Multi-label semantic segmentation algorithm for stroke extraction in calligraphy teaching

Jiayun Yu[a], Dingyu Li*[b], Zhanyang Xu[b], Jinghong Wang[b], Wei Lin[c]
[a]School of Education Science, Nanjing Normal University, Nanjing, Jiangsu, China; [b]School of Software, Nanjing University of Information Science & Technology, Nanjing, Jiangsu, China, [c]Nanjing Technology R&D Center of Jiangsu Children's Spring Interconnection Education Technology Co., Ltd, Nanjing, Jiangsu, China

ABSTRACT

Chinese characters, as the fundamental medium of communication within Chinese culture, stand out due to their intricate structures. Strokes, the basic elements of Chinese characters, are crucial for assessing Chinese handwriting. Accurate stroke extraction is essential and serves as the initial step in this evaluation. Traditional stroke extraction methods typically rely on specific rules that often fail to capture the full complexity of Chinese characters and cannot align the strokes according to the sequence used in template characters during assessments. To address these challenges, this paper redefines stroke extraction as a multi-label semantic segmentation task and introduces a new model, M-TransUnet. This model utilizes a deep convolutional approach to train individual Chinese characters, maintaining the integrity of stroke structures and resolving ambiguities in stroke segment combinations. It also accurately determines the order of strokes, aiding in subsequent tasks such as stroke evaluation. Furthermore, since handwriting images are only segmented into foreground and background without additional color cues, they are prone to false positive (FP) segmentation noise. To mitigate this issue, we propose a Local Smooth Strategy on Strokes (LSSS) that diminishes noise impacts on the segmentation results.

Keywords: Chinese handwriting characters, stroke extraction, multi-label semantic segmentation, local smooth strategy

1. INTRODUCTION

Extracting strokes from hard pen Chinese characters has always been a challenging task due to the intricate structure of the characters and the numerous types of strokes that often intersect. Traditional methods typically involve determining the stroke segments near intersection areas through rule-based assignments and then segmenting the strokes into two main categories.

One approach refines the character to a skeleton, from which strokes are extracted based on pixel arrangement rules[1-3], solving issues at intersection points[4] and reassembling stroke segments by specific rules[5]. Lin et al.[6] disassembled the skeleton into stroke segments, calculating the incident direction at all intersection points and using a bi-directional graph to determine the connections needed between stroke segment pairs. Li et al.[7] simplified the basic strokes into four types: horizontal, vertical, left-falling, and right-falling strokes, classifying stroke types based on the angle between the skeleton and the horizontal line. During the processing of specific types of strokes, other types are considered noise. Liu et al.[8] utilized the CPD algorithm to align the target and template characters by their intersection and turning points, determining stroke segment affiliation and introducing segment weights to handle complex character structures.

These skeleton-based methods heavily depend on the quality of the skeleton, issues such as splitting at intersection points, and are incapable of completely avoiding errors introduced by algorithmic rules, losing crucial information like brush tips and stops. Moreover, key point alignment methods cannot ensure an equal count of points in the datasets being compared.

Another category uses the existing features of the hard pen character images for extraction. Cao et al.[9] employed the PBOD algorithm to enhance the continuity properties of strokes, improving the separation of overlapping strokes while maintaining connectivity. Shi et al.[10] derived boundary features from the first derivative of the peripheral contours of the stroke segments, dynamically adjusting the search direction to more accurately match the actual contours of the stroke segments. Li et al.[11] used horizontal and vertical differential methods to detect the contours of Chinese characters, creating

*1063765113@qq.com

International Conference on Optics, Electronics, and Communication Engineering (OECE 2024),
edited by Yang Yue, Proc. of SPIE Vol. 13395, 133953R · © 2024 SPIE · 0277-786X · Published
under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3048589

closed polygonal contours and extracting concave and convex points based on the position, angle, and distance of the concave points to pair and connect intersecting strokes. Zhang et al.[12] incorporated stroke contours into the segmentation of Chinese characters based on the skeleton, precisely locating each stroke's contour range and assigning them to the corresponding contour strokes according to the features of the intersection area.

Compared to traditional methods, these feature-based approaches utilize more information and have higher error tolerance but still suffer from the constraints of manually set rules and cannot cover all scenarios of manual hard pen writing.

Recently, neural network-based stroke extraction algorithms have emerged. Liu et al.[13] transformed stroke extraction into an instance segmentation problem, utilizing a two-stage strategy[14] and training on a comprehensive dataset of all Chinese characters to enhance efficiency and robustness. However, the nature of instance segmentation only correctly categorizes the strokes without identifying which specific stroke of the hard pen character it corresponds to, thus affecting the comparative evaluation with the template strokes. Zhang et al.[15] utilized conditional generative adversarial networks[16] to train each stroke of each character as a separate task, selecting specific strokes to generate. Due to the multitude of Chinese strokes, this method requires significant training tasks and, when extracting strokes from hard pen characters, multiple model weight loadings, resulting in lengthy extraction times.

These methods, by learning the characteristics of sample strokes, better adapt to the morphologically complex and varied nature of manual hard pen writing. Compared to conventional approaches, they offer improved robustness.

Consequently, this paper employs neural networks for stroke extraction. Semantic segmentation, as a classical branch of neural network tasks, has attracted extensive attention from researchers. The representative network, Unet[17], widely applied and studied, employs a fully convolutional encode-decode structure with skip connections, ensuring the correct representation of high-level and low-level semantics while being structurally efficient. Based on this, researchers have incorporated multiple improvements, such as TransUnet[18], which adds a self-attention[19] mechanism to Unet, capturing distant feature dependencies while maintaining high-resolution features.

However, TransUnet only addresses single-label segmentation issues and is ineffective in handling intersecting strokes. Multi-label semantic segmentation is most suited for extracting strokes from hard pen characters, and since these images are typically binary, lacking additional color information, they rely solely on spatial structure information, making them more prone to FP noise in segmentation, impacting subsequent refinement and stroke evaluation tasks.

To address these issues, this paper uses the multi-label semantic segmentation model M-TransUnet, training on Chinese characters with strokes as categories. Additionally, a Local Smooth Segmentation Strategy (LSSS) is proposed to eliminate segmentation noise while preventing the overall image from becoming blurred, thus ensuring the clarity of stroke edges.

2. METHOD

Due to the particularities of hard pen Chinese characters, intersections between two or even multiple different strokes can occur, such as the mid-segment intersection (a) and vertex intersection (b) shown in Figure 1. Vertex intersections, which cause strokes to connect, allow the intersection area to be assigned to one of the strokes without affecting the continuity of the other strokes. However, for mid-segment intersections, if the intersection area is assigned to one stroke, it results in discontinuities in the other strokes, thus disrupting the inherent continuity of the strokes.

To address this issue, one could conceptualize each stroke of every Chinese character as a separate unit task, with each task performing a binary classification of foreground versus background. This approach, however, would exponentially increase the training costs and necessitate loading model weights for each stroke during extraction. Therefore, this paper utilizes multi-label semantic segmentation for stroke extraction. Although multi-label semantic segmentation, compared to binary single-label semantic segmentation, involves learning many more features and the hard pen character images inherently lack color and grayscale information—relying solely on spatial structural information—this can lead to increased segmentation noise.

To mitigate this issue, this paper introduces a Local Smooth Segmentation Strategy (LSSS), specifically targeting stroke extraction. This strategy effectively reduces segmentation noise, thereby enhancing the clarity and continuity of the strokes while maintaining the integrity of the overall image. By integrating multiple labels and applying LSSS, the model better adapts to the complexities of segmenting hard pen Chinese characters, making it a robust solution for high-precision calligraphy digitization and analysis.

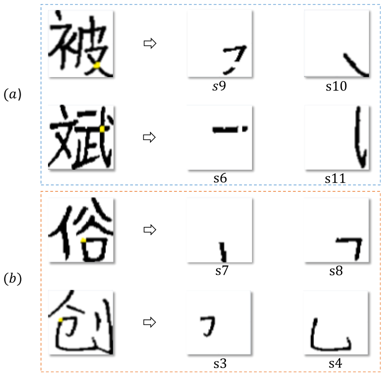

Figure 1. Stroke intersection diagram.

2.1 M-TransUnet

This paper references the TransUnet network[18], as illustrated in Figure 2, which incorporates a Transformer module into the Unet[17] architecture. Typically, the Transformer treats the input feature map as a one-dimensional sequence and primarily focuses on capturing global information, often neglecting local high-resolution features. In contrast, Unet is constrained by the depth of the network and the inherent characteristics of convolutional networks, which do not adequately address global contextual details. TransUnet synergizes the strengths of both Transformer and Unet architectures. By utilizing the self-attention mechanism, it captures global context information, while the skip-connection design of Unet preserves sufficient high-resolution features. In the task of extracting strokes from hard pen Chinese characters, where the input images are binary, distinguishing only between the foreground of the hard pen characters and the background without any color or grayscale variations, stroke extraction relies solely on the shape structure of the strokes themselves and the overall structure of the hard pen characters. The features of TransUnet are particularly well-suited for the extraction of hard pen character strokes. In tasks of single-label classification or semantic segmentation, to achieve a category with the maximum score, it is generally preferable to have a significant scoring difference among the categories. The Softmax activation function is defined as equation (1).

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{C} e^{z_c}} \tag{1}$$

Due to the exponential nature of the function e^{z_i} where $z_i > 0$ e^{z_i} is considerably large. Even minor differences between the z_i values can result in substantial differences in activation values, where larger values may suppress smaller ones. This characteristic makes the Softmax activation function particularly suitable for tasks involving single-label classification and semantic segmentation, where a distinct classification or segmentation is typically desired.

In the context of multi-label semantic segmentation tasks, each pixel may simultaneously belong to multiple categories. In such scenarios, it is undesirable for higher scoring categories to suppress those with slightly lower scores, as it is crucial for the scores of different categories to be independent of each other to determine one or more categories exceeding a given threshold. Therefore, the Sigmoid activation function, which is frequently used in binary classification tasks, becomes applicable. The Sigmoid function is represented as equation (2) and is favored in these tasks because it treats the probability of each class independently, allowing for multiple classes to be identified per pixel. This independent treatment of class scores facilitates the identification of multiple relevant categories at each pixel, enhancing the performance of multi-label semantic segmentation.

$$\sigma(z_i) = \frac{1}{1+e^{-z_i}} \tag{2}$$

Neural network outputs, which are real numbers, can be mapped to the interval $[0,1]$ using the sigmoid activation function $\sigma(z_i)$. If $\sigma(z_i) > 0.5$, the output is classified as belonging to the category represented by 1; otherwise, it is classified as belonging to the category represented by 0. In the context of multi-label multi-class tasks, this scenario can be viewed as a collection of C binary classification tasks. Therefore, in comparison to the Softmax function, the Sigmoid function is more appropriate for multi-label semantic segmentation tasks.

The Sigmoid function independently processes each class output, making it ideal for scenarios where each pixel might be associated with multiple classes. This independence allows the function to effectively handle multiple labels per instance, which is a typical requirement in multi-label segmentation. Unlike Softmax, which normalizes the output across all classes forcing them to sum to one and thus inherently making class outputs interdependent, Sigmoid treats each output separately. This characteristic is crucial in multi-label tasks where the presence of one label does not necessarily preclude the presence of another, allowing for a more nuanced and precise segmentation output.

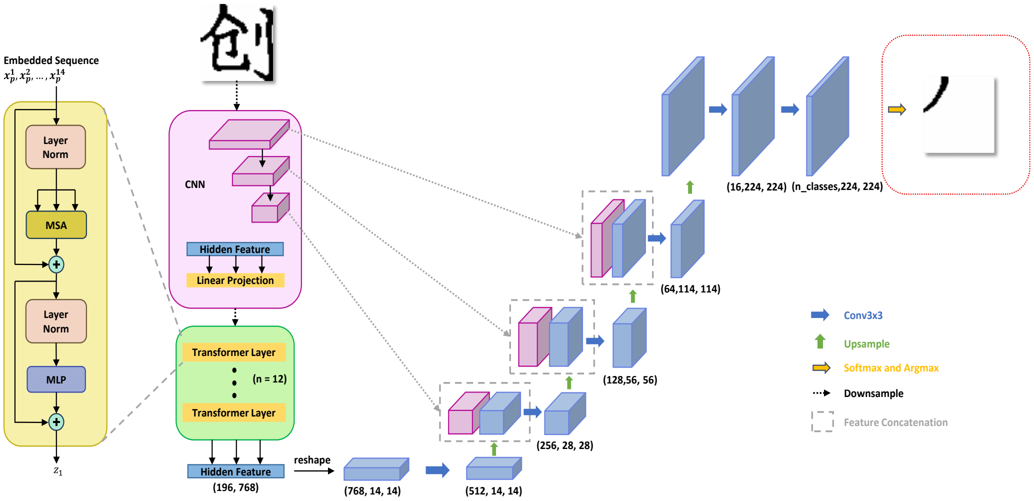

Figure 2. TransUnet network framework.

This paper introduces the Multi-Label TransUnet (M-TransUnet), which modifies the activation function of the output layer of the standard TransUnet from Softmax to Sigmoid. This modification allows for the independent calculation of scores for different categories. A threshold of 0.5 is used to determine whether a pixel belongs to the current category, as illustrated in Figure 3. Given that the background of hard pen character images is represented by white pixels and the foreground by black pixels, a score below 0.5 indicates that the pixel belongs to the current stroke category, while a score above 0.5 suggests it does not.

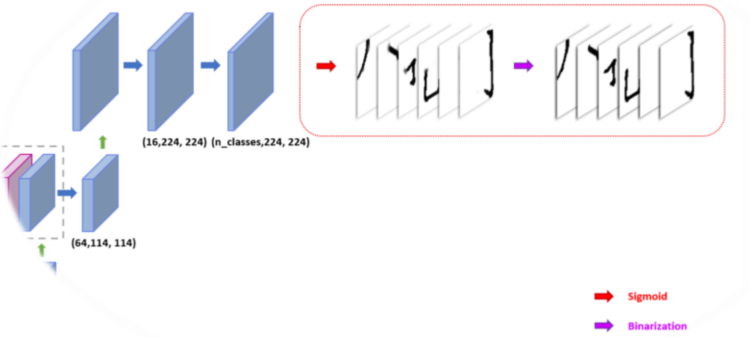

Figure 3. M-TransUnet output layer.

In multi-label tasks, which are equivalent to multiple binary classification tasks, it is appropriate to use the average binary cross-entropy function as the loss function. This function is defined as follows, where N represents the number of pixels, y_j is the true value of a single pixel, and $\hat{y}_j$ is the predicted value for that pixel:

$$L_{BCE} = \frac{\sum_{j=1}^{N} -y_j * \log(\hat{y}_j) - (1-y_j) * \log(1-\hat{y}_j)}{N} \tag{3}$$

While the binary cross-entropy function effectively represents the discrepancy between predicted values and true values at the pixel level, it does not accurately reflect the similarity between predicted masks and true masks in semantic segmentation tasks. The Dice Loss, expressed as equation (4), measures the similarity between the predicted mask and the true mask but is less sensitive to differences in individual pixels:

$$L_{DiceLoss} = 1 - \frac{2\sum_{j=1}^{N} y_j \hat{y}_j}{\sum_{j=1}^{N} y_j + \sum_{j=1}^{N} \hat{y}_j} \tag{4}$$

Therefore, this study employs a hybrid loss function, combining both binary cross-entropy and the Dice Loss, to optimize the segmentation performance effectively while focusing on the significant aspects of hard pen Chinese character stroke segmentation.

$$L_{Mix} = \frac{L_{BCE} + L_{DiceLoss}}{2} \tag{5}$$

2.2 Local smooth strategy on stroke

Although the strokes of hard pen characters follow a sequential order, there is no occurrence of strokes being occluded; hence, the segmentation results should manifest as continuous blocks of black pixels. However, the hard pen character images utilized in this study are binary, lacking additional color information and relying solely on spatial structural information. This characteristic renders them more susceptible to FP segmentation noise compared to other segmentation tasks, as illustrated in Figure 4. Such noise could adversely affect downstream tasks of stroke refinement and evaluation.

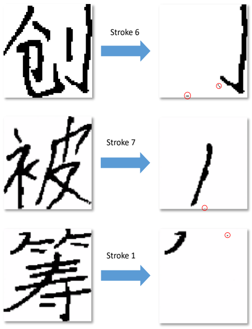

Figure 4. Segmentation noise diagram.

This paper proposes a Local Smooth Strategy on Stroke (LSSS) specifically targeting the foreground areas of hard pen character images, utilizing the inherent binary characteristics of these images. This method effectively discerns the foreground areas without the need for additional algorithms, and the segmented strokes remain as parts of the hard pen character foreground. We set a window size of $H \times H$, where H is an odd number, and categorize the pixels within the window into $\frac{H+1}{2}$ levels, treating each level as a unit. To prevent blurring at the edges of the strokes, we only compute the average of valid pixels that are within the foreground area of the hard pen characters at the level n.

$$A_n = \frac{S_n}{K_n} \tag{6}$$

Here, K_n represents the count of valid pixels, and S_n is the sum of the grayscale values of these valid pixels. The weights for different levels decrease from the innermost to the outermost level, with α as the decay coefficient. The weight for each level is given by equation (8):

$$W_n = \frac{\alpha^{n-1}}{\sum_{a=1}^{\frac{H+1}{2}} \alpha^{a-1}} \tag{7}$$

This formula results in the estimated value for the central pixel:

$$p = \sum_{n=1}^{\frac{H+1}{2}} A_n * W_n \tag{8}$$

If a level lacks valid pixels within the foreground area, its weight W^* is redistributed equally among the remaining levels:

$$p^* = p * (1 + W^*) \tag{9}$$

The algorithmic process is illustrated in Figure 5. By excluding the influence of the image background, this strategy not only eliminates small-area FP missegmentations but also ensures the clarity of the stroke edges. This approach significantly refines the precision of the segmentation output, particularly in maintaining high-quality edges which are critical for further processing and analysis of hard pen characters.

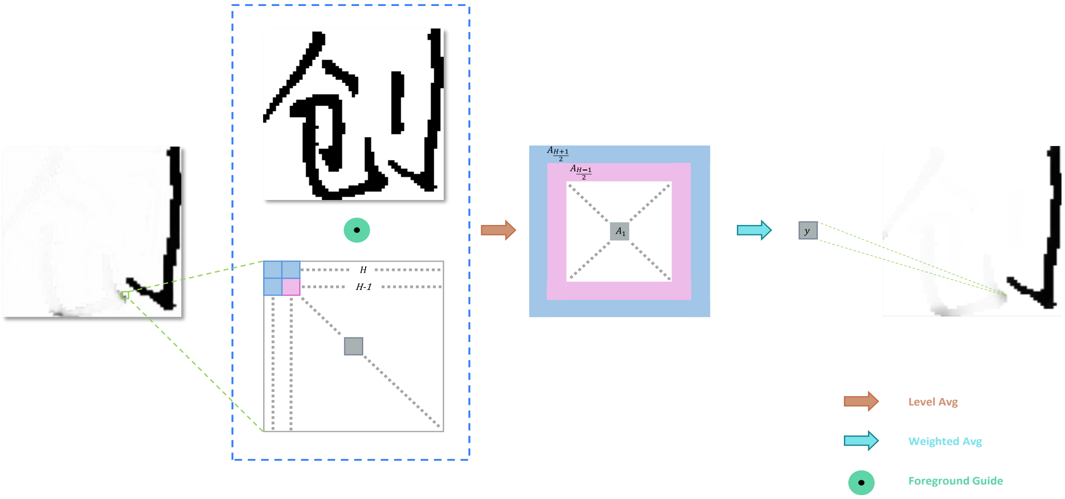

Figure 5. Algorithm flowchart.

3. EXPERIMENT

3.1 Experimental data

The dataset utilized in this study comprises images of regular script hard pen characters written by elementary and middle school students. This research aims to evaluate the impact of the number of strokes and the number of intersections per stroke on the performance of multi-label semantic segmentation. Three characters were selected for detailed examination: ‘创’, ‘被’, and ‘筹’. For each character, 240 images were utilized. All images were manually annotated to ensure high accuracy and consistency of the data.

3.2 Results

In this paper, experimental results are evaluated using the F1 Score and Accuracy metrics, which provide a robust measure of segmentation performance, especially under conditions of class imbalance between positive and negative samples. The formulas for calculating Accuracy (AC) and F1 Score are given by equations (11) and (12) respectively:

$$AC = \frac{TP+TN}{TP+TN+FP+FN} \tag{10}$$

$$F1 = \frac{2TP}{2TP+FP+FN} \tag{11}$$

Tables 1-3 present the F1 Scores and Accuracy for the segmentation of the characters ‘创’, ‘被’, and ‘筹’ using the TransUnet and M-TransUnet. These tables detail the performance for individual strokes, the average across all strokes, and the overall performance using the M-TransUnet, respectively. For the TransUnet, the mean F1 Scores achieved for the three characters were 0.9770, 0.9655, and 0.9651, indicating that the segmentation quality is comparable to manual stroke extraction in well-written hard pen characters.

The M-TransUnet, which is required to learn more complex stroke features, shows a slight decline in segmentation performance. However, the mean F1 Scores for ‘创’, ‘被’, and ‘筹’ are still substantial, at 0.9677, 0.9512, and 0.9513 respectively. These results underscore the efficacy of the M-TransUnet in handling more demanding multi-label segmentation tasks, while maintaining high levels of accuracy and precision in stroke segmentation.

Table 1. Stroke extraction performance for ‘创’.

	AC	F1
TransUnet-Avg	0.9988	0.9770
M-TransUnet-Avg	0.9977	0.9677

Table 2. Stroke extraction performance for ‘被’.

	AC	F1
TransUnet-Avg	0.9984	0.9655
M-TransUnet-Avg	0.9967	0.9512

Table 3. Stroke extraction performance for ‘筹’.

	AC	F1
TransUnet-Avg	0.9985	0.9651
M-TransUnet-Avg	0.9975	0.9513

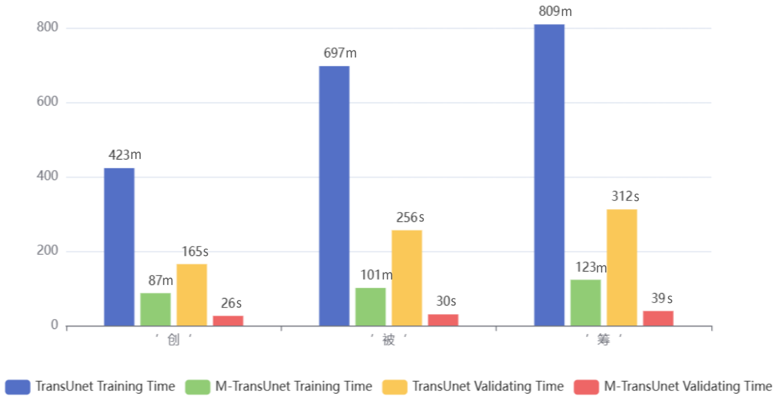

Figure 6. Training and validating time comparison chart.

Additionally, this study documented the efficiency comparison between single-label and multi-label semantic segmentation. In single-label semantic segmentation tasks, each stroke requires training a separate model, which not only consumes substantial training time but also increases time costs during prediction due to the need to load multiple models. In contrast, multi-label semantic segmentation allows for simultaneous processing of multiple strokes with just a single training session, significantly saving training time and reducing model loading times during prediction.

This experiment compared the time consumed for training and predicting the character ‘创’ using TransUnet, with results presented in Figure 6. Here, training times are measured in minutes (m), and prediction times in seconds (s), on a platform equipped with an RTX2060 GPU.

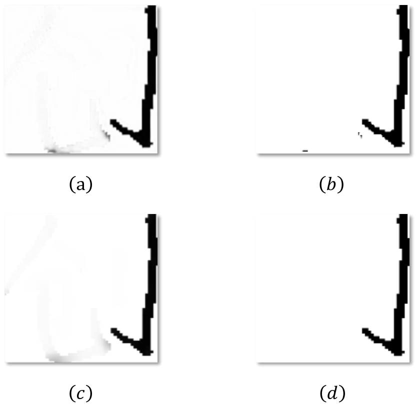

Figure 7. LSSS effect diagram.

Figure 7 illustrates typical false positive (FP) missegmentations from the test set, specifically the fourth stroke of the character ‘创’. For small-area segmentation noise (Figure 7a), the pixels themselves are closer to a grayscale value of 0, while the surrounding pixels predominantly approximate a grayscale value of 1. After conversion to hard labels, as shown in Figure 7b, the result is predominantly background. After processing with the Local Smooth Segmentation Strategy (LSSS), the pixel grayscale values within the foreground area of the hard pen characters become smoother, and the gradient at the foreground-background interface remains unaffected. The missegmented pixels, which are close to a grayscale value of 0, are influenced by the surrounding pixels which are close to a grayscale value of 1, causing their grayscale values to also approach 1 (Figure 7c). After conversion to hard labels, as depicted in Figure 7d, these adjustments produce a more accurate representation of the intended strokes, significantly reducing FP errors and enhancing the overall segmentation quality. Figure 8 demonstrates the final results of stroke extraction achieved by the algorithms discussed in this paper.

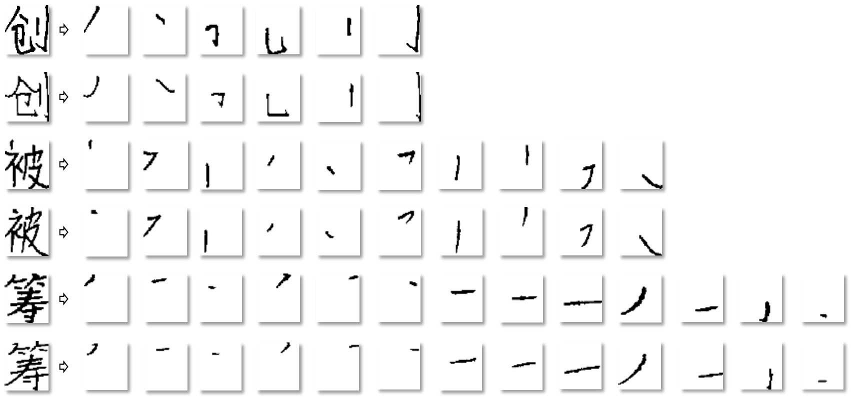

Figure 8. Test set stroke extraction results diagram.

This study has compiled statistics on the proportion of strokes with an F1 score exceeding 0.88 in the validation set for the characters ‘创’, ‘被’, and ‘筹’ using the M-TransUnet and the method by Zhang et al.[15] (CGAN), to serve as a measure of correct extraction rates. Additionally, the correct extraction rates based on the dataset used in this paper for the method by Xu et al.[20] (Traditional) are presented in Table 4.

Table 4. Stroke extraction accuracy.

Algorithm	**创**	**被**	**筹**
Traditional	83.19%	71.33%	61.09%
CGAN	98.92%	97.31%	96.31%
M-TransUnet	97.84%	96.53%	94.95%

Compared to traditional methods, the approach presented in this paper significantly improves accuracy. While there is a slight decrease in accuracy compared to the method by Zhang et al., the benefits of multi-label semantic segmentation, particularly in terms of reduced training and prediction times, are substantial. Liu et al.[13] segment strokes using instance segmentation, which does not capture the order of strokes in hard pen characters and cannot prevent issues of over-detection and under-detection. Consequently, it is not possible to match and compare the segmented strokes with template strokes in subsequent processes.

In contrast, the method developed in this study utilizes categorical dimension information in labels during training to establish the order of strokes in hard pen characters. As a result, the output strokes can be directly compared and evaluated against the corresponding template strokes, providing a clear advantage in terms of usability and functional performance.

4. CONCLUSION

In this paper, we have transformed the task of extracting strokes from hard pen characters into a problem of semantic segmentation. By utilizing a morphologically diverse dataset, we enabled the neural network to learn deep features of the strokes, enhancing both robustness and accuracy. The adoption of the multi-label semantic segmentation network, M-TransUnet, in place of a single-label network significantly reduces training time. Additionally, the sequence of labels facilitates the determination of stroke order, which is crucial for downstream stroke evaluation tasks.

To address segmentation noise, we introduced the Local Smooth Segmentation Strategy (LSSS), specifically targeted at strokes. This strategy not only eliminates segmentation noise but also prevents blurring at the edges of the strokes, thereby obviating the need for further post-processing in downstream refinement and evaluation tasks.

Experimental results validate the feasibility of our approach, underscoring its significant implications for the intelligent evaluation of hard pen characters. This method ensures efficient and accurate stroke extraction, which is essential for automated systems involved in the educational and technological assessment of handwriting quality.

REFERENCES

[1] Xu, Z. Y., Liang Y., Zhang, Q., Le, D. and Ebroul, I., "Decomposition and matching: Towards efficient automatic Chinese character stroke extraction," In 2016 Visual Communications and Image Processing (VCIP), 1-4 (2016).

[2] Fan, Y. F., Li, C. C., et al., "Extraction of handwritten Chinese character strokes based on local information," Journal of Inner Mongolia Normal University: Natural Science Chinese Edition, 52(2), 181-188 (2023).

[3] Xun, E. D., Lv, X. C., An, W. H. and Sun, Y. N., "Stroke restoration in handwritten Chinese character images for writing instruction (doctoral dissertation)," Journal of Peking University (Natural Science Edition), (2015).

[4] Liao, C. W. and Huang, J. S., "Stroke segmentation by bernstein-bezier curve fitting," Pattern Recognition, 23(5), 475-484 (1990).

[5] Fan, K. C. and Wu, W. H., "A run-length-coding-based approach to stroke extraction of Chinese characters," Pattern Recognition, 33(11), 1881-1895 (2000).

[6] Lin, F. and Tang, X., "Off-line handwritten Chinese character stroke extraction," 2002 International Conference on Pattern Recognition, 3, 249-252 (2002).

[7] Li, J. H., Wang, H., Yan, W. Z., et al., "A new method for Chinese character thinning and stroke extraction," Proceedings of the Sixth National Conference on Information Acquisition and Processing, Chinese Society for Instrument and Control (Editorial Office of Chinese Journal of Scientific Instrument), 1, 230-233 (2008).

[8] Liu, X. C., Li, Z. F., Jiang, J. and Li, Y., "Study on the extraction of regular script strokes based on the CPD algorithm and pen segment weight," Computer Applications and Software, 39(8), 204-212 (2022).

[9] Cao, R. and Tan, C. L., "A model of stroke extraction from Chinese character images," Proceedings 15th International Conference on Pattern Recognition, 4, 368-371 (2000).

[10] Shi, W., Fu, Y., Chen, A. L., et al., "A dynamic method for Chinese character stroke segment extraction," Computer Applications Research, 25(7), 1998-2000 (2008).

[11] Li, C., Wang, J. Q., and Li, B., "Algorithm on strokes separation for Chinese characters based on edge," Computer Science, 40(7), 307-311 (2013).

[12] Zhang, X. F. and Liu, J. Y., "Extraction of Calligraphy Strokes Using Crawler Method," Journal of Computer-Aided Design and Computer Graphics, 2016 (2), 301-309 (2016).

[13] Liu, L., Lin, K., Huang, S., et al., "Instance segmentation for Chinese character stroke extraction, datasets, and benchmarks," arXiv preprint arXiv:2210.13826, (2022).

[14] He, K., Gkioxari, G., Dollár, P., et al., "Mask R-CNN," Proceedings of the IEEE International Conference on Computer Vision, 2961-2969 (2017).

[15] Zhang, W., Zhang, X. and Wan, Y. J., "Stroke segmentation of calligraphy characters based on conditional generative adversarial networks," Acta Automatica Sinica, 48(7), 1861-1868 (2022).

[16] Isola, P., Zhu, J. Y., Zhou, T., et al., "Image-to-image translation with conditional adversarial networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1125-1134 (2017).

[17] Ronneberger, O., Fischer, P. and Brox, T., "U-Net: Convolutional networks for biomedical image segmentation," Medical Image Computing and Computer-Assisted Intervention: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III, 18, Springer International Publishing, 234-241 (2015).

[18] Chen, J., Lu, Y., Yu, Q., et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306 (2021).

[19] Vaswani, A., Shazeer, N., Parmar, N., et al., "Attention is all you need," Advances in Neural Information Processing Systems, 30 (2017).

[20] Xu, Z., Liang, Y., Zhang, Q., et al., "Decomposition and matching: Towards efficient automatic Chinese character stroke extraction," 2016 Visual Communications and Image Processing (VCIP), 1-4 (2016).