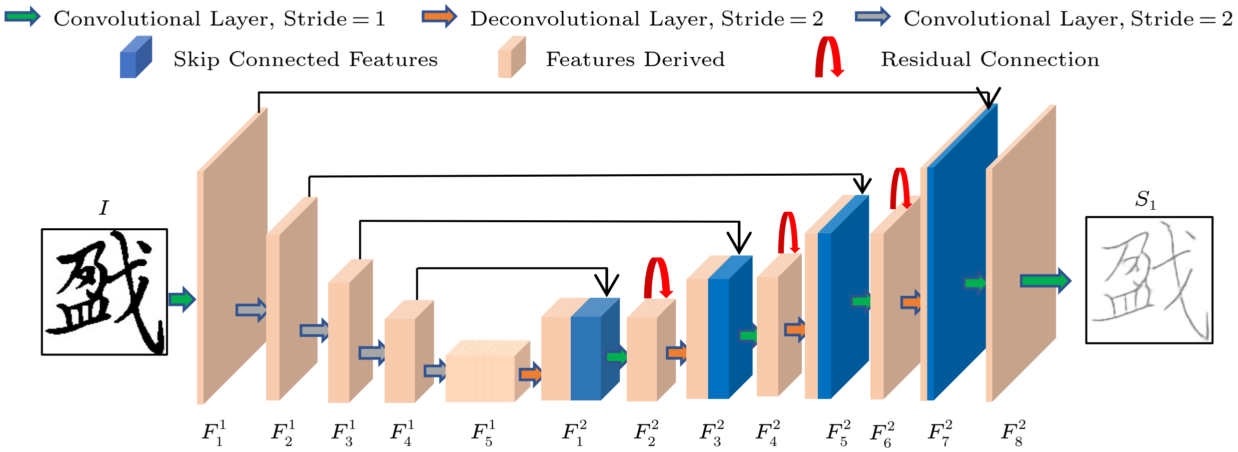


Fig.3. Architectural detail of the pre-generation network, G-net. The dimensions of feature maps are specified in Table 1.

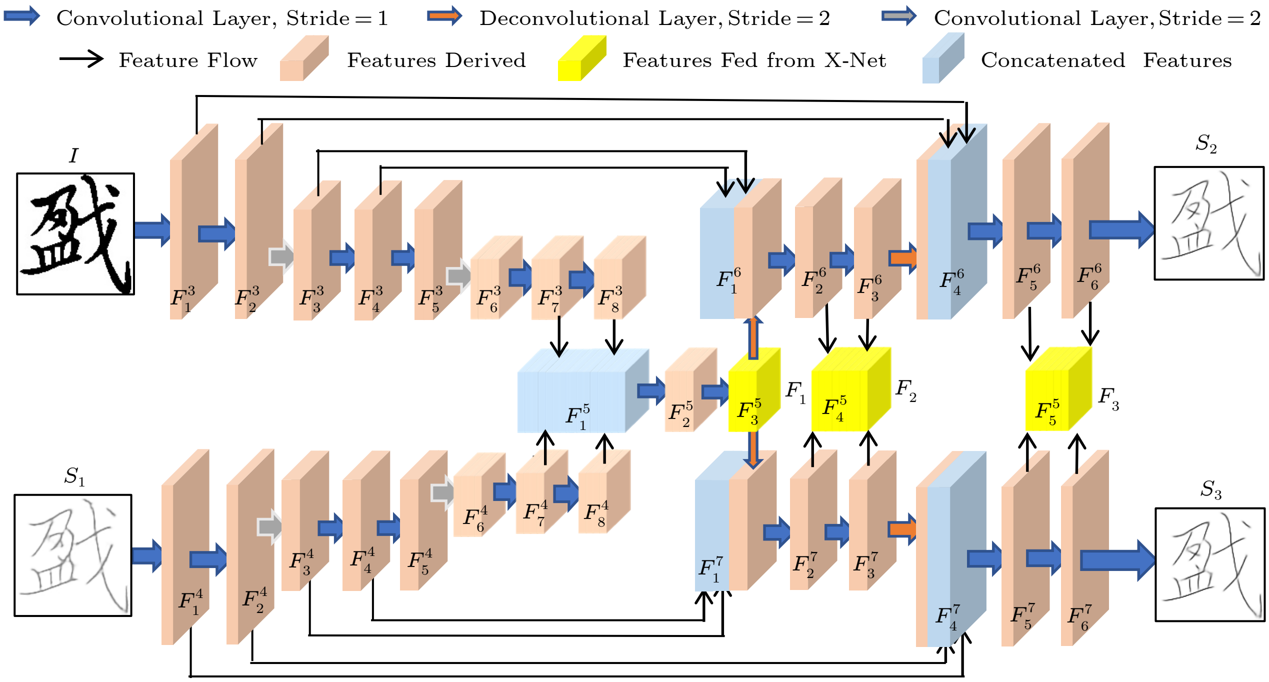


Fig.4. Architectural detail of the refined X-net. The dimensions for feature maps are specified in Table 2.

particularly complicated, so that increasing the number of parameters of the network is not effective. Under the above considerations, it is a natural idea to make detailed corrections to a preliminary result, which leads to our three-staged model. However, the selection of the networks in the subsequent stages still requires a careful design.

The pre-generation G-net produces a preliminary version of the skeleton S_1 for an input character image I, which is subsequently fed to the X-net along with the original character image I. One branch of the X-net takes S_1 as its guidance information to extract a refined skeleton S_2 for I while the other branch of the X-net treats I as a reference to refine the preliminary skeletonization result S_1, thus producing the output S_3. Such a crossover procedure implemented by the X-net allows us to fully exploit potentially useful information from multiple sources. Finally, the F-net employs a convolutional structure nested with a multiresolution synthesis paradigm to synthesize the ultimate output of the network S_4 by utilizing the three generative feature maps F_1, F_2, and F_3 (see details in Fig.2 and Figs.4–5) derived by the X-net, respectively.

3.1 Stage 1: Pre-Generation Network, G-Net

As mentioned earlier, we regard the Chinese character skeletonization problem as an image-to-image translation task because of the considerable resem-

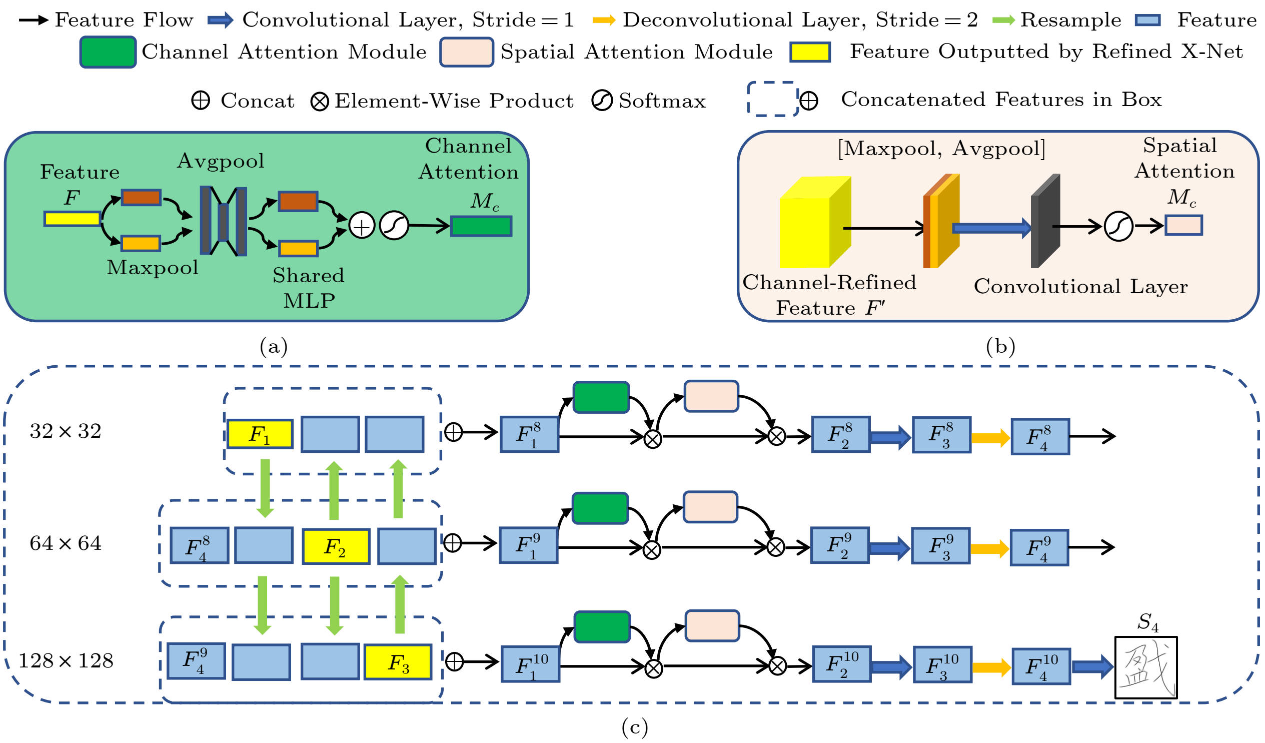


Fig.5. Architectural detail of the attention-based multiresolution feature fusion net, F-net. The dimensions for feature maps are specified in Table 3. (a) Channel attention module. (b) Spatial attention module. (c) F-net. MLP: multi-layer perceptron.

Table 1. Dimensions of Feature Maps in G-Net

Feature Name	Resolution	Number of Channels
F_1^1	128	8
F_2^1	64	16
F_3^1	32	32
F_4^1	16	64
F_5^1	8	128
F_1^2	16	128
F_2^2	16	64
F_3^2	32	64
F_4^2	32	32
F_5^2	64	32
F_6^2	64	16
F_7^2	128	16
F_8^2	128	8

Table 2. Dimensions of Feature Maps in X-Net

Feature Name	Resolution	Number of Channels
F_1^3, F_2^3	128	8
F_3^3, F_4^3, F_5^3	64	16
F_6^3, F_7^3, F_8^3	32	32
F_1^4, F_2^4	128	8
F_3^4, F_4^4, F_5^4	64	16
F_6^4, F_7^4, F_8^4	32	32
F_1^5	32	128
F_2^5, F_3^5	32	32
F_4^5	64	64
F_5^5	128	32
F_1^6	64	48
F_2^6, F_3^6	64	16
F_4^6	128	24
F_5^6, F_6^6	128	8
F_1^7	64	48
F_2^7, F_3^7	64	16
F_4^7	128	24
F_5^7, F_6^7	128	8

blance between the contour of a character and its skeleton. The pre-generation network, G-net, which is responsible for the first stage of the skeleton generation task in the proposed model, is architectured based upon the backbone of the U-net design. Such a choice is deliberately made because U-net is able to sensitively respond to local characteristics in an input image as supported by the skip connections embedded in the U-net. We argue this property of U-net is particularly desirable for our character skeletonization task because of the heavy influence of local shape characteristics of a stroke on its underlying skeleton, especially the skeletons of its intersecting strokes. In fact, the choice of this backbone network is not common in the field of skeletonization, for example the

Table 3. Dimensions of Feature Maps in F-Net

Feature Name	Resolution	Number of Channels
F_1^8, F_2^8	32	112
F_3^8	32	64
F_4^8	64	32
F_1^9, F_2^9	64	144
F_3^9	64	64
F_4^9	128	32
F_1^{10}, F_2^{10}	128	144
F_3^{10}	128	64
F_4^{10}	128	8

object skeletonization[14–18].

In our design for the G-net, we modify the original U-net architecture by adding a series of residual connections to the network to better preserve the detail and sharpness of its skeletonization result. The effectiveness of the adding operation is empirically validated through our experiments. The end design of the G-net first conducts an encoding process by downsampling the input image of a character with a resolution of 128 × 128 pixels to the resolution of 8 × 8 pixels through five consecutive convolutional layers with a step size of 2. A subsequent decoding process is executed by the G-net to reproduce an image at the original resolution of input with the aid of skip connections. In the decoding phase, a residual connection is introduced accompanying each skip connection in the original U-net design considering the sensitivity of the skeleton to localized details. Since an ideal skeleton shall assume only a single pixel width, we use a step size of 2 in the de-convolution layers to mitigate the degradation of visual quality during the upsampling process.

Table 4. Performance of Different Versions of Proposed Model on Dataset Kaiti9574

Model	FID	OFM	OAHD	OHD
S.1	80.7	0.726	0.552	6.06
S.2U	77.1	0.728	0.542	5.26
S.2M	72.4	0.754	0.481	4.60
S.2	67.8	0.760	0.472	4.50
S.3WOA	66.4	0.772	0.448	4.27
S.3WOMR	64.7	0.761	0.472	4.95
S.3M	68.7	0.768	0.459	4.72
S.3	**61.8**	0.774	0.446	4.21
S3.D	63.7	**0.777**	**0.438**	**4.02**

Note: The numbers that indicate the best performance are presented in bold. A larger OFM value, a smaller FID value, a smaller OAHD value and a smaller OHD value all indicate a better skeletonization result.

Table 5. Performance of Different Versions of Proposed Model on Dataset HW

Model	FID	OFM	OAHD	OHD
S.1	39.8	0.891	0.266	3.82
S.2U	34.2	0.896	0.246	3.80
S.2M	27.9	0.889	0.262	3.22
S.2	28.0	0.911	0.208	3.14
S.3WOA	19.2	0.920	0.190	3.06
S.3WOMR	20.6	0.919	0.192	3.08
S.3M	20.1	0.915	0.187	3.00
S.3	18.8	0.923	0.179	**2.98**
S3.D	**18.6**	**0.925**	**0.174**	3.01

Table 6. Performance of Different Versions of Proposed Model on Dataset SkeletonMF

Model	FID	OFM	OAHD	OHD
S.1	165.2	0.498	1.263	10.54
S.2U	153.5	0.509	1.212	10.12
S.2M	155.6	0.515	1.211	9.95
S.2	154.8	0.520	1.162	9.84
S.3WOA	152.9	0.520	1.149	9.73
S.3WOMR	148.8	0.522	1.162	9.76
S.3M	149.1	0.521	1.153	9.75
S.3	148.5	0.525	1.143	9.69
S3.D	**144.9**	**0.529**	**1.125**	**9.58**

3.2 Stage 2: X-Net

The second stage of the network, as implemented through the X-net, aims to refine the preliminary skeletonization result S_1 generated in the first stage by the G-net. To fully exploit visually useful information latent in both S_1 and the original input image of a character I, we introduce the X-net. It is noted that simply concatenating S_1 and I for feeding the result into a deep network with a single input branch will not achieve the same quality of skeletonizing as X-net with two input branches, which is both empirically verified through our experiments (see S.2U in Tables 4–6) and analyzed as follows: when S_1 is perceptually close to the end skeletonization result, the network tends to ignore visual information provided by I during its training, and vice versa. In either situation, the network is inclined to ignore one of the input sources. To adequately explore potentially useful information from both sources, we introduce the X-net. The main purpose of introducing the X-net is to thoroughly mine useful information for the skeleton in each input image through a specific network structure. An observation that replacing a certain input in S.2U with a blank image sometimes does not heavily affect the final result contributes to this special design.

X-net has two input and two output branches. Any combination of an input and an output branch forms an encoder-decoder pathway with skip connections. The network takes the original character image I as one of its input, transformed by features extracted from the preliminary skeleton extraction result S_1. The latter is fed in by the G-net as the second input to the X-net. After performing the above encoding process, X-net decodes the encoded feature map into a refined skeleton image S_2. In a similar way, the input image S_1 is transformed by features extracted from I to generate another encoded feature map, followed by a similar decoding process used in the above to produce another refined version of the skeleton image S_3. By utilizing such interwound pairs of encoding-decoding transformations, visual clues latent in both I and S_1 are thoroughly mined to produce two refined skeletonization results as the output of the model in its second stage. The skip connections between the paired input branch and output branch force the network to use the corresponding input image as the main reference for generating the skeleton (for example, I for generating S_2), and to use the features of the other input (S_1 for generating the S_2) as supplementary information because it only participates in the generated bottom layer (F_3^5 in Fig.4). This design allows that both two inputs contribute to each generated skeleton and the valid information carried by each input is sufficiently explored.

The detailed construction of the X-net is composed by three parts, encoding, fusion, and decoding. In the first encoding part, an input image of 128×128 pixels is down-sampled progressively through multiple convolutional layers. After downsampling with a reduction factor of 4, we obtain a feature map of the resolution of 32×32. It is noted that even though the two encoder branches of the X-net, shown as the two branches positioned in the left side of Fig.4 have the same structure, they do not share any parameters due to the differences in shapes and visual features of I and S_1 respectively. The second fusion part is the only step in which two branches of the X-net exchange information. As shown in Fig.2, features with a resolution of 32×32 generated by the two encoding branches are concatenated and then fed into the convolutional layers for a more thorough fusion. Finally, in the decoding part, low-resolution encoded features are re-sampled to the resolution of 128×128. In the decoding process, features from the encoding part are also incorporated through skip connections provided by the X-net. Like the two encoding branches, the two decoding branches also do not share their parameters to preserve the possible difference in the two refined skeletonization results. It is noted that the F-net, the downstream network to the X-net, does not directly utilize the skeletons S_2 and S_3 outputted by the X-net. Instead, it only uses the three sets of features extracted by X-net, $F_1(32 \times 32)$, $F_2(64 \times 64)$ and $F_3(128 \times 128)$, all produced through concatenating intermediary features generated by the two decoding branches.

To explore the effectiveness of the two output branches utilized in the X-net, we experiment with an alternative design where the two output branches of X-net are merged into one branch so that the X-net is reduced into a Y-net. All the relevant skip connections originally forked into the two output branches are also merged into one skip connection. Another similar attempt is to directly concentrate I and S_1 and feed the results to another G-net, which can be considered further merging the two input branches of the Y-net. Benchmarked experimental results demonstrate the advantage of using the X-net in our model design. We attribute this performance advantage to the independent skeleton refinement processes carried out by the two output branches in the X-net.

3.3 Stage 3: F-Net

Finally, the F-net takes the three feature maps, F_1, F_2 and F_3, all derived by the X-net, as its input to produce the final skeletonization result S_4. The main aim of the F-net is to refine the intermediate skeletonization results generated by the X-net to attain richer synthesis details and better stability. The effectiveness of the F-net is based on the following two points. 1) It combines the features of the two output branches from the X-net (mainly expressing the skeleton-related information in I and S_1) for generation. 2) It simultaneously employs features of multiple resolutions, and explores them with a specially designed multi-resolution framework.

The F-net is built upon a deep convolutional structure nested with a multiresolution synthesis paradigm in that the convolutional structure takes as its input a set of multiresolution feature maps resampled from the raw input feature maps, F_1, F_2 and F_3, at various resolutions. We adopt this multiresolution processing paradigm according to the empirical understanding gained through our explorative experiments that suggest the choice of a particular resolution at which a character image is skeletonized could introduce a profound impact on the quality of the end skeletonization result. This trait is frequently exhibited by learning-based image processing algorithms[17]. More specifically, skeletonization results produced at a lower resolution tend to attain a more accurate depiction of the overall structure of a character, however, at the expense of missing fine details; conversely, skeletons inferred at higher resolutions are more likely to capture minute details of a character, yet at the risk of overlooking the global characteristics of a character.

As illustrated in Fig.5, inside the F-net, the feature maps F_1 (32×32), F_2 (64×64) and F_3 (128×128) are first transformed to the other two resolutions via either an interpolation or a resampling procedure such that each feature map ends up with three versions at the resolutions of 32×32, 64×64, and 128×128, respectively. We denote the feature map F_i $(i=1,2,3)$ at the resolution of j^2 $(j=32,64,128)$ as $F_i(j\times j)$.

Next, inspired by the design of the convolutional block attention module proposed in [49], $F_1(32\times 32)$ is concatenated with $F_2(32\times 32)$ and $F_3(32\times 32)$. The result is additionally processed by a channel attention module, a spatial attention module, and a series of convolutional layers sequentially, to derive an overall feature map, $F_{1\oplus 2\oplus 3}(32\times 32)$, at the resolution of 32×32. $F_{1\oplus 2\oplus 3}(32\times 32)$ is then upsampled to the resolution of 64×64, resulting in $F_{1\oplus 2\oplus 3}(64\times 64)$. $F_{1\oplus 2\oplus 3}(64\times 64)$ is subsequently concatenated with $F_1(64\times 64), F_2(64\times 64)$, and $F_3(64\times 64)$, the result of which is similarly processed by the aforementioned channel attention module, spatial attention module, and another set of convolutional layers to derive an integrated feature map, $F_{1\oplus 2\oplus 3}(128\times 128)$, at the resolution of 128×128. Lastly, $F_{1\oplus 2\oplus 3}(128\times 128)$ is concatenated with $F_1(128\times 128), F_2(128\times 128)$, and $F_3(128\times 128)$, followed by similar transformations carried out by the above two attention modules and the convolution layers to yield the final skeletonization result S_4.

3.4 Overall Optimization Objective

Following the treatment adopted by the majority of previous work on the detection and extraction of skeletons from character images[1], the proposed approach also models the skeletonization operation as a pixel-level binary classification task in which each image pixel is individually determined regarding whether it belongs to the skeleton region of a character or not. Under this modeling perspective, we employ the loss function defined as follows:

$$loss_{\text{total1}} = \sum_{i=1}^{4} \alpha_i \times loss_{\text{CE}}(S_i, GTS), \qquad (1)$$

where GTS is the ground truth skeleton corresponding to an input character image I, $loss_{\text{CE}}$ is the cross entropy loss, and α_i $(i=1,2,3,4)$ are the coefficients corresponding to each loss term respectively.

It is also noted that the above loss function (1) does not consider the severity of a particular classification error in its measurement. Intuitively, if a pixel staying close to yet not belonging to the skeleton of a character is mistakenly classified as a skeletal pixel, the severity of such an error should be smaller than that of classifying a pixel distant from the skeleton as a skeletal pixel. Unfortunately, the cross entropy loss term as employed in (1) does not differentiate these two situations, overlooking valuable feedbacks that can be otherwise leveraged to guide the optimization of a machine learning-based solution.

To address the requirement, we introduce a novel distance-based loss function, which has not been used in prior studies in the field. For efficient evaluation of the distance-based loss function, a distance map image needs to be first derived where each pixel position in the map is assigned a distance value that records the position's shortest distance to the skeleton of a concerned character (see Fig.6). Formally, let $\mathbb{P}$ be the set of all skeletal points in a skeleton image S. The pixel value $pixel(q_j)$ of any point q_j in the distance map Q for S is defined as follows:

$$pixel(q_j) = \frac{0.9}{D} \times \min(\min_{p_i \in \mathbb{P}} d(q_j, p_i), D),$$

where D is the threshold parameter, and $d(q_j, p_i)$ is the Euclidean distance between points q_j and p_i. To encourage the proposed network to focus on correctly classifying skeletal points in marginal situations where most errors tend to occur, the threshold parameter D is introduced such that those pixel positions too distant away from the skeleton would not occupy too much attention from the network. Such a tactic essentially helps the network better learn from nega-

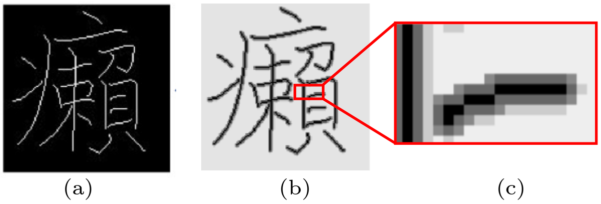


Fig.6. Examples of (a) a skeleton and (b) its corresponding distance map $(M=3)$ with (c) a partial enlarged view.

tive and positive samples, otherwise significantly unbalanced between prospects and background.

Once a distance map GTD is prepared for a given character image I, we can efficiently evaluate the following distance map based loss function:

$$loss_{\text{dis}}(S_i) = loss_{\text{MSE}}(S_i, GTD), \tag{2}$$

where $loss_{\text{MSE}}$ is the mean squared error term. And the total loss function for any candidate skeletonization result is

$$loss_{\text{total2}} = \sum_{i=1}^{3} \alpha_i \times loss_{\text{dis}}(S_i) + \alpha_4 \times loss_{\text{CE}}(S_4, GTS).$$

The above loss function (2) takes into account the geometric severity of the errors, but still does not adequately reflect the impact of the errors on the skeleton topology. Errors that make two separated strokes intersect change the original topology of the skeleton; thus they are more serious in human perception than simply shortening a stroke and may significantly affect the performance of the skeletons on downstream tasks such as handwriting recognition. However, the cross entropy loss and distance map based loss cannot give proper feedback considering the impact of errors on the skeleton structure. Alignment errors, e.g., a small displacement or rotation, belong to such an important type of errors that seriously affects the values of cross entropy loss and distance map based loss but does not change the skeleton topology.

Based on the above considerations, we introduce the contextual loss proposed in [24], which employs a novel feature-based method to compare the similarity between images without the requirement of spatial alignment. Specifically, given the generated image x and ground truth y, the corresponding collection of features $X = \{x_i\}_{i=1}^{N}$ and $Y = \{y_j\}_{j=1}^{N}$ are derived by utilizing a pre-trained model to represent the images x, y respectively, where N is the count of feature maps. The contextual similarity between two images x and y is defined as below:

$$\begin{aligned} CX(x,y) &= CX(X,Y) \\ &= -\log\left(\frac{1}{N}\sum_{j}\max_{i} CX_{ij}\right), \end{aligned} \tag{3}$$

where CX_{ij} is the similarity between x_i and y_j. Let d_{ij} be the cosine distance between x_i and y_j, and the similarity CX_{ij} between x_i and y_j is defined as follows:

$$\begin{aligned} \tilde{d_{ij}} &= \frac{d_{ij}}{\min\limits_{k} d_{ik} + \epsilon}, \\ w_{ij} &= \exp\left(\frac{1-\tilde{d_{ij}}}{h}\right), \\ CX_{ij} &= w_{ij} / \sum_{k} w_{ik}, \end{aligned} \tag{4}$$

where ϵ and h are fixed hyper-parameters. (3) employs best similarity $\max\limits_{i} CX_{ij}$ to measure the similarity between y_j and X instead of using spatially aligned CX_{jj} directly, and consequently enforces the model to pay more attention to structural similarity rather than strict spatial correspondence, which is also emphasized by the design of CX_{ij} in (4). Intuitively, features tend to represent the informative skeleton points instead of the background points, and alleviate the imbalance between the number of positive and negative points in the skeleton images.

The design of contextual loss requires a pre-trained model to derive feature maps from generated skeleton and target skeletons, while VGG19[25] training on image-net employed by the original paper[24] brings no promising improvement. Considering the significant difference between skeleton images and real world images, we propose to utilize auto-encoder pre-trained on skeleton data, which does not require any additional information other than the skeleton images. Such a model is particularly effective in extracting the skeleton features without causing difficulties in data collection or model training. Meanwhile, the auto-encoder is dedicated to reconstructing the input skeletons from the extracted representations, which therefore helps to encode all the valid information in the skeleton images and ensures that the comparison between images is comprehensive enough.

Specifically, the auto-encoder consists of an encoder E_{cx} and a decoder D_{cx}, where E_{cx} includes convolutional layers $\{E_{\text{cx},l}\}_{l=1}^{M}$ with a step size of 2, where M is a hyper-parameter. Given a skeleton image s, the sequence of feature maps $\{\phi^l(s)\}_{l=1}^{M}$ is computed step by step as $\phi^l(s) = E_{\text{cx},l}(\phi^{l-1}(s))$ where $\phi^0(s) = s$. The followed decoder D_{cx} reconstructs the final features $\phi^M(s)$ into the skeleton image s_{fake}. We utilize the cross entropy loss to train E_{cx} and D_{cx} as below:

$$loss_{\text{auto}}(s_{\text{fake}}, s) = loss_{\text{CE}}(s_{\text{fake}}, s),$$

where $loss_{\text{CE}}$ is the cross entropy loss. The final contextual loss between generated image x and target image y using the above-mentioned auto-encoder is computed as:

$$loss_{\mathrm{CX}}(x,y)=\sum_{l}CX(\phi^{l}(x),\phi^{l}(y)),$$

where the value range of l is an optional hyper-parameter.

We apply contextual loss as a supplement of the aforementioned losses in the third stage and the corresponding overall loss under the ground truth skeleton GTS is as follows:

$$loss_{\mathrm{total3}}=\sum_{i=1}^{3}\alpha_i\times loss_{\mathrm{dis}}(S_i)+\alpha_4\times loss_{\mathrm{CE}}(S_4,GTS)+\\ \alpha_5\times loss_{\mathrm{CX}}(S_4,GTS).$$

4 Experiments

4.1 Implementation Detail

4.1.1 Datasets

Three datasets, Kaiti9574, HW, and SkeletonMF are used in our experiments. The Kaiti9574 dataset, collected from the Make-Me-a-Hanzi project[③], contains images of 9 574 Kaiti characters and their corresponding skeletons. We randomly select 7 000 characters for the training. The SkeletonMF dataset provided in [28] contains 27 fonts, each of which has 639 characters. We randomly select 500 characters from each font of the dataset to make up a training set of a total size of 13 500 sample characters. The HW dataset is obtained from the data released by [27], which carries a total of 220 000 online trajectories for handwritten Chinese characters. We randomly select 140 000 characters as training samples in our experiment.

4.1.2 Experimental Setup

The proposed model is implemented in PyTorch. All experiments are conducted on NVIDIA GeForce RTX 2080 Ti GPU. The learning rate used for training the proposed network has an initial value of 0.000 2. Unless otherwise specified, each model training takes 10 epochs.

To verify the effectiveness of various components in the proposed model, ablation analysis is conducted. A series of variations of the proposed model are hence introduced. S.1 refers to the version of the proposed model that only executes its first stage and uses the output S_1 from the pre-generation network G-net as its result. Similarly, S.2 and S.3 refer to the version of the proposed model that uses the output of stage 2 and stage 3 as the model output, respectively. For the two outputs generated by stage 2 of the model, the one that yields a higher performance metric is used in the following analysis. Note that S.3 is the full model proposed. S.2M refers to a version of the proposed model that uses the output of a Y-net instead of the X-net as its output (see details discussed in Section 3) while S.3D refers to a version that turns the ground truth of the first two stages into the distance map as described in Section 3. Finally, S.3C refers to a version that adds contextual loss (see Subsection 3.4) to S.3D with an auto-encoder pre-trained on skeleton data, while S.3CV that adopts contextual loss with VGG19 pre-trained on image-net is also trained for comparison. Four new variants of the proposed model, S.2U, S.3M, S.3WOA, and S.3WOMR, are also trained to more carefully verify the effectiveness of the model's components. S.2U refers to a version that employs modified U-net to replace the X-net of S.2 in stage 2 (see Section 3), S.3M is an extended version of S.2M which takes the features obtained by the Y-net in stage 2 as the input of the F-net, and S.3WOA, S.3WOMR refer to the variants of S.3 without CAM, SAM (see Subsection 3.3) and without the multiresolution synthesis paradigm, respectively.

4.1.3 Evaluation Metrics

Four experimental metrics, including the frechet inception distance (FID)[26], F-measure, Hausdorff distance (HD), and average Hausdorff distance (AHD), are used to evaluate the performance of the proposed model, its variants and peer methods.

The F-measure gives a description of accuracy in the pixel-wise level, while the Hausdorff distance gives a description of geometric similarity. Assuming that p_s and $\widetilde{p}_s$ are the two sets of skeleton pixels in a ground truth skeleton and the corresponding skeletonization result, respectively, for p_s and $\widetilde{p}_s$, HD is computed as:

$$HD(p_s,\widetilde{p}_s)=\max(\max_{b\in p_s}\min_{a\in\widetilde{p}_s}d(b,a),\max_{a\in\widetilde{p}_s}\min_{b\in p_s}d(a,b)),$$

where $d(x,y)$ is the Euclidean distance between pixels a and b. In order to reflect the overall quality of an extracted skeleton, we also introduce the average Hausdorff distance (AHD) as:

[③] https://github.com/skishore/makemeahanzi, Oct. 2021.

$$ahd(p_s, \widetilde{p}_s) = \frac{1}{|p_s|} \sum_{b \in p_s} \min_{a \in \widetilde{p}_s} d(b, a),$$
$$AHD(p_s, \widetilde{p}_s) = ahd(p_s, \widetilde{p}_s) + ahd(\widetilde{p}_s, p_s). \quad (5)$$

The probability prediction map outputted by the network needs to be binarized before being compared with the ground truth. The binarization threshold τ will seriously affect the result. For a fair and more comprehensive comparison, we explore a range of the threshold value. We let $\tau = 0.01, 0.02, \ldots, 0.99$ and report the resulting model performance using both the precision-recall curve and the average Hausdorff distance curve. In the average Hausdorff distance curve, the two terms in (5), $ahd(p_s, \widetilde{p}_s)$ and $ahd(\widetilde{p}_s, p_s)$, are treated as the coordinates of a point on the curve. Lastly, we also calculate an optimal F-measure (OFM), an optimal HD score (OHD), and an optimal AHD score (OAHD) as described in (5), which stand for the highest F-measure, HD score, and AHD score encountered during the exhaustive search of the aforesaid threshold τ, respectively.

We also adopt the frechet inception distance (FID)[26] commonly used in the image generation tasks for assessing the generation quality. FID is employed to measure the distance between images in the feature level, and is consequently closer to human perception than pixel-wise metrics. A smaller FID indicates better generated results in terms of the similarity between features of a generated image and features of the corresponding ground truth images.

4.2 Ablation Study

4.2.1 Effectiveness of Three-Staged Model

An ablation study is conducted to explore the respective contribution of each proposed algorithmic module in the new method to its end skeletonization capability. Specifically, we compare the relative performance among nine alternative versions of the proposed model, S.1, S.2U, S.2M, S.2, S.3WOA, S.3WOMR, S.3M, S.3, and S.3D (see Subsection 4.1 for details), using the three experimental datasets.

Tables 4–6 show the respective performance of these nine versions of the proposed model quantitatively evaluated by using FID, OFM, OHD, and OAHD. It is also noted that the last three numerical metrics are sensitive to the particular binarization threshold applied at the final output stage of the proposed network (see Section 3). To comprehensively explore the relative performance among the five model versions under a variety of binarization thresholds, we further derive and report the precision-recall curves and the average Hausdorff distance curves in Fig.7 for five main model versions applied in experiments.

According to the performance measurements both numerically reported in Tables 4–6 and graphically illustrated in terms of the precision-recall curve and the average Hausdorff distance curve in Fig.7, we can see that S.2 significantly outperforms S.1 in all experiments conducted over the three datasets. As S.2 differs from S.1 only in that the X-net is employed in S.2 but not S.1, the above performance advantage shows the usefulness of the X-net. It is also recognized from Tables 4–6 and Fig.7 that S.2 is consistently superior to S.2M, as well as S.2U. The difference between S.2 and S.2M is that S.2 employs an X-net in its second stage while S.2M adopts a Y-net instead, and the X-net differs from the Y-net only in that the former network has two output branches while the latter network has a single branch. And thus such performance advantage demonstrates that the two output branches of the X-net both contribute meaningfully to the end capability of the proposed skeletonization method. A similar comparison also occurs between S.2M and S.2U that S.2M is consistent and significantly better than S.2U, which verifies the necessity of two separate input branches in the X-net and the Y-net and supports the analysis and discussion in Section 3. S.3 and S.3M are extended versions of S.2 and S.2M respectively, and the advantages of S.3 in numerical evaluation metrics reveal that not only the skeleton, but also the features extracted by the X-net outperform those by the Y-net. It is worth noting that S.3M still has a significant improvement over S.2M. This indicates the advantages of using F-net in stage 3, not only considering two output branches, but also using multi-resolution features as input and performing special processing on it. The performance degradation of S.3WOA and S.3WOMR compared with S.3 in Tables 4–6 indicates that the two important components of the F-net, the multi-resolution structure and attention mechanism, are necessary and effective components, because the only difference between S.3WOA and S.3 is that S.3WOA does not employ the attention module while the only difference between S.3WOMR and S.3 is that S.3WOMR does not employ the multiresolution convolutional structure. Lastly, Tables 4–6 and Fig.7 also show that S.3D generally outperforms S.3,

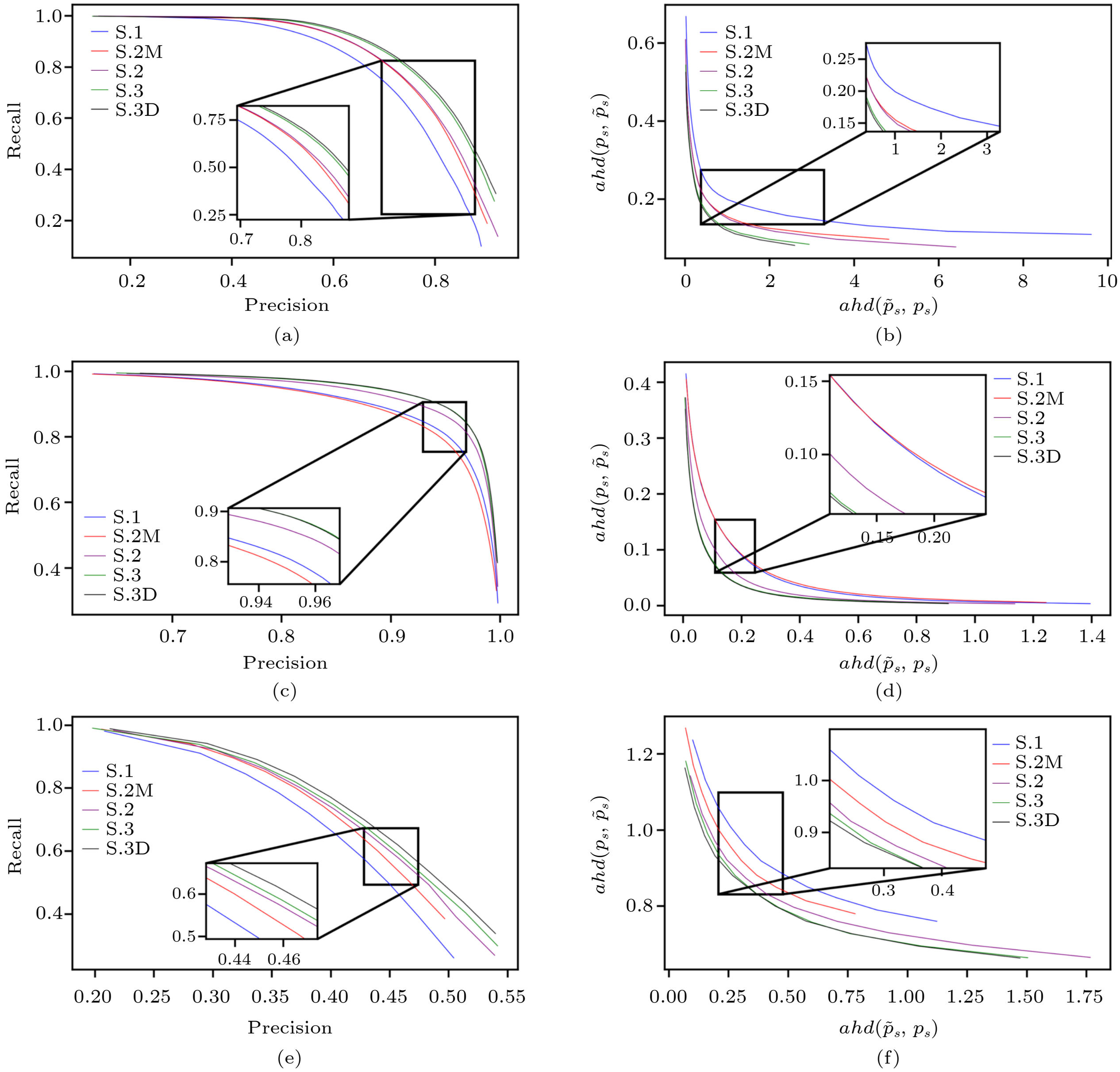


Fig.7. Precision-recall curves on (a) Kaiti9574, (c) HW, and (e) SkeletonMF, and the average Hausdorff distance curves on (b) Kaiti9574, (d) HW, and (f) SkeletonMF.

which indicates that the distance map based loss function defined in (2) indeed helps improve the overall skeletonization capability of the proposed method.

Finally, to give an intuitive demonstration on the gradual refinement effects attained by individual components in the model, Fig.8 shows a set of skeletonization results progressively produced by various stages of the proposed model.

4.2.2 Effectiveness of Contextual Loss

We also conduct experiments to illustrate the influence of contextual loss on skeleton generation. Three versions of the proposed model, S.3C, S.3CV and S.3D (see Subsection 4.1), are trained on three datasets and evaluated by metrics FID, OFM, OHD, and OAHD. S.3C and S.3CV utilize contextual loss with the pre-training model using skeleton data and image-net data as described in Subsection 3.4, and S.3D is trained without contextual loss.

The results are shown in Tables 7–9. The numbers that indicate the best performance is presented in bold. A larger OFM value, a smaller FID value, a smaller OAHD value and a smaller OHD value all indicate a better skeletonization result.

According to the numerical performance, S.3C

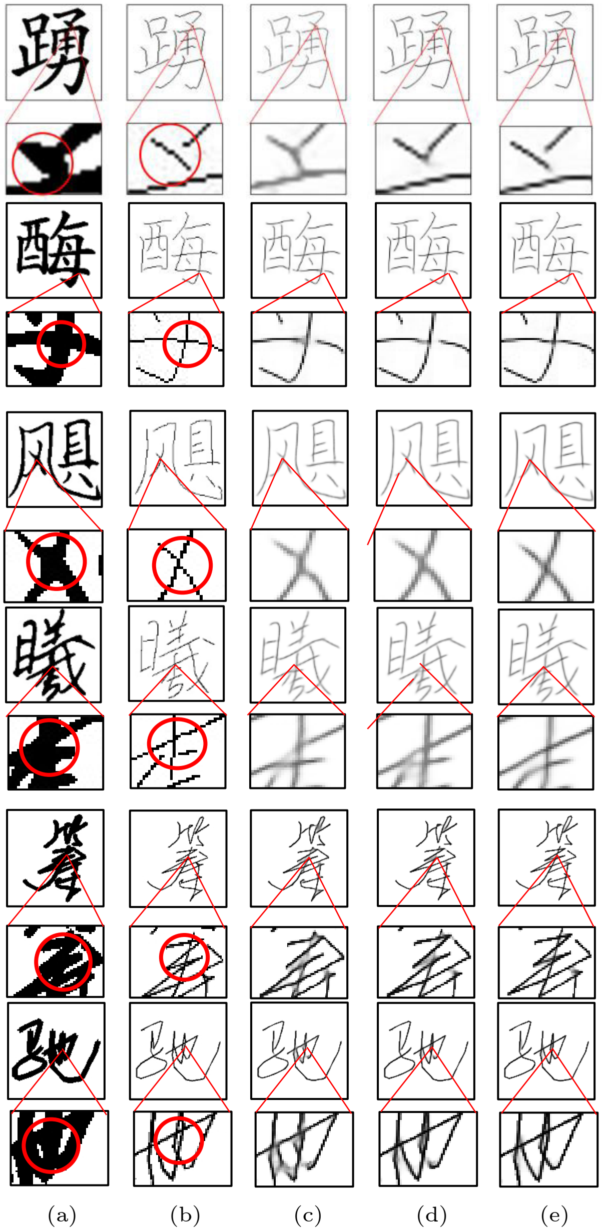


Fig.8. Results by the proposed model in three stages from different datasets. (a) Input. (b) Ground truth. (c) Output of S.1. (d) Output of S.2. (e) Output of S.3.

Table 7. Effectiveness of Contextual Loss on Dataset Kaiti9574

Model	FID	OFM	OAHD	OHD
S.3D	63.7	**0.777**	**0.438**	4.02
S.3C	**56.9**	0.774	0.442	**3.98**
S.3CV	62.7	0.759	0.471	4.05

Table 8. Effectiveness of Contextual Loss on Dataset HW

Model	FID	OFM	OAHD	OHD
S.3D	18.6	**0.925**	**0.174**	3.01
S.3C	**17.3**	0.924	0.176	**3.00**
S.3CV	17.9	0.914	0.185	3.67

Table 9. Effectiveness of Contextual Loss on Dataset SkeletonMF

Model	FID	OFM	OAHD	OHD
S.3D	144.9	**0.529**	**1.125**	9.58
S.3C	**124.1**	0.524	1.148	**9.45**
S.3CV	133.3	0.507	1.388	10.05

performs consistently better than S.3D on FID, which is considered closer to human perception, and performs similarly to S.3D on the other metrics. Compared with S.3D, S.3C only adds contextual loss, and thus the above advantages indicate that the skeletonization performance of the proposed model can be effectively improved by introducing the contextual loss. It is worth noting that S.3CV only employs VGG19 as the original work[38] trained with the image-net data to replace the auto-encoder trained with the skeleton data in S.3C, but its performance is significantly reduced. This verifies the necessity of using skeleton data instead of real-world images for pre-training.

Furthermore, we explore the impact of reducing the number of training samples on S.3D and S.3C by conducting experiments over the Kaiti9574 dataset. We employ 7 000, 1 000, 200, and 40 training samples of Kaiti9574 to train S.3D and S.3C, and the respective performance is evaluated in Table 10. The results indicate that the reduction in the data size impairs the performance of the model on all metrics; however, S.3C is less affected. For example, S.3C performs slightly weaker than S.3D on OFM with 7 000 training samples (0.774 vs 0.777), but performs similarly with 1 000 (0.728 vs 0.727), and clearly outperforms S.3D with 200 or 40 training samples (0.705, 0.674 vs 0.691, 0.643). This expanded experiment shows that the contextual loss has stronger advantages in the scenario with a small size of the dataset.

Table 10. Impact of Training Set Size on S.3D and S.3C over Dataset Kaiti9574

Model	FID	OFM	OAHD	OHD
S.3D (7 000)	63.7	0.777	0.43	4.02
S.3C (7 000)	56.9	0.774	0.44	3.98
S.3D (1 000)	81.4	0.727	0.59	5.93
S.3C (1 000)	66.4	0.728	0.54	4.87
S.3D (200)	90.0	0.691	0.64	7.12
S.3C (200)	74.9	0.705	0.61	7.45
S.3D (40)	107.9	0.643	0.77	8.39
S.3C (40)	82.6	0.674	0.71	8.63

Note: 7 000, 1 000, 200, and 40 are the number of training samples.

4.3 Comparison with State-of-the-Art Peer Methods

Next, we compare the performance of the proposed model and multiple state-of-the-art peer methods, including the classical thinning algorithm ZhangSuen introduced in [13] and two deep learning based skeletonization models— HED[17] and SegNet[16]. In this set of comparative experiments, we use the full version of the proposed model, S.3D, since it attains the best skeletonization results according to the finding obtained in the above ablation study. Tables 11–13 show the respective performance of all concerned methods under comparison as quantitatively evaluated using the three performance metrics, OFM, OHD, and OAHD in experiments conducted over the three datasets. S3.D indicates the proposed model. The numbers that indicate the best performance is presented in bold. A larger OFM value, a smaller FID value, a smaller OAHD value and a smaller OHD value all indicate a better skeletonization result. These results consistently reveal the superiority of the proposed model compared with all peer solutions.

Table 11. Comparison with Peer Methods on Dataset Kaiti9574

Model	FID	OFM	OAHD	OHD
ZhangSuen[13]	74.8	0.427	1.37	10.45
HED[17]	145.1	0.740	0.57	6.45
SegNet[16]	94.4	0.726	0.57	5.74
S3.D	**63.7**	**0.777**	**0.44**	**4.02**

Table 12. Comparison with Peer Methods on Dataset HW

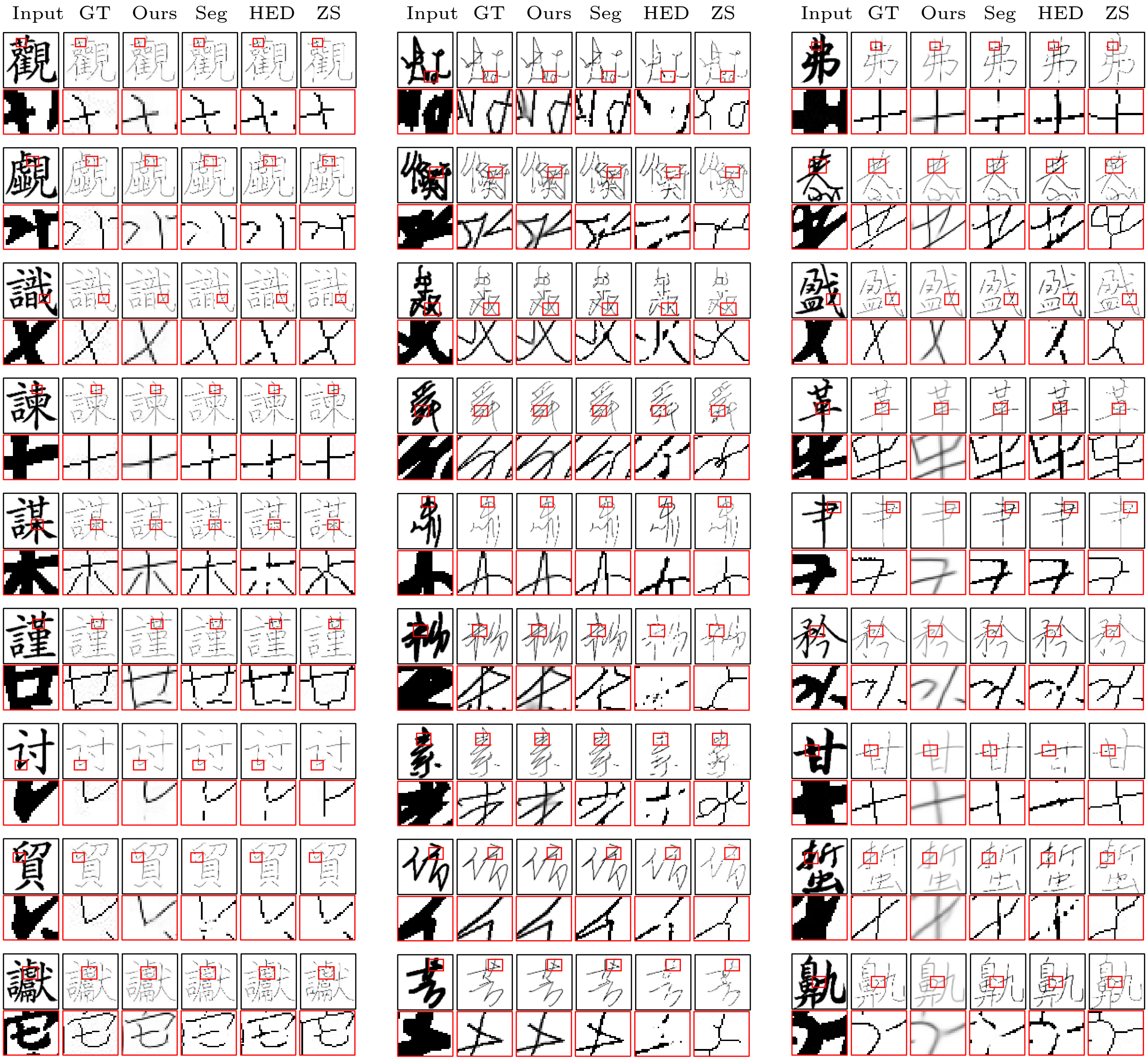
Model	FID	OFM	OAHD	OHD
ZhangSuen[13]	114.5	0.452	1.65	9.67
HED[17]	90.5	0.355	2.28	11.34
SegNet[16]	29.5	0.893	0.28	5.41
S3.D	**18.6**	**0.925**	**0.17**	**3.01**

Table 13. Comparison with Peer Methods on Dataset SkeletonMF

Model	FID	OFM	OAHD	OHD
ZhangSuen[13]	190.2	0.313	2.13	12.82
HED[17]	218.0	0.490	1.77	13.21
SegNet[16]	170.4	0.485	1.64	11.30
S3.D	**144.9**	**0.529**	**1.13**	**9.56**

Among the three existing methods, it is noted that the traditional thinning algorithm is generally inferior to the other two methods in conducting these experiments, except for the experiment carried out over the HW dataset, where the thinning algorithm outperforms the HED method. For the two peer deep learning based skeletonization algorithms, HED and SegNet perform comparatively over the SkeletonMF and Kaiti9574 datasets; yet in experiments executed over the HW dataset, SegNet outperforms the HED algorithm. Considering the fact that characters in HW, all of which are handwritten, display much more cursive and versatile shapes and structures than characters in the other two datasets, where all characters are printed in some standard font, the above experimental results suggest that SegNet is more capable than HED in coping with handwritten characters or characters in versatile styles. In comparison with all the three peer methods, the proposed approach achieves a consistent and significant lead in conducting all experiments carried out over these three datasets, according to the three numerical performance metrics reported in Tables 11–13.

To intuitively demonstrate the relative performance among all peer methods including the proposed approach, Fig.9 lists a few results selected from the above comparison experiments where areas displaying the most erroneous skeletonization results with respect to the corresponding ground truth skeleton are shown in a zoomed-in view. It is easy to notice that skeletons extracted by the ZhangSuen algorithm[13] tend to be continuous and stable, which however are error-prone at stroke intersections. The HED algorithm suffers from the same difficulty in skeletonizing overlapping strokes, and produces fragile skeletonization results for handwritten characters. This observation is consistent with the finding obtained from the quantitative performance analysis discussed in the above. The SegNet algorithm performs most competently among all the three existing solution approaches. It is able to extract skeletons even from relatively complex or cursive characters. However, SegNet fails to retain continuity and details in its skeletonization results, partly because of the frequent breakpoints undesirably generated. In comparison with all these peer methods, the proposed approach achieves a marked advantage in satisfactorily extracting skeletons from characters, both in printed fonts and cursively written by hand, while preserving the continuity of the resulting skeletons with rich details. The proposed model also noticeably outperforms all peer solutions in skeletonizing characters with overlapping strokes.



(a) (b) (c)

Fig.9. Skeletons generated by three peer methods and the proposed approach for character images from three datasets. (a) Kaiti9574, (b) HW and (c) SkeletonMF. From left to right are the input images, ground truth skeletons (GT) and results by proposed model (Ours), SegNet (Seg)[16], HED[17] and ZhangSuen (ZS)[13].

4.4 Skeletonizing Characters in Small Samples

Considering the label-intensive operations needed for acquiring groundtruth skeletons, a model's capability of learning from a small number of training samples to tackle the skeletonization task is particularly appealing. To explore such capability, we conduct another experiment over the Kaiti9574 dataset where the size of the training set is progressively reduced from 7 000 to 1 000, 200, and 40, respectively. The proposed model and all the three peer methods are applied in this experiment. Fig.10 shows results of this experiment. Except for the classical thinning algorithm ZhangSuen, which does not depend on any training data, all the other three machine learning-based skeletonization methods experience a decayed performance when the size of training samples shrinks. Among the three learning-based methods, skeletons produced by the proposed model get slightly blurred, yet remain at a high visual quality; in contrast, the quality of skeletons produced by the other two learning-based peer models declines significantly when the size of training samples drops.

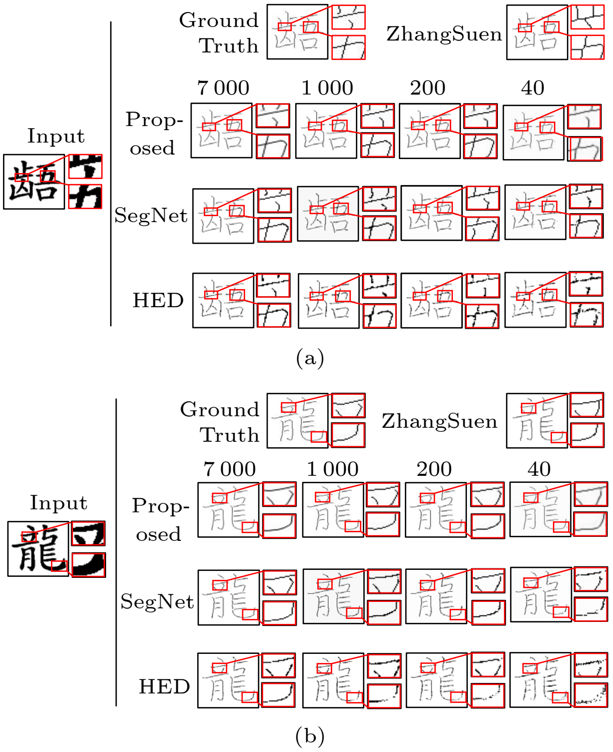


Fig.10. Results generated by the models trained using a progressively smaller set of 7 000, 1 000, 200, and 40 training samples, respectively, on the Kaiti9574 dataset. ZhangSuen[13], SegNet[16], HED[17] are peer methods in which ZhangSuen does not need training samples. (a) Character “齠”. (b) Character “龍”.

4.5 Skeletonizing Characters in Newly Encountered Styles

To explore the generalization capability of the proposed method in skeletonizing characters in newly encountered styles, we carry out two additional experiments.

In the first experiment, real-world calligraphic images are used for evaluation, the results of which are shown in Fig.11. These calligraphic characters exhibit fuzzy boundaries with uneven edges, often accompanied by heavy background noises, whose shapes and structures often deviate noticeably from those printed in standard fonts, all of which make their skeletonization operations much more challenging. It is also noted that calligraphic characters can be written in a vast number of personal styles, which provide a good test-bed to explore an algorithm's capability of skeletonizing characters written in styles not previously witnessed.

When conducting this experiment, we train all three learning-based skeletonization models using the SkeletonMF dataset because the dataset carries 27 diverse looking fonts, making an algorithm more likely to learn to skeletonize characters in style not encountered previously. This hypothesis is supported by additional experiments where either of the other two datasets is used as a training source, which produces compromised outcomes. It is noted that none of the calligraphic writing styles encountered in this experiment is covered in the training dataset. From Fig.11, we can observe that the proposed model achieves visually noticeable advantages over all the three peer methods. The peer method, SegNet, produces much poorer results in this experiment than the proposed method despite SegNet's relatively decent performance in earlier experiments involving characters with previously encountered styles.

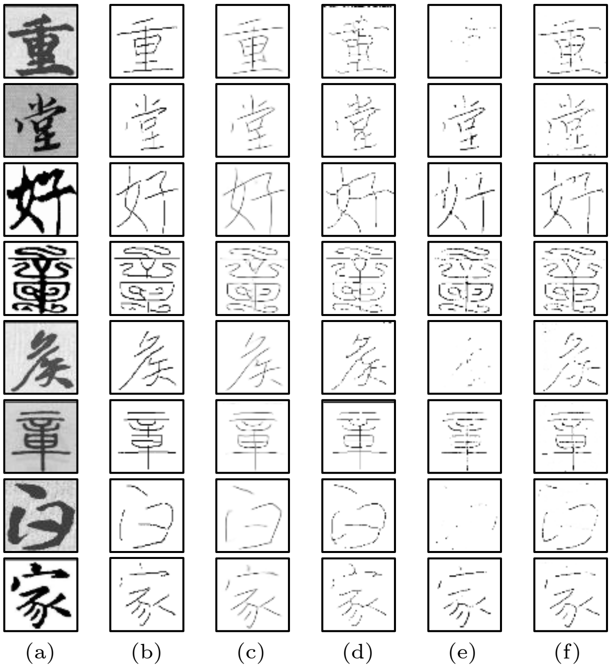


Fig.11. Skeletons extracted from (a) real-world calligraphic characters by different methods. (b) Ground truth. (c) Proposed model. (d) ZhangSuen[13]. (e) SegNet (Seg)[16]. (f) HED[17].

In the second experiment, we test the performance of the proposed model in comparison with that of the peer methods using a cursive handwriting dataset, CASIA-OFFHWDB1.1[27], introduced in [1]. Again, all the three learning-based skeletonization models using the SkeletonMF dataset due to the same empirical reason explain the above. Similarly, none of the handwriting styles encountered in this experiment is covered in the training dataset. Fig.12 gives a few skeletonization results generated by the proposed method in comparison with those by the peer approaches, where the proposed method delivers visually more plausible results. It is noted that no ground truth skeletons are provided in the CASIA-OFFH-

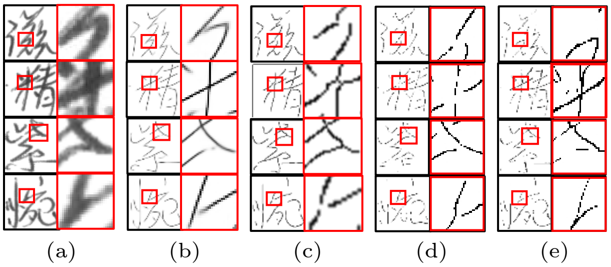


Fig.12. Skeletons extracted for character images in the CASIA-OFFHWDB1.1 dataset[27]. (a) Input. (b) Proposed model. (c) FNCBS[1]. (d)Seg[16]. (e) HED[17].

WDB1.1 dataset. To quantitatively explore the relative performance among all peer solutions, we formulate a panel of 10 human evaluators, proficient in recognizing cursive Chinese handwritings. Each evaluator is invited to assess the visual quality of a character skeletonization result in terms of its structural correctness, the consistency with corresponding input character image (conformance), and the overall plausibility of the skeleton as perceived by the human reader. The assessment outcome is expressed using a five-point Likert scale from 1 (poor) to 5 (excellent). The evaluation results are shown in Table 14, which convincingly demonstrates the superiority of the proposed method to all peer methods compared.

Table 14. Mean of Subjective Opinions Scores for Skeletons Generated by Different Methods

	Structure	Conformance	Plausibility
Proposed	3.8	3.7	3.9
FNCBS[1]	2.7	2.2	2.5
SegNet[16]	1.3	1.5	1.2
HED[17]	1.6	1.8	1.5

Finally, Fig.13 shows skeletons extracted using the proposed approach for the Chinese character “鼎” in 28 writing styles, which comprehensively demonstrates the morphological diversity of Chinese characters as well as the proposed model's ability to cope with such versatile styles.

4.6 Impact on Handwriting Chinese Characters Recognition

To measure the generation quality in another view and explore its impact on downstream tasks, we choose the handwritten Chinese character recognition task, a widely-used Chinese character related task, for testing the generated skeleton of different methods. Specifically, we train ResNet-50[50] by taking the ground truth skeletons in the HW dataset as the in-

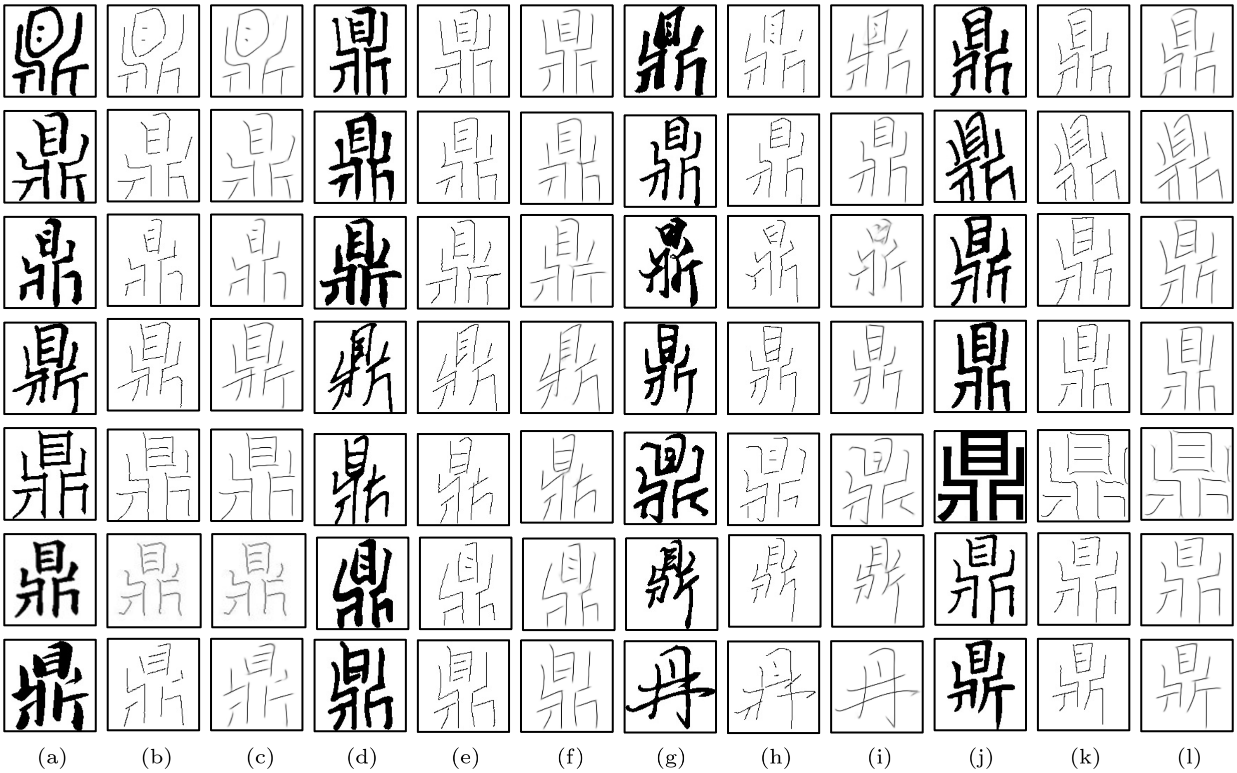


Fig.13. Skeletonization results generated by the proposed model regarding the Chinese characters “鼎” in 28 font styles. (a), (d), (g), and (j) are input images of model, (b), (e), (h), and (k) are corresponding ground truth skeletons, and (c), (f), (i), and (l) are model results.

put of the model and their corresponding character classification labels as the target output. After the training, we feed the skeleton images obtained by different methods to the ResNet-50 classifier and calculate the classification accuracy of the generated skeletons. Since only the ground truth skeletons are used in the training, a higher accuracy indicates that the input skeletons are more similar to the ground truth in the view of recognition, and this fact partially illustrates the potential of the models to be applied in recognition tasks.

The top-1 and the top-5 recognition accuracy reported in Table 15 indicate that skeletons obtained by the proposed model are easily recognized by the model (with the top-1 accuracy of 94.6% and the top-5 accuracy of 99.8%) and are relatively close to the ground truth (with the top-1 accuracy of 96.6% and the top-5 accuracy of 99.8%). Although SegNet performs the best among the three peer methods, its accuracy still has a considerable gap compared with the proposed model. The top-1 recognition accuracy of the rest two methods is less than 50%, indicating their errors have a considerable impact on the correctness of the topology or structure of the skeleton, which is consistent with our previous analysis. We also conduct comparisons between different versions of the proposed model, whose results are displayed in Table 16. The gradually increasing top-1 and top-5 accuracy with the use of more components effectively illustrates the effectiveness of the proposed model.

Table 15. Recognition Accuracy (%) of Skeletons Extracted by Three Peer Methods and Proposed Model

Model	Top-1	Top-5
Ground Truth	**96.6**	**99.8**
Proposed	94.6	99.8
SegNet[16]	84.1	94.1
HED[17]	36.4	51.8
ZhangSuen[13]	46.2	67.4

Table 16. Recognition Accuracy (%) of Skeletons Obtained by Different Versions of Proposed Model

Model	Top-1	Top-5
S.3C	**94.6**	**99.8**
S.3D	**94.6**	99.6
S.3	93.6	99.4
S.2	92.0	99.0
S.2M	93.0	99.2
S.1	90.4	98.2

It is worth noting that the simplest model version S.1 does not outperform SegNet on pixel-wise metrics OFM, OAHD, and OHD, but S.1 significantly exceeds SegNet in the recognition task (90.4% top-1 accuracy compared with 84.1%). This further demonstrates that the proposed model has stronger potential to be applied to downstream tasks compared with the peer methods.

5 Conclusions

This research proposed a novel deep generative model capable of extracting high-quality skeletons of Chinese characters following an image-to-image translation approach. The new model comprises three sequential processing stages, empowered by three deep-learning modules respectively, including an improved G-net module, an adapted X-net module, and a custom-designed convolutional module augmented by an attention mechanism as well as a multiscaled generative pathway. Simultaneously, by using a newly introduced contextual loss with modification as a supplement for pixel-wise loss, the ability of the proposed model to generate better skeletons is further enhanced. As a whole, such a multistage processing pipeline is able to progressively improve the skeletonization result of a character. Its effectiveness is comprehensively demonstrated in comparison with peer methods and different versions of the proposed model, including comparison of numerical metrics and amouts of generated images. The experimental results reported throughout this paper showed that the proposed model is superior to existing methods in generation quality and capabilities in coping with less data and various styles.

Results of an ablation study additionally revealed the respective usefulness of the three constituent modules and newly introduced loss in the proposed generative model. The new model is also particularly well-suited for processing characters with only a small number of training samples, a.k.a. the small-sample learning capability of the new method, as well as characters written in versatile styles previously unseen to the algorithm, a.k.a. the transfer learning capability of the method.

The outstanding performance of our model in the handwriting recognition task showed its promising potential and broad prospects for better completion of downstream tasks. In the future, we will further utilize skeletons and features obtained by the proposed model to enhance the performance of downstream tasks, including handwriting recognition as shown in this paper, and other tasks, e.g., image restoration and segmentation, and style learning and transfer.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] Wang T Q, Liu C L. Fully convolutional network based skeletonization for handwritten Chinese characters. In *Proc. the 32nd AAAI Conference on Artificial Intelligence*, Apr. 2018, pp.2540–2547. DOI: 10.1609/aaai.v32i1.11868.

[2] Xu L, Wang Y X, Li X X, Pan M. Recognition of handwritten Chinese characters based on concept learning. *IEEE Access*, 2019, 7: 102039–102053. DOI: 10.1109/ACCESS.2019.2930799.

[3] Yu K, Wu J Q, Zhuang Y T. Skeleton-based recognition of Chinese calligraphic character image. In *Proc. the 9th Pacific-Rim Conference on Multimedia*, Dec. 2008, pp.228–237. DOI: 10.1007/978-3-540-89796-5_24.

[4] Sun B, Hua S J, Li S T, Sun J. Graph-matching-based character recognition for Chinese seal images. *Science China Information Sciences*, 2019, 62(9): 192102. DOI: 10.1007/s11432-018-9724-7.

[5] Jiang Y, Lian Z H, Tang Y M, Xiao J G. DCFont: An end-to-end deep Chinese font generation system. In *Proc. the 2017 SIGGRAPH Asia Technical Briefs*, Nov. 2017, Article No. 22. DOI: 10.1145/3145749.3149440.

[6] Azadi S, Fisher M, Kim V, Wang Z W, Shechtman E, Darrell T. Multi-content GAN for few-shot font style transfer. In *Proc. the 31st Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.7564–7573. DOI: 10.1109/CVPR.2018.00789.

[7] Zhang Y X, Zhang Y, Cai W B. Separating style and content for generalized style transfer. In *Proc. the 31st Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.8447–8455. DOI: 10.1109/CVPR.2018.00881.

[8] Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In *Proc. the 27th International Conference on Neural Information Processing Systems*, Dec. 2014, pp.2672–2680.

[9] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv: 1411.1784, 2014. https://arxiv.org/abs/1411.1784, Nov. 2023.

[10] Gatys L A, Ecker A S, Bethge M. A neural algorithm of artistic style. arXiv: 1508.06576, 2015. https://arxiv.org/abs/1508.06576, Nov. 2023.

[11] Jiang Y, Lian Z H, Tang Y M, Xiao J G. SCFont: Structure-guided Chinese font generation via deep stacked networks. In *Proc. the 33rd AAAI Conference on Artificial Intelligence*, Jul. 2019, pp.4015–4022. DOI: 10.1609/aaai.v33i01.33014015.

[12] Yuan T L, Zhu Z, Xu K, Li C J, Mu T J, Hu S M. A large Chinese text dataset in the wild. *Journal of Computer Science and Technology*, 2019, 34(3): 509–521. DOI: 10.1007/s11390-019-1923-y.

[13] Zhang T Y, Suen C Y. A fast parallel algorithm for thinning digital patterns. *Communications of the ACM*, 1984, 27(3): 236–239. DOI: 10.1145/357994.358023.

[14] Pujari A K, Mitra C, Mishra S. A new parallel thinning algorithm with stroke correction for Odia characters. In *Proc. the 2nd International Conference on Advanced Computing, Networking and Informatics—Volume 1*, Jun. 2014, pp.413–419. DOI: 10.1007/978-3-319-07353-8_48.

[15] Dong J W, Chen Y M, Yang Z J, Ling B W K. A parallel thinning algorithm based on stroke continuity detection. *Signal, Image and Video Processing*, 2017, 11(5): 873–879. DOI: 10.1007/s11760-016-1034-y.

[16] Badrinarayanan V, Kendall A, Cipolla R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2017, 39(12): 2481–2495. DOI: 10.1109/TPAMI.2016.2644615.

[17] Xie S N, Tu Z W. Holistically-nested edge detection. In *Proc. the 2015 International Conference on Computer Vision*, Dec. 2015, pp.1395–1403. DOI: 10.1109/ICCV.2015.164.

[18] Ke W, Chen J, Jiao J B, Zhao G Y, Ye Q X. SRN: Side-output residual network for object symmetry detection in the wild. In *Proc. the 2017 Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp.302–310. DOI: 10.1109/CVPR.2017.40.

[19] Liu C, Ke W, Qin F, Ye Q X. Linear span network for object skeleton detection. In *Proc. the 15th European Conference on Computer Vision*, Sept. 2018, pp.136–151. DOI: 10.1007/978-3-030-01216-8_9.

[20] Wang Y K, Xu Y C, Tsogkas S, Bai X, Dickinson S, Siddiqi K. DeepFlux for skeletons in the wild. In *Proc. the 2019 Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.5282–5291. DOI: 10.1109/CVPR.2019.00543.

[21] Zhao K, Shen W, Gao S H, Li D D, Cheng M M. Hi-Fi: Hierarchical feature integration for skeleton detection. arXiv: 1801.01849, 2018. https://arxiv.org/abs/1801.01849, Nov. 2023.

[22] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2017, 39(4): 640–651. DOI: 10.1109/tpami.2016.2572683.

[23] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In *Proc. the 2016 Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp.770–778. DOI: 10.1109/CVPR.2016.90.

[24] Mechrez R, Talmi I, Zelnik-Manor L. The contextual loss for image transformation with non-aligned data. In *Proc. the 15th European Conference on Computer Vision*, Sept. 2018, pp.800–815. DOI: 10.1007/978-3-030-01264-9_47.

[25] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014. https://arxiv.org/abs/1409.1556, Nov. 2023.

[26] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proc. the*

31st International Conference on Neural Information Processing Systems, Dec. 2017, pp.6629–6640.

[27] Liu C L, Yin F, Wang D H, Wang Q F. CASIA online and offline Chinese handwriting databases. In *Proc. the 2011 International Conference on Document Analysis and Recognition*, Sept. 2011, pp.37–41. DOI: 10.1109/ICDAR.2011.17.

[28] Lian Z H, Zhao B, Chen X D, Xiao J G. EasyFont: A style learning-based system to easily build your large-scale handwriting fonts. *ACM Trans. Graphics*, 2018, 38(1): Article No. 6. DOI: 10.1145/3213767.

[29] Isola P, Zhu J Y, Zhou T H, Efros A A. Image-to-image translation with conditional adversarial networks. In *Proc. the 30th Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp.5967–5976. DOI: 10.1109/CVPR.2017.632.

[30] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. the 2017 International Conference on Computer Vision*, Oct. 2017, pp.1501–1510. DOI: 10.1109/ICCV.2017.167.

[31] Tang H, Xu D, Sebe N, Wang Y Z, Corso J J, Yan Y. Multi-channel attention selection GAN with cascaded semantic guidance for cross-view image translation. In *Proc. the 2019 Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.2417–2426. DOI: 10.1109/CVPR.2019.00252.

[32] Regmi K, Borji A. Cross-view image synthesis using conditional GANs. In *Proc. the 31st Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp.3501–3510. DOI: 10.1109/CVPR.2018.00369.

[33] Chen K, Pang J M, Wang J Q, Xiong Y, Li X X, Sun S Y, Feng W S, Liu Z W, Shi J P, Ouyang W L, Loy C C, Lin D H. Hybrid task cascade for instance segmentation. In *Proc. the 32nd Conference on Computer Vision and Pattern Recognition*, Jun. 2019, pp.4969–4978. DOI: 10.1109/CVPR.2019.00511.

[34] Liu X B, Qiao Y L, Xiong Y H, Cai Z H, Liu P. Cascade conditional generative adversarial nets for spatial-spectral hyperspectral sample generation. *Science China Information Sciences*, 2020, 63(4): 140306. DOI: 10.1007/s11432-019-2798-9.

[35] Shin H C, Roberts K, Lu L, Demner-Fushman D, Yao J H, Summers R M. Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation. In *Proc. the 2016 Conference on Computer Vision and Pattern Recognition*, Jun. 2016, pp.2497–2506. DOI: 10.1109/CVPR.2016.274.

[36] Cui Z, Chang H, Shan S G, Zhong B N, Chen X L. Deep network cascade for image super-resolution. In *Proc. the 13th European Conference on Computer Vision*, Sept. 2014, pp.49–64. DOI: 10.1007/978-3-319-10602-1_4.

[37] Huang Y X, He M C, Jin L W, Wang Y P. RD-GAN: Few/zero-shot Chinese character style transfer via radical decomposition and rendering. In *Proc. the 16th European Conference on Computer Vision*, Aug. 2020, pp.156–172. DOI: 10.1007/978-3-030-58539-6_10.

[38] Park S, Chun S, Cha J, Lee B, Shim H. Few-shot font generation with localized style representations and factorization. arXiv: 2009.11042, 2020. https://arxiv.org/abs/2009.11042, Nov. 2023.

[39] Gao Y M, Wu J Q. GAN-based unpaired Chinese character image translation via skeleton transformation and stroke rendering. In *Proc. the 34th AAAI Conference on Artificial Intelligence*, Feb. 2020, pp.646–653. DOI: 10.1609/aaai.v34i01.5405.

[40] Sun D Y, Ren T Z, Li C X, Su H, Zhu J. Learning to write stylized Chinese characters by reading a handful of examples. arXiv: 1712.06424, 2017. https://arxiv.org/abs/1712.06424, Nov. 2023.

[41] Zhang J W, Chen D N, Han G Q, Li G Z, He J T, Liu Z M, Ruan Z H. SSNet: Structure-semantic Net for Chinese typography generation based on image translation. *Neurocomputing*, 2020, 371: 15–26. DOI: 10.1016/j.neucom.2019.08.072.

[42] Xu S H, Lau F C M, Cheung W K, Pan Y H. Automatic generation of artistic Chinese calligraphy. *IEEE Intelligent Systems*, 2005, 20(3): 32–39. DOI: 10.1109/MIS.2005.41.

[43] Xu S H, Jin T, Jiang H, Lau F C M. Automatic generation of personal Chinese handwriting by capturing the characteristics of personal handwriting. In *Proc. the 21st Innovative Applications of Artificial Intelligence Conference*, Jul. 2009, pp.191–196.

[44] Xu S H, Jiang H, Jin T, Lau F C M, Pan Y H. Automatic generation of Chinese calligraphic writings with style imitation. *IEEE Intelligent Systems*, 2009, 24(2): 44–53. DOI: 10.1109/MIS.2009.23.

[45] Xu S H, Jiang H, Lau F C M, Pan Y H. An intelligent system for Chinese calligraphy. In *Proc. the 22nd National Conference on Artificial Intelligence*, Jul. 2007, pp.1578–1583.

[46] Li B, Chen H H, Chen Y C, Dai Y C, He M Y. Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In *Proc. the 2017 IEEE International Conference on Multimedia and Expo Workshops*, Jul. 2017, pp.613–616. DOI: 10.1109/ICMEW.2017.8026283.

[47] Xu W J, Parmar G, Tu Z W. Geometry-aware end-to-end skeleton detection. In *Proc. the 30th British Machine Vision Conference*, Sept. 2019, pp.28.1–28.13. DOI: 10.5244/C.33.28.

[48] Lin T Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In *Proc. the 30th Conference on Computer Vision and Pattern Recognition*, Jul. 2017, pp.2117–2125. DOI: 10.1109/CVPR.2017.106.

[49] Woo S, Park J, Lee J Y *et al.* CBAM: Convolutional block attention module. In *Proc. the 15th European Conference on Computer Vision*, Sept. 2018, pp.3–19. DOI: 10.1007/978-3-030-01234-2_1.

[50] He K M, Zhang X Y, Ren S Q, Sun J. Identity mappings in deep residual networks. In *Proc. the 14th European Conference on Computer Vision*, Oct. 2016, pp.630–645. DOI: 10.1007/978-3-319-46493-0_38.

Ye-Chuan Tian received his B.S. degree in mathematics and applied mathematics from Xi'an Jiaotong University, Xi'an, in 2018. He is a Ph.D. candidate at the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an. His current research interests include deep learning based Chinese character representation and generation.

Song-Hua Xu received his Ph.D. degree in computer science from Yale University, New Haven, in 2010. His research interests include healthcare informatics, information retrieval, knowledge management and discovery, intelligent web and social media, visual analytics, and user interface design.

Cheickna Sylla received his M.S. and Ph.D. degrees in applied operations research and industrial engineering from the State University of New York, Buffalo, in 1983. He is currently a professor of Decision Sciences and MIS (Management Information System) at the Martin Tuchman School of Management, New Jersey Institute of Technology, Newark. His research interests include decision sciences, management information systems, knowledge based systems, analysis and evaluation of human machine systems, supply chain management, and distribution logistics. He has published over 86 peer reviewed journal and conference articles.