Two-Stage Deep Learning Framework for Automatic Evaluation of Hard-Pen Chinese Calligraphy

Wei Shi
School of Science
Xi'an Technological University
Xi'an, China
E-mail: 2902876593@qq.com

Yu Lei
School of Chinese Calligraphy
Xi'an Technological University
Xi'an, China
E-mail: leiyu@xatu.edu.cn

Yuran Zhang
School of Science
Xi'an Technological University
Xi'an, China
E-mail: 2784068234@qq.com

Yujie Zhao
School of Chinese Calligraphy
Xi'an Technological University
Xi'an, China
E-mail: 3521307633@qq.com

Yuankui Ma
School of Science
Xi'an Technological University
Xi'an, China
E-mail: mayuankui@xatu.edu.cn

***Abstract*—Intelligent evaluation of hard-pen Chinese calligraphy is crucial for smart education, online learning, and automated grading systems. Traditional manual assessment is inefficient, subjective, and difficult to scale. A novel two-stage deep learning framework for automatic handwriting quality assessment is proposed in this study. In the first stage, a convolutional AutoEncoder is employed to generate pseudo-labels based on reconstruction error, enabling unsupervised evaluation without manual annotation. In the second stage, supervised regression networks, including ResNet18 and MobileNetV2, are trained to predict handwriting scores from images, capturing complex structural features. It is demonstrated through extensive experiments on the CASIA-HWDB dataset that stable and reasonable score distributions are achieved by the proposed method, with predicted scores predominantly concentrated in the 650–700 range. Furthermore, comparative analysis between ResNet18 and MobileNetV2 illustrates the trade-off between model accuracy and computational efficiency, highlighting the feasibility of deploying lightweight models in resource-constrained scenarios. Overall, an effective and scalable solution for automated calligraphy evaluation is provided by the proposed framework.**

Keywords-Deep Learning; Autoencoder; Resnet18; Mobilenetv2; Chinese Calligraphy Evaluation; Pseudo-Label; Two-Stage Assessment

I. INTRODUCTION

The evaluation of handwritten Chinese characters plays an important role in education, handwriting training, and intelligent learning systems. As one of the most important carriers of Chinese culture, Chinese calligraphy not only reflects writing skills but also demonstrates the writer's understanding of character structure, stroke arrangement, and aesthetic expression. With the rapid development of digital education and intelligent tutoring systems, automatic handwriting evaluation has attracted increasing attention from both academic researchers and educational practitioners. A recent comprehensive survey reviewed traditional and machine-learning-based handwritten Chinese character evaluation methods and highlighted the growing demand for intelligent assessment systems [1].

Traditional handwriting assessment mainly relies on expert evaluation and manual scoring. Although experienced teachers can provide comprehensive feedback on writing quality, such evaluation methods are often time-consuming, labor-intensive, and highly subjective. Furthermore, manual assessment becomes increasingly impractical when dealing with large-scale handwriting datasets. These limitations have motivated researchers to explore automated handwriting evaluation methods capable of providing objective and efficient assessment results.

Recent advances in artificial intelligence and deep learning have significantly improved the performance of image analysis and pattern recognition systems. Convolutional Neural Networks (CNNs) have achieved remarkable success in various computer vision tasks, including image classification, object detection, and handwriting recognition. Inspired by these achievements, several studies have attempted to apply deep learning techniques to handwriting quality evaluation. For example, deep learning-based penmanship assessment systems have demonstrated promising performance in predicting aesthetic and legibility scores of Chinese handwriting [2]. Beyond conventional CNN architectures, active learning frameworks have also been introduced for handwriting quality assessment, showing the effectiveness of combining statistical features and deep learning techniques [3]. In addition, Transformer-based models have recently been applied to Chinese calligraphy evaluation, further demonstrating the adaptability of deep neural networks in handwriting scoring tasks [4]. Siamese-network-based quality evaluation approaches have also been proposed, indicating the potential of similarity-based handwriting assessment methods [5]. Existing approaches generally focus on extracting visual features from handwriting images and learning scoring functions through supervised training. However, most of these methods heavily depend on manually annotated datasets, which are expensive to construct and difficult to standardize due to variations in individual evaluation criteria.

For Chinese handwriting assessment, the lack of publicly available datasets containing reliable quality labels remains a significant challenge. Unlike handwriting recognition tasks, where character category labels can be easily obtained, handwriting quality evaluation requires expert judgment and professional knowledge. Consequently, acquiring large-scale annotated datasets is costly and often introduces subjective bias. This limitation restricts the practical application of fully supervised deep learning approaches in handwriting quality assessment.

To address the above challenges, a two-stage deep learning framework for automatic evaluation of hard-pen Chinese calligraphy is proposed in this study. In the first stage, a convolutional AutoEncoder is employed to learn structural representations of handwritten characters. AutoEncoder-based representation learning has been widely used for extracting latent structural features from image data and learning meaningful low-dimensional representations without manual annotation [6]. The reconstruction error generated by the AutoEncoder is then transformed into pseudo-labels, providing an unsupervised measurement of handwriting quality without requiring manual annotation. In the second stage, the generated pseudo-labels are used to train supervised regression networks, including ResNet18 [7] and MobileNetV2 [8], to predict handwriting quality scores directly from input images. By combining unsupervised representation learning and supervised score prediction, the proposed framework effectively reduces annotation costs while maintaining reliable evaluation performance.

Compared with existing handwriting evaluation methods, the proposed approach offers several advantages. First, it eliminates the dependence on manually labeled quality scores by introducing an AutoEncoder-based pseudo-label generation strategy. Second, the two-stage framework enables deep neural networks to learn handwriting quality representations from large-scale unlabeled datasets. Third, both heavyweight and lightweight regression models are investigated, providing valuable insights into the trade-off between prediction accuracy and computational efficiency.

The main contributions of this work can be summarized as follows:

1. A pseudo-label generation method based on AutoEncoder reconstruction error is proposed to achieve handwriting quality assessment without manual annotation.

2. A two-stage deep learning framework integrating unsupervised scoring and supervised regression learning is developed for automatic hard-pen Chinese calligraphy evaluation.
3. A comparative study of ResNet18 and MobileNetV2 is conducted to analyze the balance between evaluation accuracy and computational efficiency for practical deployment scenarios.

The remainder of this paper is organized as follows. Section 2 presents the proposed methodology, including pseudo-label generation and regression model design. Section 3 describes the experimental setup and evaluation results. Finally, Section 4 concludes the paper and discusses future research directions.

II. METHODOLOGY

A. Overall Framework

To achieve automatic evaluation of hard-pen Chinese calligraphy without manual annotation, a two-stage deep learning framework is proposed in this study. The overall workflow consists of four main components: data preprocessing, pseudo-label generation, supervised score learning, and result evaluation.

Unlike traditional handwriting assessment methods that rely on expert scoring, the proposed framework first utilizes an AutoEncoder to learn the intrinsic structural characteristics of handwritten Chinese characters. The reconstruction error generated by the AutoEncoder is then transformed into pseudo-labels, which are used as handwriting quality indicators.

Subsequently, the generated pseudo-labels are employed to train supervised regression models. ResNet18 and MobileNetV2 are selected as representative deep neural networks to learn the mapping relationship between handwritten character images and handwriting quality scores. Finally, the trained models are used to predict handwriting quality and provide quantitative evaluation results.

The proposed framework combines the advantages of unsupervised representation learning and supervised regression modeling. It effectively reduces the dependence on manually labeled data and enables large-scale handwriting quality assessment.

To provide a clear overview of the proposed handwriting evaluation framework, the complete processing pipeline is illustrated in Fig. 1. The framework consists of four main stages, including image preprocessing, pseudo-label generation, supervised regression learning, and score prediction. By combining unsupervised feature learning with supervised score regression, the proposed method can effectively evaluate handwriting quality without requiring manually annotated labels.

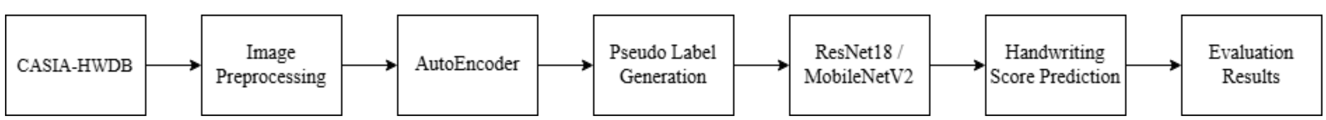

Figure 1. Overall framework of the proposed two-stage handwriting evaluation system

B. AutoEncoder-Based Pseudo-Label Generation

The reconstruction errors produced by the convolutional AutoEncoder serve as pseudo-labels for supervised regression, following principles similar to recent self-supervised handwriting verification approaches [9].

Handwriting quality is closely related to character structure, stroke arrangement, and spatial layout. To automatically evaluate handwriting quality without manual annotation, a convolutional AutoEncoder is adopted to generate pseudo-labels.

A convolutional AutoEncoder is employed to extract latent structural representations from handwritten characters. The architecture of the AutoEncoder is presented in Fig. 2. Through the encoding and decoding processes, the model learns to reconstruct input images while preserving essential stroke and structural information. The reconstruction error generated during this process serves as the basis for pseudo-label generation.

Figure 2. Architecture of the convolutional AutoEncoder

The AutoEncoder consists of two main components: an encoder and a decoder. The encoder compresses the input handwriting image into a low-dimensional latent representation, while the decoder reconstructs the original image from the extracted features. Through this reconstruction process, the network learns the underlying structural characteristics of handwritten Chinese characters.

The reconstruction performance is measured using Mean Squared Error (MSE), which can be expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2 \qquad (1)$$

where x_i denotes the original pixel value, $\hat{x}_i$represents the reconstructed pixel value, and $N$is the total number of pixels.

The pseudo-label generation procedure plays a critical role in the proposed framework. As shown in Fig. 3, the reconstruction error produced by the AutoEncoder is first calculated and then transformed through logarithmic mapping and normalization operations. The resulting pseudo-labels provide an unsupervised measurement of handwriting quality and are subsequently used for regression model training.

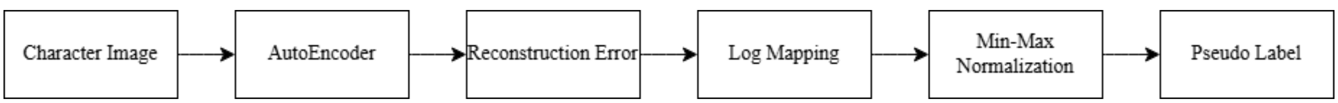

Figure 3. Pseudo-label Generation Strategy

The basic assumption of this work is that well-written characters with regular structures and clear strokes can be reconstructed more accurately by the AutoEncoder. Therefore, such samples produce lower reconstruction errors. In contrast, characters with distorted structures, irregular stroke arrangements, or poor writing quality tend to generate larger reconstruction errors.

After training, the AutoEncoder is applied to all handwriting samples in the dataset. The reconstruction errors become handwriting quality scores through logarithmic mapping and Min-Max normalization. These scores serve as pseudo-labels for subsequent supervised learning.

Compared with traditional manual scoring approaches, the proposed pseudo-label generation strategy significantly reduces annotation costs while maintaining objective and consistent evaluation criteria.

TABLE I. NETWORK ARCHITECTURE OF THE AUTOENCODER

Layer	Output Size
Conv1	32×32×32
Conv2	64×16×16
DeConv1	32×32×32
DeConv2	1×64×64

C. ResNet18-Based Regression Model

To further learn the relationship between handwriting appearance and quality scores, ResNet18 is employed as the first supervised regression model.

ResNet18 introduces residual learning mechanisms that effectively alleviate the degradation problem encountered in deep neural networks. Its strong feature extraction capability makes it suitable for capturing complex stroke patterns and structural characteristics of handwritten Chinese characters.

Since the original ResNet18 architecture is designed for RGB image classification tasks, several modifications are introduced to adapt it to the handwriting evaluation task. The overall modified architecture is shown in Fig. 4. First, the input layer is modified to accept single-channel grayscale images instead of three-channel color images. Second, the original classification layer is replaced with a regression layer containing a single output node. Consequently, the network predicts continuous handwriting quality scores rather than category labels.

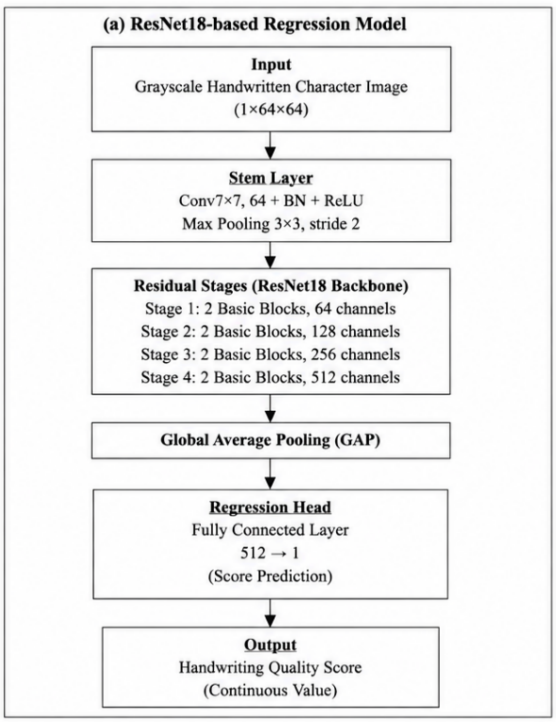

Figure 4. ResNet18-based Regression Model

During training, ResNet18 learns the mapping relationship between handwriting images and pseudo-labels generated by the AutoEncoder. By minimizing prediction errors, the network gradually acquires the ability to evaluate handwriting quality directly from image data.

D. MobileNetV2-Based Regression Model

Although ResNet18 demonstrates strong performance in feature extraction, its relatively large number of parameters may limit its deployment in resource-constrained environments. Therefore, MobileNetV2 is investigated as a lightweight alternative.

MobileNetV2 adopts depthwise separable convolutions and inverted residual structures to significantly reduce computational complexity while maintaining effective feature representation capability. Compared with conventional convolutional neural networks, MobileNetV2 requires fewer parameters and lower computational cost.

Similar to the modifications applied to ResNet18, the first convolutional layer of MobileNetV2 is adapted for grayscale handwriting images, and the final classification layer is replaced by a single-node regression output. The overall architecture of the MobileNetV2-based regression model is shown in Fig. 5. The network is then trained using the pseudo-labels generated by the AutoEncoder.

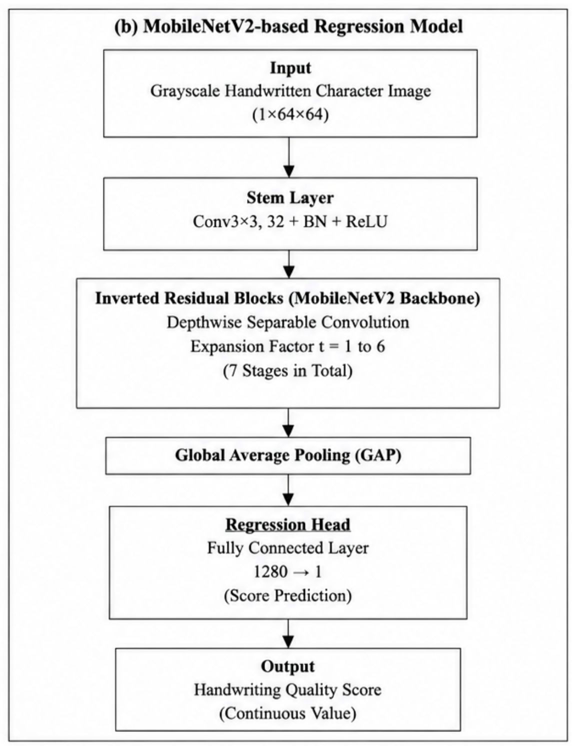

Figure 5. MobileNetV2-based Regression Model

The lightweight nature of MobileNetV2 makes it particularly suitable for deployment on mobile devices and embedded platforms. Through comparison with ResNet18, the effectiveness of lightweight networks for handwriting quality evaluation can be systematically analyzed.

E. Loss Function and Optimization Strategy

Handwriting quality prediction is formulated as a regression problem. Therefore, Mean Squared Error (MSE) is selected as the loss function:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

where y_i denotes the pseudo-label generated by the AutoEncoder and $\hat{y}_i$ represents the predicted score produced by the regression model.

To optimize network parameters efficiently, the Adam optimizer is employed during training. Adam combines adaptive learning rate adjustment and momentum optimization, providing fast convergence and stable training performance for large-scale datasets.

Compared with other automated evaluation frameworks that leverage feature interpretability for aesthetic assessment [10], the proposed two-stage framework provides a balance between unsupervised feature extraction and supervised regression for effective quality scoring.

Both ResNet18 and MobileNetV2 are trained by minimizing the difference between predicted scores and pseudo-labels. Through iterative parameter updates, the models gradually learn effective handwriting quality representations and achieve reliable score prediction performance.

III. Experiments and Results

A. Dataset and Preprocessing

The experiments are conducted on the CASIA-HWDB dataset, which contains over 450,000 handwritten Chinese characters across 3,755 classes. This dataset is widely used for handwriting recognition and calligraphy analysis.

Prior to training, the images undergo preprocessing:

1) *Grayscale conversion:* Color information is not required; converting to single channel reduces computation.

2) *Resize and Padding:* All images are resized to 64×64 pixels. Padding ensures the original aspect ratio is preserved, preventing distortions in stroke structure.

3) *Tensor transformation:* Images are converted to PyTorch tensors, with pixel values normalized to the [0,1] range and dimensions set to C×H×W for model input.

The dataset statistics are summarized in Table 2.

TABLE II. DATASET STATISTICS

Dataset	Characters	Samples
CASIA-HWDB	3755	450,000+

B. AutoEncoder Results

The convolutional AutoEncoder is first trained to generate pseudo-labels based on reconstruction error. The AutoEncoder compresses each input character into latent feature representations and reconstructs it back to the original image. The reconstruction error is computed using the mean squared error (MSE) defined in Equation (1).

The reconstruction error serves as an indicator of handwriting quality. Characters with well-formed strokes and regular structures tend to be reconstructed with higher fidelity, resulting in lower reconstruction errors. In contrast, characters with irregular stroke patterns or structural distortions are more difficult to reconstruct accurately, leading to higher reconstruction errors.

After training, the AutoEncoder is applied to the entire dataset to compute reconstruction errors for all samples. These errors are then transformed into pseudo-labels through logarithmic mapping and Min-Max normalization. The resulting values are used as surrogate handwriting quality scores for subsequent supervised regression training.

Fig. 6 illustrates representative examples of original and reconstructed handwritten characters, demonstrating that the AutoEncoder successfully captures structural and stroke-level information. Characters with regular and well-proportioned structures are reconstructed with high fidelity, whereas distorted or irregular characters exhibit noticeable reconstruction errors. This visualization highlights the model's capability to extract meaningful latent representations and provides a basis for generating pseudo-labels used in the subsequent supervised regression stage. The comparison emphasizes the effectiveness of the AutoEncoder in identifying structural quality, which is critical for reliable handwriting evaluation.

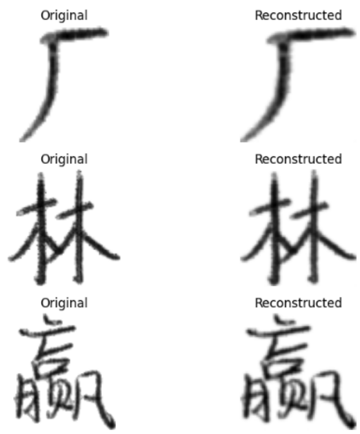

Figure 6. Reconstruction examples generated by the AutoEncoder

C. Supervised Regression Models

The pseudo-labels generated by AutoEncoder are then used to train two supervised networks, ResNet18 and MobileNetV2, for score prediction.

Training Setup:

- Loss function: Mean Squared Error (MSE)
- Optimizer: Adam
- Input: 64×64 grayscale images
- Output: single-node regression predicting handwriting score.

1) Training and Validation Loss.

Fig. 7 shows the training and validation loss curves of both ResNet18 and MobileNetV2 regression models. ResNet18 converges faster and achieves a lower final loss, indicating superior feature extraction capability and more precise handwriting quality prediction. MobileNetV2, although slightly higher in loss, maintains stable convergence with significantly fewer parameters, highlighting its efficiency and suitability for deployment in resource-constrained environments. The loss curves also demonstrate that both models avoid overfitting, providing robust and consistent evaluation performance throughout training.

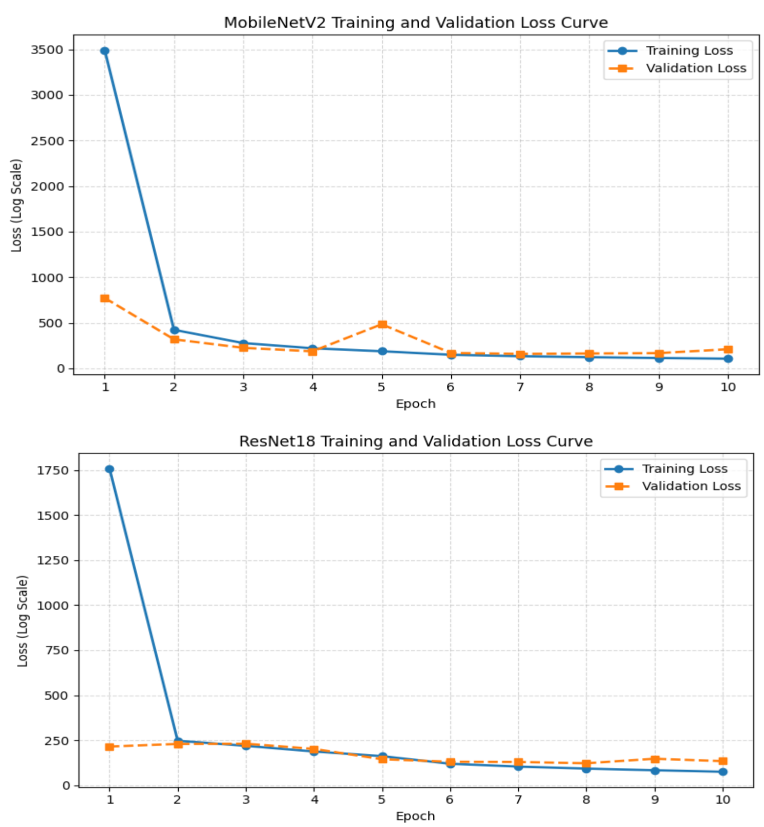

Figure 7. Training and validation loss curves

2) Score Distribution

The predicted handwriting scores for both models are analyzed to evaluate the consistency and stability of the framework. As illustrated in Fig. 8, scores are concentrated primarily in the 650–700 range. This distribution indicates that the proposed approach consistently distinguishes high-quality characters from lower-quality ones. The concentrated score range also demonstrates the stability of the models' predictions across diverse handwriting styles, which is essential for reliable automatic assessment in educational and intelligent tutoring applications.

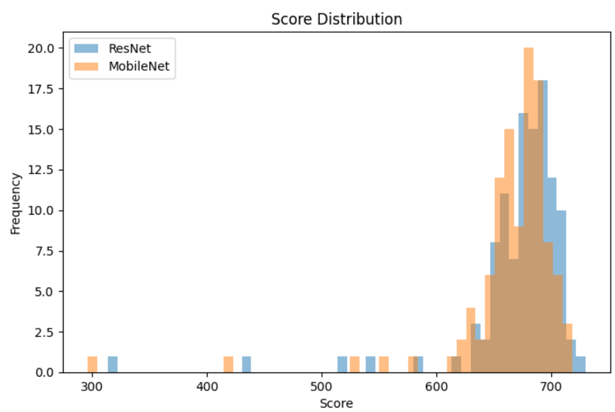

Figure 8. Score Distribution

3) High-score and Low-score Sample Analysis

To evaluate the effectiveness and interpretability of the proposed scoring models, representative high-scoring and low-scoring handwritten characters are visualized in Fig. 9. It is observed that high-scoring samples typically exhibit clear, well-structured strokes, balanced composition, and consistent spatial organization, reflecting good handwriting quality. In contrast, low-scoring samples are characterized by irregular stroke formations, disconnected components, or structural distortions, indicating poor writing quality or instability in character construction.

This comparative visualization provides qualitative evidence that both ResNet18 and MobileNetV2 are capable of capturing fine-grained variations in handwriting structure. More importantly, the consistency between visual quality differences and predicted scores demonstrates that the learned regression models are not only able to produce numerical evaluations, but also implicitly learn meaningful representations of stroke clarity and structural regularity. Therefore, the results offer interpretable support for the effectiveness of the proposed two-stage framework in distinguishing handwriting quality across diverse samples.

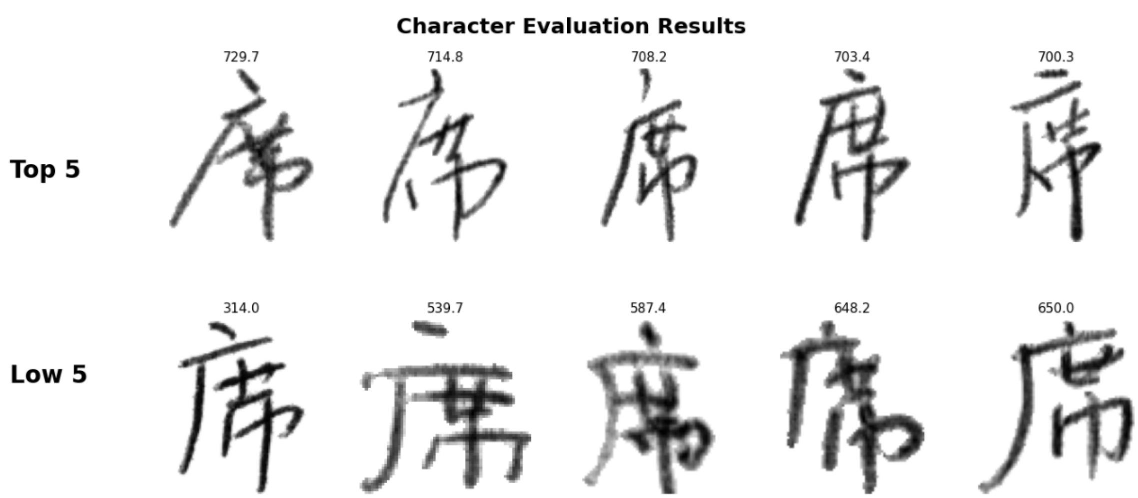

Figure 9. Character Evaluation Results

D. Model Comparison

Table 3 summarizes the performance comparison between ResNet18 and MobileNetV2.

- ResNet18: better performance on complex characters, higher computational cost.
- MobileNetV2: lightweight, fast, suitable for deployment on resource-limited devices.

TABLE III. PERFORMANCE COMPARISON OF RESNET18 AND MOBILENETV2

Model	Final Loss	Parameters	Characteristics
ResNet18	74.82	11.2M	High Accuracy
MobileNetV2	105.59	3.5M	Lightweight, Efficient

E. Discussion

The central assumption of the proposed method is that the reconstruction error generated by the convolutional AutoEncoder can serve as a reliable indicator of handwriting quality. Characters with well-formed strokes, balanced structures, and regular spatial layouts are easier for the AutoEncoder to reconstruct, resulting in lower reconstruction errors. In contrast, characters with irregular stroke arrangements, disconnected strokes, or distorted structures produce higher reconstruction errors. This assumption aligns with the principle of anomaly detection, where samples that deviate from the learned distribution yield larger reconstruction errors.

Several factors support the validity of this approach. First, the AutoEncoder is trained on a large-scale dataset containing diverse handwriting styles. As a result, it learns the dominant structural patterns and common stroke configurations present in standard handwriting. Characters that conform to these learned patterns are reconstructed accurately, whereas outliers or poorly written samples deviate from the learned distribution, producing larger errors. Second, the use of convolutional layers allows the model to capture local spatial relationships, stroke connectivity, and topological structures, which are essential for evaluating character quality. Third, normalization and logarithmic mapping of reconstruction errors ensure that the pseudo-labels are numerically stable and suitable for downstream supervised regression training.

The pseudo-label strategy effectively bridges the gap between unsupervised feature learning and supervised score prediction. By converting reconstruction errors into normalized quality scores, the model generates objective and consistent labels without requiring manual annotation. These pseudo-labels provide a sufficient training signal for regression networks such as ResNet18 and MobileNetV2, enabling accurate prediction of handwriting quality across large datasets.

Potential limitations of this approach include sensitivity to dataset bias and variations in writing styles. If the training dataset contains predominantly standard or ideal handwriting, the model may assign disproportionately high reconstruction errors to unconventional but still acceptable writing styles. To mitigate this, a sufficiently diverse dataset and appropriate preprocessing techniques, such as padding and aspect ratio preservation, are employed to ensure fair evaluation of all handwriting samples.

In summary, the AutoEncoder reconstruction error provides a theoretically justified and empirically validated metric for handwriting quality assessment. When combined with supervised regression models, this pseudo-label strategy forms a robust framework for automatic evaluation of hard-pen Chinese calligraphy.

F. Summary

The experiments confirm that the proposed two-stage framework:

1) Generates effective pseudo-labels using AutoEncoder reconstruction error

2) Accurately predicts handwriting scores using supervised networks

3) Provides a practical trade-off between model accuracy and efficiency

Overall, the framework demonstrates both scalability and effectiveness for automated hard-pen Chinese calligraphy evaluation.

IV. CONCLUSIONS

A novel two-stage deep learning framework for automatic evaluation of hard-pen Chinese calligraphy is proposed in this study. By integrating an AutoEncoder-based pseudo-label generation strategy with supervised regression networks, including ResNet18 and MobileNetV2, the framework effectively eliminates the need for manual annotation while providing accurate and scalable handwriting quality assessment.

Experimental results on the CASIA-HWDB dataset demonstrate that the framework can reliably predict handwriting quality scores. The reconstruction error of the AutoEncoder serves as an effective unsupervised metric, capturing structural irregularities and stroke variations in handwritten characters. Subsequent supervised regression models accurately map character images to predicted quality scores, with ResNet18 achieving higher accuracy and MobileNetV2 providing a lightweight alternative suitable for deployment in resource-constrained environments.

The comparative analysis between heavyweight and lightweight networks provides insights into the trade-offs between model accuracy and computational efficiency. These findings indicate that the proposed framework can be adapted to diverse practical scenarios, including educational applications and mobile handwriting assessment tools.

Future work may include extending the framework to multiple handwriting styles and calligraphy types, incorporating explainable AI techniques to provide interpretable evaluation criteria, and integrating the system into intelligent tutoring platforms for real-time feedback.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the developers and maintainers of the CASIA-HWDB dataset for making large-scale Chinese handwriting data publicly available. Their contributions have greatly facilitated research in handwriting analysis and evaluation.

Funding: This research was funded by the College Students' Innovative Entrepreneurial Training Plan Program (Nos. S202510702129 and X202510702200).

REFERENCES

[1] Chen, W., Su, J., Song, W. et al. Quality evaluation methods of handwritten Chinese characters: a comprehensive survey. Multimedia Systems 30, 194 (2024).

[2] Xu, Z., Mittal, P.S., Ahmed, M.M. et al. Assessing penmanship of Chinese handwriting: a deep learning-based approach. Read Writ 38, 723–743 (2025).

[3] Koushik, K.S., Nair, B.J.B., Rani, N.S. et al. HQA2LFS-handwriting quality assessment using an active learning framework in smartphones. Sci Rep 16, 8186 (2026).

[4] Yan, F.; Lan, X.; Zhang, H.; Li, L. Intelligent Evaluation of Chinese Hard-Pen Calligraphy Using a Siamese Transformer Network. Appl. Sci. 2024, 14, 2051.

[5] Chen, W., Zhu, G., Li, Y. et al. SiamHCC: a novel siamese network for quality evaluation of handwritten Chinese characters. Multimedia Systems 31, 303 (2025).

[6] Torabi H, Mirtaheri S L, Greco S. Practical autoencoder based anomaly detection by using vector reconstruction error[J]. Cybersecurity, 2023, 6(1): 1.

[7] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[8] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.

[9] Chauhan M, Hashemi M A, Satbhai A, et al. Self-supervised learning based handwriting verification [C]//IET Conference Proceedings CP887. Stevenage, UK: The Institution of Engineering and Technology, 2024, 2024(10): 170-177.

[10] Yoshida S, Yatoh N, Muneyasu M. Aesthetic evaluation of chinese calligraphy using tabnet: Interpretability and novel features for improved accuracy[J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2025, 108(3): 357-361.