Journal of computer science and technology: Instruction for authors. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY

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

Abstract Character skeletons provide valuable information for a number of character processing tasks, e.g. optical character recognition, image restoration and segmentation, style learning and transfer, as well as identity authentication and verification based on handwriting analysis, to name but a few. However, automatically skeletonizing Chinese characters poses a steep computational challenge due to the large volume of Chinese characters, the great variability in ways strokes may combine spatially and topologically to compose a character, and the wide assortment of writing styles employed by billions of writers in everyday use. Traditional image analysis and machine learning approaches developed before the deep learning era are error-prone and fragile in dealing with characters with intersecting strokes, a phenomenon exhibited in the majority of characters. Current deep learning-based approaches require a heavy amount of manually labelled training samples, which imposes serious limitations on the scalability and generability of an algorithm in learning to skeletonize characters in versatile styles.

To overcome the above caveats, this paper introduces a novel three-staged deep generative model to extract high-quality skeletons of Chinese characters. The new model presents a novel image-to-image translation approach, which significantly reduces the model's demand for training samples, an algorithm trait particularly appealing for learning to skeletonize characters in versatile styles. The new model is built upon an improved G-net, an adapted X-net, and a newly proposed F-net, which comprises a deep convolutional structure nested with a multiresolution synthesis paradigm, coupled with a channel attention module and another spatial attention module. Empowered by the three networks, each responsible for a sequential processing stage of the model, the new algorithm is able to progressively extract skeletons of Chinese characters in versatile styles with a high quality and generability. Trained with a newly introduced modificated contextual loss served as a supplement for pixel-wise loss, the proposed model's skeletonization ability is further enhanced. Comprehensive experimental results convincingly demonstrate that the new approach outperforms two state-of-the-art deep learning methods and a classical thinning algorithm for character skeletonization. The performance advantage of the new model is especially evident when applied to skeletonize characters in cursive handwriting styles and calligraphic work.

Keywords Contextual Loss, Convolution with Attention, Skeletonization of Characters in Versatile Styles, Three-Staged Skeletonization, X-net

1 Introduction

The skeleton of a character conveys rich and essential structural and shape clues for informing the computational processing of its image, e.g. optical character recognition, image restoration and segmentation, style learning and transfer, as well as identity authentication and verification based on handwriting analysis. Skeletons of Chinese characters are no exception in providing these essential merits. For example, their skeletons have been utilized to significantly improve the recognition performance of handwritten and calligraphic characters [1-3]. Skeletons can also bring valuable informational aid to improve the quality of style transfers among characters [4-6], greatly enhancing the capability of standard deep generative models [7-9] in dealing with Chinese characters in versatile styles and complex structures. These empirical advantages are further supported by the understanding that skeletons provide priors as a secondary input, which are generally useful for augmenting an end-to-end deep style learning and transfer model [10].

Unfortunately, the plentiful merits of skeletons of Chinese characters are largely underexplored in today's Chinese character processing because of the inhibitive cost of manual acquisition of these skeletons as well as the limitation of today's algorithms in their automatic

Regular Paper

extraction, both tasks of which are severely hampered due to the large volume of Chinese characters and the wide assortment of writing styles deployed in everyday use. For example, the standard character set used in mainland China, GB2312-80, includes 6763 most frequently used characters while the total number of Chinese characters is above 50,000 [11]. The form of a Chinese character can also be heavily influenced when written in different styles. Fig.13. shows the appearance of a Chinese character, "Ding", an ancient vessel with two handles and three or four legs that symbolizes the throne of a kingdom, written in 28 well-recognized styles, the example of which demonstrates the wide variation in the potential look of a character in versatile styles. Countless handwritten styles have been invented through the long history of the language use [12], the fact of which further adds to the difficulty of the task.

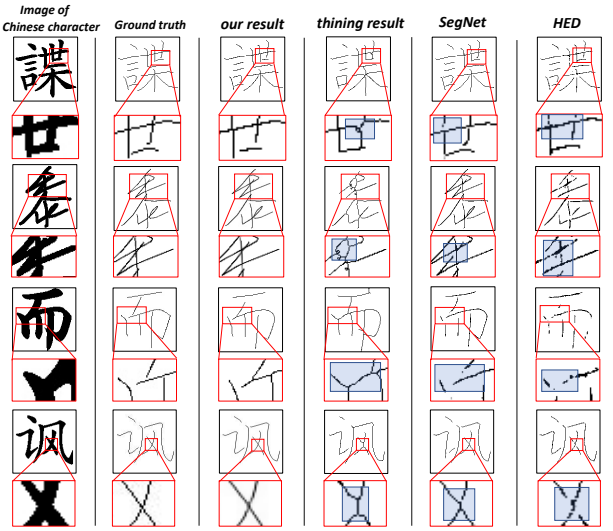

Fig.1. Skeletonization results by the proposed method in comparison with the ground truth skeletons and skeletons generated by three peer methods—a popularly used thinning algorithm (ZhangSuen [13]) and two leading deep learning solutions, SegNet [21] and HED [16]. Areas displaying problematic skeletonization results are shaded in blue.

To computationally skeletonize a Chinese character, 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 its middle axis. Followup research modified the thinning rule used in the original work, resulting in a large number of algorithm variants, e.g. [14, 15]. The benefit of the class of thinning algorithms is their data free nature, meaning that no training data is required, whose role is replaced by smartly designed heuristics carefully yet manually encoded into an algorithm. Unfortunately, no known heuristics are able to reliably cope with the diversity and complexity of all stroke combination scenarios in Chinese characters. Hence, thinning algorithms are generally error-prone and fragile in processing characters with intersecting strokes, a phenomenon occurring in the majority of characters. Fig.1. shows skeletons extracted for a few sample characters using the algorithm of [13] where the enlarged inserts show erroneous skeletonization results in areas of overlapping strokes.

Most recently, [1] introduced a deep learning-based solution for skeletonizing Chinese characters to address the above limitation. Their method relies upon a handwritten Chinese character recognition task to pre-train its model. The derived features are subsequently repurposed for the skeletonization task. To successfully conduct the first recognition task, their method requires the provision of the category information for a large number of Chinese characters. Such a heavy demand for training data severely limits the transferability and generalizability of their method in processing characters of versatile styles. On a separate line of research, holistically-nested edge detector (HED) [16] was designed for edge detection, which was found to have good effects on detecting skeletons of objects and thus became a popularly used model. Recent methods of object skeleton detection [17-20] are mainly based on

HED. However, these revised approaches still fail to reliably extract character skeletons. SegNet [21] is a classic algorithm for semantic segmentation, developed under an image-to-image translation framework. Compared with the similar method [22, 23], it has fewer layers and faster inference speed, which is suitable for the task of skeleton extraction because it does not require much high-level semantic information. Considering the single-pixel width nature of the character skeleton, [1] uses binary and thinning to post-process the probability map output by the network, resulting in unnatural distortion as shown in Fig.12, while skeletons extracted by the SegNet and HED methods have many breakpoints and uneven skeleton trajectories.

To overcome shortcomings of the above state-of-the-art solutions, this study introduces a novel deep generative model to extract high-quality skeletons of Chinese characters following an image-to-image translation approach, 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 respectively leverages a modified G-net, an adapted X-net, and a newly proposed F-net, which comprises a deep convolutional structure nested with a multiresolution synthesis paradigm, coupled with a channel attention module and another spatial attention module, to conduct the skeltonization task with progressively improved quality. The main contribution of the proposed Chinese character skeletonization solution lies in its novel deep neural network design. Enabled by its newly introduced three stages of the progressive image-to-image generation process, along with the specific deep neural network structure carefully designed for each stage, the proposed solution is capable of attaining a noticeably superior quality in its skeletonization results in comparison with results generated by state-of-the-art peer methods. Benefited by such an innovative deep network design, the proposed approach is the first algorithm capable of effectively skeletonizing Chinese characters in versatile styles with a satisfactory quality, a task beyond the reach of all existing algorithms to our best knowledge.

In addition to the above contributions stated in the earlier version [1] accepted by CVM 2021 [2], we newly introduce the contextual loss [39] which calculates the similarity of images by comparing their derived feature maps without requirement of spatial alignment. An autoencoder trained with skeleton data is proposed to obtain features for computing of contextual loss, which evidently outperforms the VGG19 [41] used in original work[39]. Empowered by the contextual loss and specially designed feature acquisition method, the proposed model achieve better performance under the measure of newly added Frechet Inception Distance (FID) [40], which is considered to be close to human perception and is therefore widely used in image generation tasks.

It is also noteworthy that the new algorithm requires a much smaller size of training samples than its peer methods to attain at least a comparable visual quality. For example, the leading deep learning-based peer method [1] needs 1.121 million training pairs of Chinese character images and their associated skeletons to learn to skeletonize handwritten characters when applied to the dataset of [24] while the proposed approach only needs to be trained using 0.14 million such training pairs, a reduction of training size by 87.5%. For two other smaller datasets, including skeletonMF [25] with 13500 pairs of training samples and Kaiti9574 [26] with

[1]please refer to followed conference-version.pdf for earlier version and detailed-extension.pdf for the detail of extensions

[2]The 9th international conference on Computational Visual Media

7000 pairs of training samples, the peer method [1] cannot be adequately trained to obtain decent skeletonization results. In contrast, the proposed method is still able to produce visually satisfactory results as shown later in the experimentation section of this paper. It is noted that for the Kaiti9574 dataset, the smallest among the three, the new method is able to learn using only 40 training samples to produce visually acceptable results (see Fig.10.). These results consistently demonstrate the capability of the new approach in learning using a much smaller sample size than its peer deep learning alternatives.

The rest of the paper is 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 with detail. Section 4 shows the experimental results of the new algorithm in comparison with results by state-of-the-art peer methods. Finally, Section 5 concludes this paper.

2 Related Work

2.1 Image-to-Image Translation

Image-to-image translation aims to learn some 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 in the other domain without distorting the underlying image content. For example, the PIX2PIX work [27] applies conditional adversarial networks as a general-purpose solution to tackle image-to-image translation tasks. Their approach 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 [28], cross-view image translation [29, 30], as well as font generation [4-6]. In this study, we perceive the image of a character and its corresponding skeleton as an object manifested in two image domains, under which perspective the character skeletononization task is converted as an image to 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 which circumstance a cascade of networks is sometimes leveraged to progressively synthesize a desirable result, e.g. [10, 29, 31-34]. For our character skeletonization task, by definition, the width of a skeleton needs to be of a single pixel width. It is therefore difficult for a single generative network to accurately produce a skeleton in one shot. Hence, utilizing a cascade of generative networks may provide a more capable solution, the idea of which inspires the three-staged generative deep network approach proposed in this paper. The comprehensive experimental results presented later in this paper indeed verify the effectiveness of such a design approach.

2.2 Skeleton Detection

Skeleton extraction and detection are intensively investigated in computer vision and image processing, under tasks such as action detection [35] and natural object skeletonization [20, 36], 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, many of which are based on an end-to-end edge detection algorithm, named HED [16] and many of its variants and improvements [17-20,

36]. Among the followup works of HED, [17] leverages a bidirectional residual learning scheme; [20] adopts a hierarchical fusion procedure; [36] employs a geometry-based loss function. Very limited attention has been paid however to extracting skeletons of Chinese characters. To our best knowledge, [1] is the only deep learning-based approach solely developed for the purpose of skeletonizing Chinese characters. As discussed earlier, their approach 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.

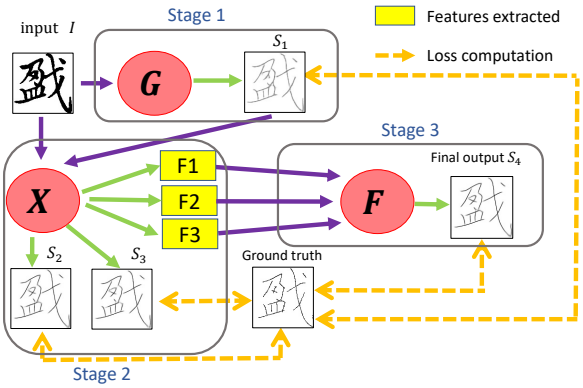

Fig.2. Architecture of the proposed model. See detail of G, X and F respectively in Fig. 3–5.

3 The Proposed Model

Fig.2. shows the main architecture of the three-staged proposed model for skeletonizing Chinese characters in versatile styles. G, X and F respectively represent a pre-generation network, a refined X-net and multiresolution feature fusion net (see detail in Fig. 3–5). These three networks, which are respectively referred to as the G-Net, X-net, and F-net from now on, sequentially power each of the three key generation stages of the new model.

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 of 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 respectively derived by the X-net.

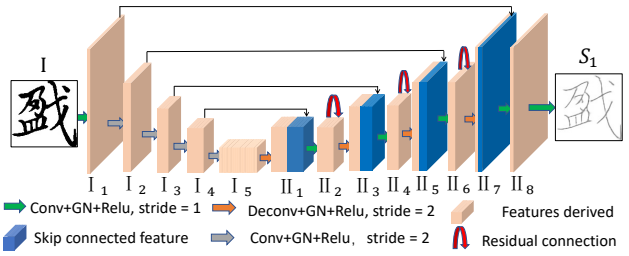

Fig.3. Architectural detail of the pre-generation network, G. G outputs a preliminary version of the skeleton S_1 from an input character image I. The dimensions for various feature maps generated by G are specified in Table 1 following indices provided by Roman letters with subscripts.

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 resemblance 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 of the ability of U-net in sensitively responding 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

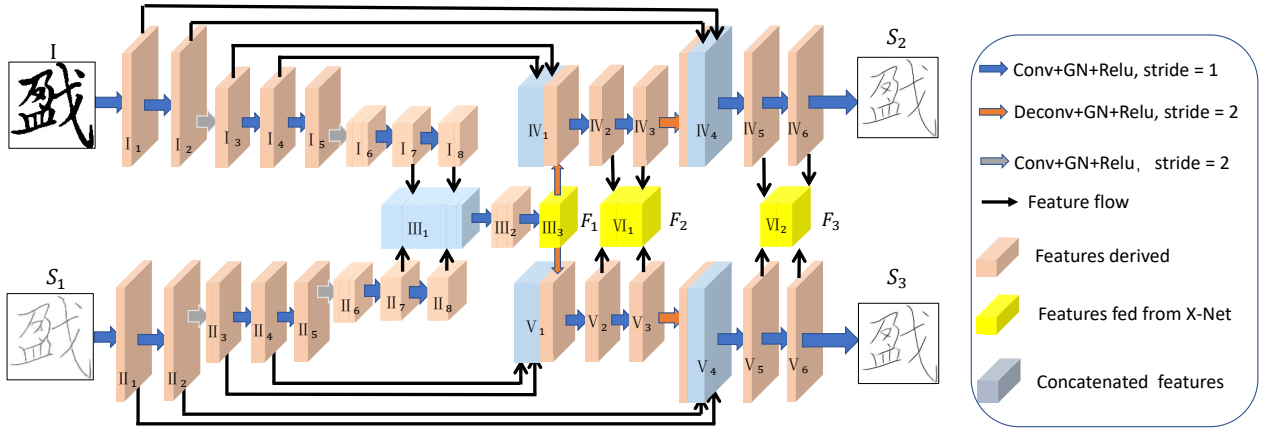

Fig.4. Architectural detail of the refined X-net. The two input branches of the network respectively accept the image of a Chinese character I and a preliminary version of its skeleton S_1 extracted by the upstream network (G). The network generates a pair of output S_2 and S_3 through a visual information encoding and exchange procedure. In addition, the network also derives three sets of features F_1, F_2 and F_3, to be used by its downstream network F. The dimensions for various feature maps generated by X-net are specified in Table 1 following indices provided by Roman letters with subscripts.

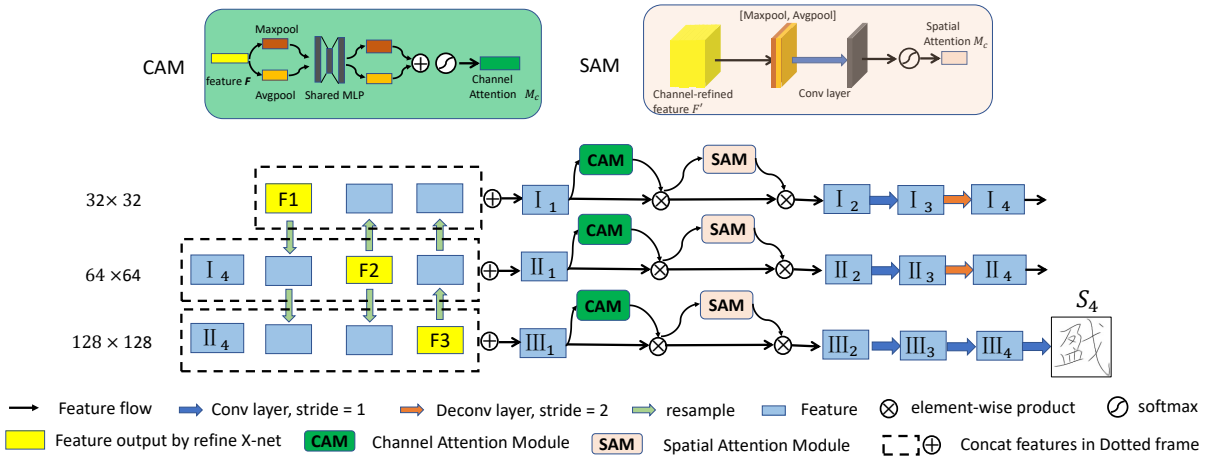

Fig.5. Architectural detail of the attention-based multiresolution feature fusion net, F. F_1, F_2 and F_3 are features extracted by the refined X-net as shown in Fig.4. while S_4 is the final output of the model. The channel attention module (CAM) and spatial attention module (SAM) used in F are both adopted from the Convolutional Block Attention Module proposed in [38].

its underlying skeleton as well as the skeletons of its intersecting strokes, if they spatially close by inside the character image.

In our design for the G-net, we modified 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 effectivenss of which is empricially validated through our experiments. The end design of 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 5 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 for the reason discussed in the above. Since an ideal skeleton shall assume only a single pixel width, we use a step size of 2

in the deconvolution layers to mitigate degradation of visual quality during the upsampling process.

Table 1. Dimensions of feature maps in the proposed model. FN: feature name, Res: resolution, CN: channel number.

Dimensions for features in the G-net.

FN	I_1	I_2	I_3	I_4	I_5	II_1	II_2	II_3	II_4	II_5	II_6	II_7	II_8
Res	128	64	32	16	8	16	16	32	32	64	64	128	128
CN	8	16	32	64	128	128	64	64	32	32	16	16	8

Dimensions for features in the X-net

FN	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	II_1	II_2	II_3
Res	128	128	64	64	64	32	32	32	128	128	64
CN	8	8	16	16	16	32	32	32	8	8	16
FN	II_4	II_5	II_6	II_7	II_8	III_1	III_2	III_3	III_4	III_5	IV_1
Res	64	64	32	32	32	32	32	32	64	128	64
CN	16	16	32	32	32	128	32	32	64	32	48
FN	IV_2	IV_3	IV_4	IV_5	IV_6	V_1	V_2	V_3	V_4	V_5	V_6
Res	64	64	128	128	128	64	64	64	128	128	128
CN	16	16	24	8	8	48	16	16	24	8	8

Dimensions for features in the F-net

FN	I_1	I_2	I_3	I_4	II_1	II_2	II_3	II_4	III_1	III_2	III_3	III_4
Res	32	32	32	64	64	64	64	128	128	128	128	128
CN	112	112	64	32	144	144	64	32	144	144	64	8

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 both in 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 into a deep network will not accomplish the aim with the same quality, which is both empirically verified through our experiments 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 source. To adequately explore potentially useful information from both sources, we introduce the X-net.

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 of which 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 detailed construction of the X-net is composed by three parts, including an encoding, a fusion, and a decoding part. 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. and labelled with Roman letters of I's and II's have the same structure, they do not share any parameters to the respect 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 the figure, features with a resolution of 32×32 respectively generated by the two encoding branches are concatenated first and then fed into the convolutional layers for 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 con-

nections 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 output by the X-net. Instead, it only uses the three sets of features extracted by X-net, F_1 (32×32), F_2 (64×64) and F_3 (128×128), both 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. Benchmarked experimental results demonstrate the advantage of using 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.

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 profound impact on the quality of the end skeletonization result, a trait frequently exhibited by deep learning-based image processing algorithms [19]. 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 detail; 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 [38], $F_1(32\times32)$ is concatenated with $F_2(32\times32)$ and $F_3(32\times32)$, the result of which 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, $F_{1\oplus2\oplus3}(32\times32)$, at the resolution of 32×32. $F_{1\oplus2\oplus3}(32\times32)$ is then upsampled to the resolution of 64×64, resulting in $F_{1\oplus2\oplus3}(64\times64)$. $F_{1\oplus2\oplus3}(64\times64)$ is subsequently concatenated with $F_1(64\times64)$, $F_2(64\times64)$, $F_3(64\times64)$, 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, $F_{1\oplus2\oplus3}(128\times128)$, at the the resolution

of 128×128. Lastly, $F_{1 \oplus 2 \oplus 3}(128 \times 128)$ is concatenated with $F_1(128 \times 128), F_2(128 \times 128), F_3(128 \times 128)$, followed by similar transformations carried out by the above two attention modules and the convolution layers to yield the final skeletonization result S_4.

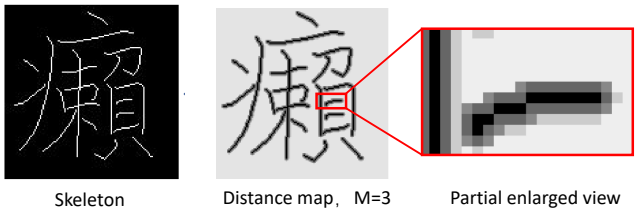

Fig.6. Example of skeleton and corresponding distance map with partial enlarged view.

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, the following loss function is employed due to its wide adoption in the literature [1]:

$$loss_{total1} = \sum_{i=1}^{4} \alpha_i \cdot loss_{CE}(S_i, GTS) \tag{1}$$

where GTS is the ground truth skeleton corresponding to an input character image I, $loss_{CE}$ 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 Eq. (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 Eq. (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 caveat, we introduce a novel distance-based loss function, which has not been used in prior studies at 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 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:

$$pixel(q_j) = \frac{0.9}{D} \times min(\min_{p_i \in P} dis(q_j, p_i), D) \tag{2}$$

where D is the threshold parameter. 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 negative and positive samples, otherwise significantly unbalanced in our problem.

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

$$loss_{dis}(S_i) = loss_{MSE}(S_i, GTD) \tag{3}$$

where $loss_{MSE}$ is the mean squared error term and total loss function for any candidate skeletonization result GTS:

$$loss_{total2} = \sum_{i=1}^{3} \alpha_i \cdot loss_{dis}(S_i) + \alpha_4 \cdot loss_{CE}(S_4, GTS) \tag{4}$$

The above loss function Eq. (3) takes into account the geometric severity of the errors, but still does not adequately reflect the impact of the errors on the skeleton topology. Consider the errors that make two separated strokes intersect, they change the original topology of the skeleton, thus 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 [39], which employs a novel feature-based method to compare the similarity between images without 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. The contextual similarity between images is defined as below:

$$CX(x, y) = CX(X, Y) = -log(\frac{1}{N}\sum_j \max_i CX_{ij}) \quad (5)$$

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, the similarity CX_{ij} between x_i and y_j is defined as follows:

$$\begin{aligned} \tilde{d_{ij}} &= \frac{d_{ij}}{\min_k d_{ik} + \epsilon} \\ w_{ij} &= exp(\frac{1 - \tilde{d_{ij}}}{h}) \\ CX_{ij} &= w_{ij} / \sum_k w_{ik} \end{aligned} \quad (6)$$

where ϵ and h is fixed hyper-parameters. Eq. (5) employs best similarity $\max_i CX_{ij}$ to measure the similarity between y_j and X instead of using spatially aligned CX_{jj} directly, and consequently enforce the model to pay more attention to structural similarity rather than strict spatial correspondence, which also emphasized by the design of CX_{ij} in Eq. (6). Intuitively, features tend to represent the informative skeleton points instead of background point, 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 skeleton, while VGG19 [41] training on image-net employed by original paper [39] brings no promising improvement. Considering the significant difference between skeleton images and real world images, we propose to utilize autoencoder 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 autoencoder is dedicated to reconstruct the input skeleton, thus the acquired features tend to include all skeleton information, which also ensures that the comparison between images is comprehensive enough.

Specifically, The autoencoder consists of an encoder E_{cx} and a decoder D_{cx}, where E_{cx} includes five convolutional layers $\{E_{cx,l}\}_{l=1}^M$ with a step size of 2. Given a skeleton image s, the sequence of feature maps $\{\phi^l(s)\}_{l=1}^M$ is computed step by step as $\phi^l(s) = E_{cx,l}(\phi^{l-1}(s))$ 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 the E_{cx} and D_{cx} as below:

$$loss_{auto}(s_{fake}, s) = loss_{CE}(s_{fake}, s) \quad (7)$$

The final contextual loss between generated image x ad target image y using the above-mentioned autoencoder is computed as:

$$loss_{CX}(x,y) = \sum_{l} CX(\phi^l(x), \phi^l(y)) \tag{8}$$

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

We apply contextual loss as a supplement of the aforementioned losses in third stage and the corresponding overall loss for candidate skeletonization result GTS is as follow:

$$\begin{aligned} loss_{total3} = &\sum_{i=1}^{3} \alpha_i \cdot loss_{dis}(S_i) + \alpha_4 \cdot loss_{CE}(S_4, GTS) \\ &+ \alpha_5 \cdot loss_{CX}(S_4, GTS) \end{aligned} \tag{9}$$

4 Experimentation

4.1 Implementation Detail

4.1.1 Datasets

Three data sets, Kaiti9574, HW and SkeleonMF are used in our experiments. The Kaiti9574 dataset, collected from the Make-Me-a-Hanzi project [26], contains images of 9574 kaiti characters and their corresponding skeletons. 7000 randomly selected characters are used for the training purpose. The SkeletonMF dataset provided in [25] contains 27 fonts, each of which has 639 characters. 500 characters are randmoly selected 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 release by [1], which carries a total of 220,000 online trajectories for handwritten Chinese characters. 140,000 randomly selected characters are used as training samples in our experiment.

4.1.2 Experimental Setup

The proposed model is implemented in PyTorch. All experiments were conducted on Nvidia GeForce RTX 2080 Ti GPU. The learning rate used for training the proposed network has an initial value of 0.0002. Unless otherwise specified, each model training takes 10 epochs.

To verify the effectiveness of various components in the proposed model, ablation analysis was conducted. A series of variations of the proposed model is 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 respectively refer to the version of the proposed model that uses the output of Stage 2 and Stage 3 as the model output. For the two outputs generated by Stage 2 of the model, the one that yields a higher performance metric is used as the output for S.2. Note that S.3 is the full model proposed. Finally, 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 detail discussed in Sec. 3) while S.3D refers to a version that turns the ground truth of the first two stages into distance map as described in Sec. 3. Finally, the S.3C refers to a version that adds contextual loss (see Sec. 3.4) to S.3D with autoencoder pre-trained on skeleton data, while S.3CV that adopts contextual loss with VGG19 pre-trained on image-net is also trained for comparison.

4.1.3 Evaluation Metrics

Three experimental metrics, including the Frechet Inception Distance (FID) [40], 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.

F-measure gives a description of accuracy in the pixel-wise level, which however could not represent the geometric similarity between two skeletons, a consideration equally important for our task. Hausdorff distance is a measure to describe the similarity between two sets of points. Assume that p_s and $\widetilde{p}_s$ are respectively the

two sets of skeleton pixels in a ground truth skeleton and the corresponding skeletonization result. For p_s and $\widetilde{p}_s$, HD is computed as:

$$HD(p_s, \widetilde{p}_s) = max(\max_{b \in p_s} \min_{a \in \widetilde{p}_s} d(b,a), \max_{a \in \widetilde{p}_s} \min_{b \in p_s} d(a,b)), \tag{10}$$

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:

$$\begin{aligned} &AHD(p_s, \widetilde{p}_s) = ahd(p_s, \widetilde{p}_s) + ahd(\widetilde{p}_s, p_s) \\ &= \frac{1}{|p_s|} \sum_{b \in p_s} \min_{a \in \widetilde{p}_s} d(b,a) + \frac{1}{|\widetilde{p}_s|} \sum_{a \in \widetilde{p}_s} \min_{b \in p_s} d(a,b). \end{aligned} \tag{11}$$

The probability prediction map output 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 rang of the threshold value, i.e. let $\tau = 0.01, 0.02, \cdots, 0.99$ 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 Eq. (11), $ahd(p_s, \widetilde{p}_s)$ and $ahd(\widetilde{p}_s, p_s)$, are treated as the coordinates of a point on curve, hence the name of the curve. Lastly, we also calculated an optimal F-measure (OFM), an optimal HD score (OHD), and an optimal AHD score (OAHD), which respectively stand for the highest F-measure, HD score and AHD score encountered during the exhaustive search of the aforesaid threshold τ.

We also adopt the Frechet Inception Distance (FID) [40] commonly-used in image generation tasks for assessing the generation quality. FID is employed to measure the distance between images in feature level, and is consequently closer to human perception than pixel-wise metrics. The smaller FID, the better the quality of the generated result in terms of the similarity between features of a generated image and those of the corresponding ground truth image.

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 compared the relative performance among five alternative versions of the proposed model, including S.1, S.2M, S.2, S.3 and S.3D (see Sec. 4.1 for detail), using the three experimental datasets.

Table 2. The FID, OFM, OAHD and OHD scores of skeletonization results by different versions of model. S.1, S.2M, S.2, S.3 are various versions of the proposed model. A larger OFM, a smaller FID, OAHD and OHD values all indicate a better skeletonization result.

kaiti9574

MODEL	FID	OFM	OAHD	OHD
S.1	80.7	0.726	0.552	6.06
S.2M	72.4	0.754	0.481	4.60
S.2	67.8	0.760	0.472	4.50
S.3	**61.8**	0.774	0.446	4.21
S3.D	63.7	**0.777**	**0.438**	**4.02**

HW

MODEL	FID	OFM	OAHD	OHD
S.1	39.8	0.891	0.26	3.82
S.2M	27.9	0.889	0.26	3.22
S.2	28.0	0.911	0.208	3.14
S.3	18.8	0.923	0.179	**2.98**
S3.D	**18.6**	**0.925**	**0.174**	3.01

skeletonMF

MODEL	FID	OFM	OAHD	OHD
S.1	165.2	0.498	1.26	10.54
S.2M	155.6	0.515	1.21	9.95
S.2	154.8	0.520	1.162	9.84
S.3	148.5	0.525	1.143	9.69
S3.D	**144.9**	**0.529**	**1.125**	**9.58**

Table 2. shows the respective performance of these five versions of the proposed model quantitatively evaluated using FID, OFM, OHD and OAHD. It is also noted that these last three numerical metrics are sensitive to the particular binarization threshold applied at the final output stage of the proposed network, (see Sec. 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 curve and average Hausdorff distance curve in Fig.7. for each model version applied in each experiment.

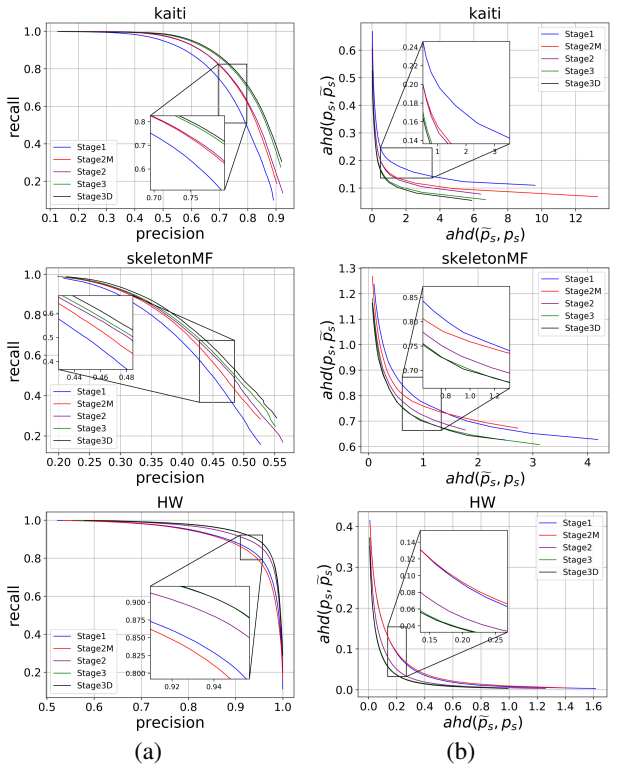

Fig.7. The precision-recall curve and average Hausdorff distance curve of different module evaluated on three dataset. (a) is precision-recall curve while (b) shows average Hausdorff distance curve.

According to performance measurements both numerically reported in Table 2. 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 an X-net is employed in S.2 but not S.1, the above performance advantage shows the usefulness of X-net. It is also recognized from Table 2. and Fig.7. that S.2 is consistently superior to S.2M. Since the only 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, 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, as X-net differs from Y-net only in that the former network has two output branches while the latter network has a single output branch. Lastly, Table 2. and Fig.7. also show that S.3D generally outperforms S.3, which indicates that the distance map-based loss function defined in Eq. (3) indeed helps improve the overall skeletonization capability of the proposed method.

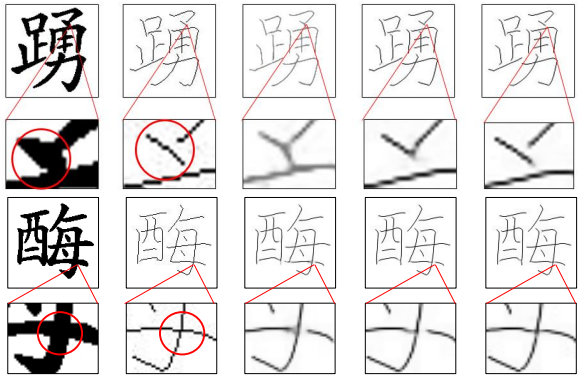

Fig.8. Results by the model in different stages. From left to right: input, ground truth, output of Stage 1, Stage 2 and Stage 3, respectively.

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 the skeleton generation. Three versions of the proposed model, S.3C, S.3CV, S.3D (see Sec.4.1), are trained on three datasets and evaluated among metrics FID, OFM, OHD and OAHD. S.3C, S.3CV utilize contextual loss with pre-training model using skeleton data and image-net data respec-

tively as described in Sec.3.4, and S.3D is trained without contextual loss.

Table 3. The FID, OFM, OAHD and OHD scores of skeletonization results by different versions of model. S.3D, S.3C, S.3CV are various versions of the proposed model. A larger OFM, a smaller FID, OAHD and OHD values all indicate a better skeletonization result.

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

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

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

Table 3. shows the respective performance of these three versions of the proposed model quantitatively evaluated using FID, OFM, OHD and OAHD on three datasets. According to the numerical performance shown in the Table 3, S.3C performs consistently better than S.3D on FID, which is considered closer to human perception, and performs similarly with S.3D on other metrics. Compared with S.3D, S.3C only adds contextual loss, thus the above advantages indicate that the skeletonization performance of 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 [39] trained with image-net data to replace the autoencoder trained with 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 kaiti9574 dataset. 7000, 1000, 200, 40 training samples of kaiti9574 are employed to train the S.3D and S.3C, and the respective performance evaluated are shown in Table 4. The result indicates that the reduction in data size impairs the performance of the model on all metrics, however, S.3C is less affected. For example, the S.3C performs slightly weaker than S.3D on OFM with 7000 training samples (0.774 compared to 0.777), but performs similarly with 1000 (0.728 compared to 0.727), and clearly outperforms S.3D with 200 or 40 training samples (0.705, 0.674 compared to 0.691, 0.643). The expanded experiment shows that the contextual loss has stronger advantages in the scenario with small size of dataset.

Table 4. The FID, OFM, OAHD and OHD scores of skeletonization results by different versions of model S.3D and S.3C on kaiti9574 dataset. 7000, 1000, 200, 40 are numbers of samples used for training. A larger OFM, a smaller FID, OAHD and OHD values all indicate a better skeletonization result.

MODEL	FID	OFM	OAHD	OHD
S.3D(7000)	63.7	0.777	0.438	4.02
S.3C(7000)	56.9	0.774	0.442	3.98
S.3D(1000)	81.4	0.727	0.593	5.931
S.3C(1000)	66.4	0.728	0.541	4.873
S.3D(200)	90.0	0.691	0.641	7.123
S.3C(200)	74.9	0.705	0.616	7.454
S.3D(40)	107.9	0.643	0.770	8.387
S.3C(40)	82.6	0.674	0.707	8.634

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

Next, we compare the performance of the proposed model with that of multiple state-of-the-art peer methods, including the classical thinning algorithm Zhang-Suen introduced in [13] and two recently proposed deep learning-based skeletonization models—HED [16] and SegNet [21]. 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.

Table 5. shows 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

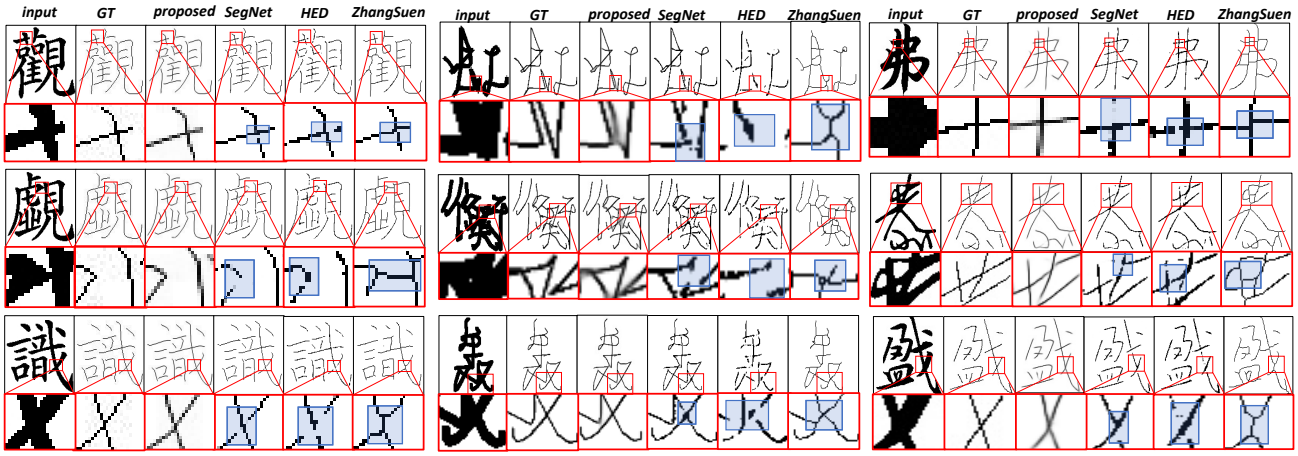

Fig. 9. Skeletons generated by three peer methods and the proposed approach for character images from the three experimental datasets—Kaiti9574 (left), HW (center), and SkeletonMF (right). For results reported for each dataset, columns from left to right show the input image, the groundtruth skeleton, and skeletons produced by the proposed approach, SegNet [21], HED [16] and Zhang-Suen [13] respectively.

three datasets. These results consistently reveal the superiority of the proposed model among all peer solutions. 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. Among the two peer deep learning-based skeletonization algorithms, HED and Seg perform comparatively over the skeletonMF and kaiti9574 datasets; yet in experiments executed over the HW dataset, Seg 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 using some standard font, the above experimental results suggest that Seg is more capable than HED in coping with handwritten characters or characters in versatile styles. In comparison with all 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 the table.

Table 5. The FID, OFM, OAHD and OHD scores of skeletonization results generated by three peer methods, ZS [13], HED [16], and Seg [21], as well as the proposed model, S.3D, in experiments conducted over the three datasets. A superior skeletonization result is indicated via a larger OFM score, a smaller FID, OAHD and OHD score.

kaiti9574

MODEL	FID	OFM	OAHD	OHD
ZS	74.8	0.427	1.37	10.45
HED	145.1	0.740	0.57	6.45
Seg	94.4	0.726	0.57	5.74
S3.D	**63.7**	**0.777**	**0.44**	**4.02**

HW

MODEL	FID	OFM	OAHD	OHD
ZS	114.5	0.452	1.65	9.67
HED	90.5	0.355	2.28	11.34
Seg	29.5	0.893	0.28	5.41
S3.D	**18.6**	**0.925**	**0.17**	**3.01**

skeletonMF

MODEL	FID	OFM	OAHD	OHD
ZS	190.2	0.313	2.13	12.82
HED	218.0	0.490	1.77	13.21
Seg	170.4	0.485	1.64	11.30
S3.D	**144.9**	**0.529**	**1.13**	**9.56**

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 groundtruth skeleton are shown

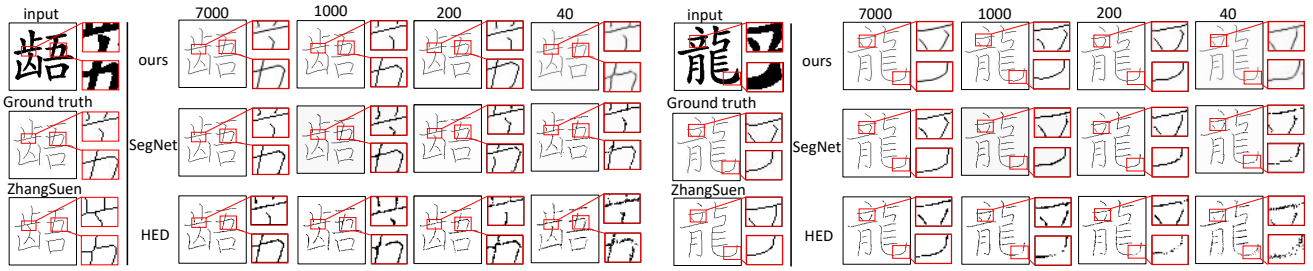

Fig. 10. Results generated by various methods using a progressively smaller set of training samples (7000, 1000, 200, 40) from the Kaiti9574 dataset.

in a zoomed-in view. It is easy to notice that skeletons extracted by the ZhangSuen algorithm 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, which also produces fragile skeletonization results for handwritten characters, an observation consistent with the finding obtained from the quantitative performance analysis discussed in the above. The Segnet algorithm performs most competently among all three existing solutions, which 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.

4.4 Skeletonizing Characters in Small Samples

Considering the label-intensive operations needed for acquiring groundtruth skeletons, a model's capability in 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 7000 to 1000, 200, and 40 respectively. The proposed model and all three peer methods are applied in this experiment. Fig. 10 shows results of this experiment. Except for the thinning algorithm, which does not depend on any training data, all 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 still remain at a high visual quality; in contrast, the quality of skeletons produced by the other two learning-based peer model declines significantly when the size of training samples drops.

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 carried out two additional experiments.

In the first experiment, real-world calligraphic images are used for evaluation purpose, 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 markedly from those printed in standard fonts, all of which makes their skeletonization operations much more challenging. It is also noted that calligraphic characters can be written in a vast number of personal styles, which provides a good test bed to explore an algorithm's capability in skeletonizing characters written in previously unwitnessed styles.

Fig.11. Skeletons extracted for real-world calligraphic characters (input) by the proposed approach (proposed) and the three peer methods with respect to the groundtruth (GT).

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 previously unencountered style. 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 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.

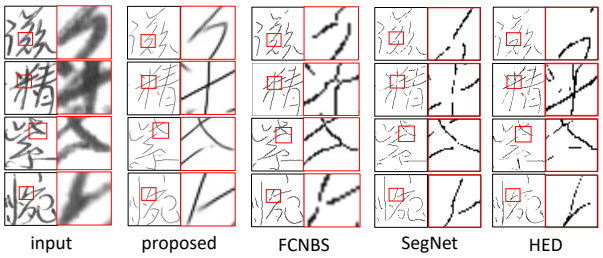

Fig.12. Skeletons extracted for character images in the CASIA-OFFHWDB1.1 dataset [24]. From left to right: input, results by the proposed model, FNCBS [1], SegNet [21] and HED [16].

Table 6. Mean of subjective opinions for skeletons generated by the proposed method in comparison with FCNBS [1], SegNet [21] and HED [16] on structure(structural correctness), conformance(conformance with input) and plausibility(overall plausibility).

	proposed	FNCBS	SegNet	HED
structure	3.8	2.7	1.3	1.6
conformance	3.7	2.2	1.5	1.8
plausibility	3.9	2.5	1.2	1.5

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 [24], introduced in [1]. Again, all three learning-based skeletonization models using the skeletonMF dataset due to the same empirical reason explained in 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 the proposed method delivers visually more plausible results. It is noted that no groundtruth skeletons are provided in the CASIA-OFFHWDB1.1 dataset. To quantitatively explore the relative performance among all peer solutions, we formulate a panel of ten 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, conformance to the input character image 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

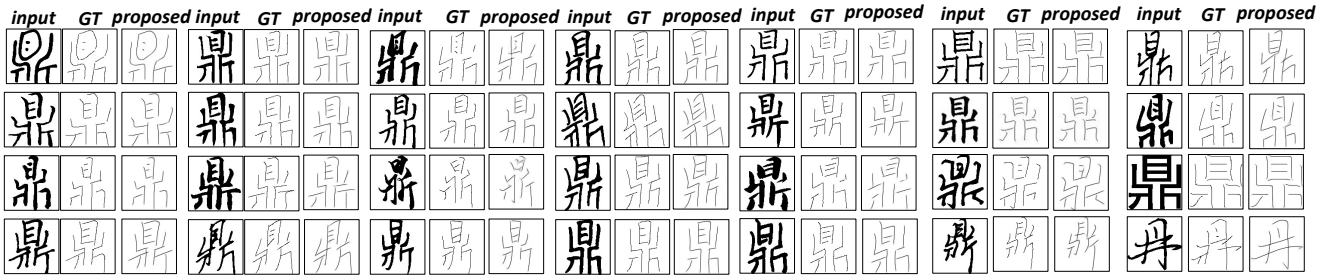

Fig.13. Skeletonization results generated by the proposed model regarding the Chinese characters 'ding' in 28 font styles. It is noteworthy that all these results are generated by a single trained network of the proposed model. The character is unseen by the model in its training.

1 (poor) to 5 (excellent). The evaluation results are shown in Table 6 , which convincingly demonstrates the superiority of the proposed method among all peer methods compared.

Finally, Fig. 13 shows skeletons extracted using the proposed approach for the Chinese character 'ding' in 28 writing styles, which comprehensively demonstrates the morphological diversity of Chinese characters as well as the proposed model's ability in coping 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 handwritten Chinese character task, a widely used Chinese character related task, for testing the generated skeleton of different methods. Specifically, we train ResNet-50 [42] by taking the ground truth skeletons in HW dataset as the input of the model and their corresponding character classification labels as the target output. After the training, we feed the skeletons obtained by using different methods to the ResNet-50 and calculate recognition accuracy. Since only the ground truth skeletons are used in the training, the higher accuracy indicates that the input skeletons are more similar to the ground truth in the view of recognition. And it partially illustrates the potential of the models to be applied in downstream tasks.

Table 7. The top1 and top5 recognition accuracy of skeletons with different sources when they are fed to a shared model. From left to right: GT (ground truth skeletons), proposed (skeletons extracted using proposed model), SegNet [21] and HED [16] and ZS [13] (skeletons extracted using peer method). The bigger accuracy indicates better performance.

	GT	proposed	SegNet	HED	ZS
Top-1	**96.6**%	94.6%	84.1%	36.4%	46.2%
Top-5	**99.8**%	99.8%	94.1%	51.8%	67.4%

Table 8. The top1 and top5 recognition accuracy of skeletons obtained by different versions of proposed model (see Sec 4.1.2). The bigger accuracy indicates better performance.

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

The top1 and top5 recognition accuracy reported in Table 7. indicate that skeletons obtained by proposed model are easily recognized by the model and achieve top-1 and top-5 accuracy relatively closed to the ground truth (94.6%, 99.8% compared to 96.6%, 99.8%). Although SegNet performs best among the three peer methods, its accuracy still has a considerable gap compared to our results. The remaining two methods can even not reach the top-1 recognition accuracy of 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 8. The gradually increasing top-1 and top-5 accuracy with the using of more components effec-

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

5 CONCLUSION

We propose 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, respectively empowered by three deep-learning modules, 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, trained using a newly introduced contextual loss with modification as a supplement for pixel-wise loss, the proposed model‘‘s skeletonization ability is further enhanced. As a whole, such a multistage processing pipeline is able to progressively improve the skeletonization result of a character, the effectiveness of which is comprehensively demonstrated by experimental results reported in this paper. Results of an ablation study additionally reveal the respective usefulness of the three constituent modules and newly introduced loss in the proposed generative model. Results of comparative experiments compellingly show that the proposed image-to-image translation framework is superior to multiple state-of-the-art peer algorithms in skeletonizing Chinese characters with significant advantages, which is evidenced by both qualitative perceptional inspections carried out by a panel of human evaluators and quantitative evaluation metrics widely adopted in the field. The new model is 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. Given the general difficulty of obtaining groundtruth skeletonization results of characters in an adequate number for machine learning purpose, the aforesaid small-sample learning and transfer learning capability of the proposed method present two of its distinguishing advantages with respect to all peer solutions known at present. The outstanding performance of our model in handwriting recognition task shows its promising potential and broad prospects for better completion of downstream tasks.

References

[1] Wang T Q, Liu C L. Fully Convolutional Network Based Skeletonization for Handwritten Chinese Characters. In *Proc. the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] L. Xu, Y. Wang, X. Li and M. Pan. Recognition of hand written chinese characters based on concept learning. *IEEE Access*, 2019,7:102039-102053.

[3] K. Yu, J. Wu, and Y. Zhuang. Skeleton-based recognition of chinese calligraphic character image. In *Pacific-Rim Conference on Multimedia*,Springer-Verlag 2008, pp.228-237.

[4] Y. Jiang, Z. Lian, Y. Tang, and J. Xiao. Dcfont: an endto-end deep chinese font generation system. In *SIGGRAPH Asia 2017 Technical Briefs*, 2017, pp.1-4.

[5] S. Azadi, M. Fisher, V. G. Kim, Z. Wang, E. Shechtman, and T. Darrell. Multi-content gan for few-shot font style transfer. In *CVPR*, 2018, pp.7564-7573.

[6] Y. Zhang, Y. Zhang, and W. Cai. Separating style and content for generalized style transfer. In *CVPR*, 2018, pp.8447-8455.

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

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

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

[10] Y. Jiang, Z. Lian, Y. Tang, and J. Xiao. GScfont: Structureguided chinese font generation via deep stacked networks. In *Proc. the 33th AAAI Conference on Artificial Intelligence*, 2019, pp.4015-4022.

[11] BBC - Languages - Real Chinese - Mini-guides - Chinese characters. Website, 2014. http://www.bbc.co.uk/languages/chinese/real_chinese/mini_guides/characters/characters_howmany.shtm.

[12] Chinese script styles - wikipedia. Website, 2008. https://en.wikipedia.org/wiki/Chinese_script_styles.

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

[14] A. K. Pujari, C. Mitra, and S. Mishra. A new parallel thinning algorithm with stroke correction for odia characters. *In Advanced Computing, Networking and Informatics-Volume1*,2014,pp:413-419.

[15] J. Dong, Y. Chen, Z. Yang, and B. W.-K. Ling. A parallel thinning algorithm based on stroke continuity detection. *Signal, Image and Video Processing*,2017, 11(5):873‘‘879.

[16] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015, pp.1395-1403.

[17] W. Ke, J. Chen, J. Jiao, G. Zhao, and Q. Ye. Srn: side-output residual network for object symmetry detection in the wild. In *CVPR*, 2017, pp.1068-1076.

[18] C. Liu, W. Ke, F. Qin, and Q. Ye. Linear span network for object skeleton detection. In *ECCV*, 2018, pp.133-148.

[19] Y. Wang, Y. Xu, S. Tsogkas, X. Bai, S. Dickinson, and K. Siddiqi. Deepflux for skeletons in the wild. In *CVPR*, 2019, pp.5287-5296.

[20] K. Zhao, W. Shen, S. Gao, D. Li, and M.-M.Cheng. Hi-fi: Hierarchical feature integration for skeleton detection. arXiv:1801.01849,2018 https://arxiv.org/abs/1801.01849.

[21] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*,2017, 39(12):2481-2495.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015, pp.3431-3440.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016, pp.770-778.

[24] C.-L. Liu, F. Yin, D.-H. Wang, and Q.-F. Wang. Casia online and offline chinese handwriting databases. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 2011, pp.37-41.

[25] Z. Lian, B. Zhao, X. Chen, and J. Xiao. Easyfont: a style learning-based system to easily build your large-scale handwriting fonts. *ACM Transactions on Graphics*,2018, 38(1):1-18.

[26] Make me a hanzi: Free, open-source chinese character data. Website, 2016. https://github.com/skishore/makemeahanzi.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017, pp.1125-1134.

[28] X. Huang and S. Belongie. Arbitrary style transfer in real time with adaptive instance normalization. In *CVPR*, 2017, pp.1501-1510.

[29] H. Tang, D. Xu, N. Sebe, Y. Wang, J. J. Corso, and Y. Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2019, pp.2417-2426.

[30] K. Regmi and A. Borji. Cross-view image synthesis using conditional gans. In *CVPR*, 2018, pp.3501-3510.

[31] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al.Hybrid task cascade for instance segmentation. In *CVPR*, 2019, pp.4974-4983.

[32] X. Liu, Y. Qiao, Y. Xiong, Z. Cai, and P. Liu. Cascade conditional generative adversarial nets for spatial-spectral hyperspectral sample generation. *Science China Information Sciences*,2020, 63(4):1-16.

[33] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers.Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In *CVPR*, 2016, pp.2497-2506.

[34] Z. Cui, H. Chang, S. Shan, B. Zhong, and X. Chen.Deep network cascade for image super-resolution. In *ECCV*, 2014, pp.49-64.

[35] B. Li, H. Chen, Y. Chen, Y. Dai, and M. He.Skeleton boxes: Solving skeleton based action detection with a single deep convolutional neural network. In *ICMEW*, 2017, pp.613-616.

[36] W. Xu, G. Parmar, and Z. Tu. Geometry-aware end-to-end skeleton detection. In *BMVC*, 2019.

[37] T.-Y. Lin, P. Doll'ar, R. Girshick, K. He, B. Hariharan, and S. Belongie.Feature pyramid networks for object detection. In *CVPR*, 2017, pp.2117-2125.

[38] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon.Cbam: Convolutional block attention module. In *ECCV*, 2018, pp.3-19.

[39] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor.The contextual loss for image transformation with non-aligned data. In *ECCV*, 2018, pp.768-783.

[40] 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*, 2017, pp.6626-6637.

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

[42] He, K., Zhang, X., Ren, S., Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, October 2016, pp.630-645.

Journal of computer science and technology: Instruction for authors. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY

Detailed Extensions

We will elaborate in this document the extensions in this version compared to the previous conference version [1] accepted by CVM 2021 (The 9th international conference on Computational Visual Media) and not yet published. Our main new contribution is introducing the feature-based contextual loss [1] with modification to enhance the performance of the proposed model on skeleton extraction task. Experimental results verify its effectiveness under the measure of newly added FID (Frechet Inception Distance) [2], which is considered close to the human perception and is commonly-used in evaluation of image generation model. We also add experiments on handwritten Chinese character recognition task to explore the possible impact of skeleton quality on downstream tasks by comparing the recognition accuracy of obtained skeletons of different models.

We report the detail of contextual loss in Sec.3.4 of manuscript [2]. The basic purpose of introducing the contextual loss is to encourage the model to pay more attention to errors that are easily perceivable by humans. For example , we hope model spend less attention on alignment errors (e.g. a small displacement or rotation), which seriously affect the value of pixel-based loss but bring no change on skeleton topology. The careful design of feature-based contextual loss help model to achieve better performance on FID. It is worth noting that when acquiring features to calculate the above loss, we innovatively employ an autoencoder trained on skeleton data instead of VGG19 [3] used in original work [1]. This design takes into account the significant differences between real world images and skeleton images, and brings no additional difficulties in data acquisition and model training. Meanwhile, the autoencoder is dedicated to reconstruct the input skeleton, thus the acquired features tend to carry all skeleton information, which also ensures that the comparison between images is comprehensive enough.

We also adopt the Frechet Inception Distance (FID) [2] commonly-used in image generation tasks as a new evaluation metric. FID is employed to measure the distance between generated images and ground truth images in feature level, and is consequently closer to human perception than pixel-wise metrics. Therefore, FID is suitable for evaluating the effect of contextual loss.

Experiments (Sec 4.2.2, Table 3. and Table 4. in manuscript) are conducted to verify the effectiveness of contextual loss. Results reported in Table 3. indicate that the skeletonization performance of proposed model on FID are effectively improved by introducing the contextual loss with autoencoder trained on skeleton data. And another version of model using VGG19 as original work [1] results a decrease in performance, which prove the necessity and effectiveness of our modification to feature acquisition approach. We also explored the impact of dataset size on the performance of contextual loss by gradually reducing the number of training samples. The expanded experiment results shown in Table 4 reveal the special advantage of contextual loss on small-scale dataset.

And extended experiments are conducted to illustrate the effect of skeleton quality on a downstream

[1] see in followed conference-version.pdf

[2] see manuscript.pdf

task, handwritten Chinese character recognition, which is widely used is everyday life (see Sec 4.6, Table 7. and Table 8. in manuscript). A handwriting recognition model is trained using ground truth skeleton, and tested on skeletons from different sources. Since only the ground truth skeletons are used in the training, the higher recognition accuracy indicates that the input skeletons are more similar to the ground truth in the view of recognition. And it partially illustrates the potential of the models to be applied in downstream tasks. Results of comparison among peer methods and among different versions of proposed model are displayed in Table 7 and Table 8. The top-1 and top-5 recognition accuracy reported in Table 7 indicate that skeletons obtained by proposed model are easily recognized by the model and achieve accuracy relatively closed to the ground truth. And the gradually increasing top-1 and top-5 accuracy in Table 8. with the using of more components clearly illustrates the effectiveness of modules in proposed model.

References

[1] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor.The contextual loss for image transformation with non-aligned data. In *ECCV*, 2018, pp.768-783.

[2] 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*, 2017, pp.6626-6637.

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