2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Ancient Chinese Character Image Retrieval Based on Self-Attention Mechanism and Multi-Scale Feature Fusion

Ye Yang

School of Cyber Security and Computer, Hebei University, Baoding 071002, China
17306819170@163.com

Abstract. The variety of fonts and shapes of ancient Chinese characters, along with the notable disparities in glyphs, pose substantial difficulties for the retrieval of ancient Chinese character images. In this research, we developed a Chinese character image retrieval model that relies on the self-attention mechanism and multi-scale feature fusion (SAMSFF). The aim is to enhance the feature extraction capacity and retrieval precision of Chinese character images in ancient documents. Firstly, an improved inverted residual module called HardFused IB was constructed using the optimized SE attention mechanism to obtain the enhanced features of key information. Secondly, the static dynamic context fusion module was used to fully use the context information between adjacent keys to improve the expressiveness and representativeness of the output features. Finally, the bilinear multi-scale feature fusion module BMSFblock was constructed to perform an adaptive fusion of the multi-layer features extracted by the designed network. The network measures the Euclidean distance between the queried and candidate images and sorts and returns the most relevant results. The mAP@-1 of the retrieval method proposed in this paper on the ancient Chinese character image dataset is 0.932. Experiments prove the model efficiently extracts ancient Chinese character image features, boosts retrieval accuracy, and excels in relevant retrieval tasks.

Keywords: Ancient Chinese Character images, Image Retrieval, Self-Attention Mechanism, Multi-Scale Feature Fusion.

1 Introduction

In the traditional study of ancient books, researchers often spend a lot of time and effort consulting many documents manually. By constructing a digitized image retrieval system for ancient Chinese character images, the retrieval efficiency of ancient literature can be significantly improved. This will help further improve the quality of research on Chinese characters in ancient books.

However, there are differences in the glyphs of Chinese characters in many ancient books under different writing styles, and there are problems such as non-standard writing or lack of consistency, which brings difficulties to the traditional text image retrieval system. Therefore, considering the traits of ancient Chinese character glyph images, this study proposes an SAMSFF-based retrieval model. Using the self-attention mechanism, it enhances the feature representation of key information. Additionally, it integrates multi-scale feature fusion technology to boost the representation of features at various levels of Chinese character images, thereby enhancing the accuracy and stability of retrieving ancient Chinese character images. During the retrieval phase, the enhanced features produced by the aforementioned modules are employed to assess the similarity between images through calculating the Euclidean distance, and the candidate images with the highest relevance are ranked according to the calculated similarity values.

The remainder of the article is structured as follows: Chapter 2 reviews related work. Chapter 3 introduces the neural network model proposed in this study. Chapter 4 presents experimental investigations and performance comparisons. Finally, Chapter 5 concludes the research.

2 Related Work

With the development of science and technology, especially computer vision and deep learning [7], significant progress has been made in ancient Chinese character image retrieval. Research in this area falls into two categories: traditional methods and deep learning-based approaches.

Among the traditional methods, Qu et al. [9] proposed an improved adaptive discriminant local alignment method for Chinese character image recognition, which successfully solved the problem of dependence on discriminant local alignment parameters and achieved better recognition results while maintaining low cost, so it was more suitable for practical application. Zhang et al. [18] proposed a calligraphic character retrieval method based on shape similarity, which achieved good results in recall and accuracy by using contour similarity. However, in the face of a large data volume of calligraphy character retrieval, the retrieval speed is slow, and the effect is not ideal. To solve this problem and inspired by the mature term frequency-inverse document frequency (TF-IDF) method in text retrieval, an adaptive calligraphy character matching and retrieval method was proposed, which was adapted to prune and match according to the eigenvalues of different query samples. Experiments demonstrate the proposed method significantly boosts retrieval efficiency [17]. Nevertheless, these traditional methods depend on manually crafted features, which makes them ill-equipped to handle the diversity of glyphs, complex handwriting, and distortions. They are also highly sensitive to noise and deformation, leading to reduced accuracy and robustness when dealing with non-standard writing.

With the continuous development of deep learning theories and methods, feature extraction based on convolutional neural networks has become a mainstream technology in image retrieval. It has shown significant advantages [12]. Considering the drawbacks of traditional Chinese character image retrieval in ancient Chinese character retrieval, Cai et al. [2] introduced a stratified retrieval approach for ancient Chinese character images, which was grounded in region saliency and skeleton matching. This approach demonstrated its outstanding performance in the retrieval of ancient Chinese character images. To solve the problem that the complex structure and diverse writing styles of ancient Chinese characters increase the difficulty of the implementation of image retrieval of ancient Chinese characters, Wang et al. [11] developed a multi-layer feature adaptive fusion model for ancient Chinese character image retrieval, which enhanced the retrieval performance to some degree.

3 Method

Fig. 1 illustrates the architecture of the proposed SAMSFF network.

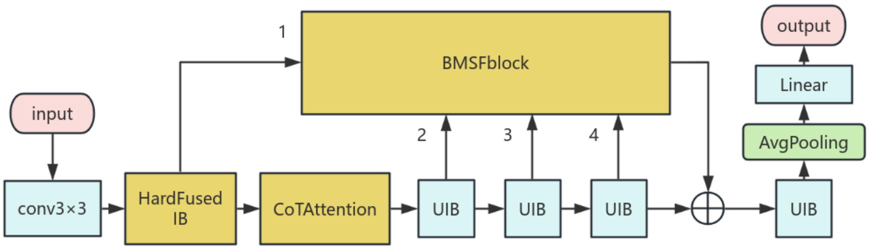

Fig. 1. The Overall Framework of the SAMSFF Network.

3.1 Improved Inverted Residual Module

The module is improved based on the Fused Inverted Bottleneck (Fused IB) module in MobileNetV4 [8]. Fig. 2 shows the improved structure of the HardFused Inverted Bottleneck (HardFused IB) module.

The Squeeze-and-Excitation (SE) attention mechanism [4] introduced plays a crucial role in image feature extraction, and the expression ability of important feature channels is enhanced by dynamically adjusting the weights between channels. Specifically, the SE module initially reduces the spatial information of every channel to a single scalar value by means of global average pooling, which is known as the Squeeze operation. Then, it generates a weight coefficient for each channel using a small, fully connected network (the Excitation operation). These weights reflect the importance of the channels, allowing the network to automatically increase the focus on key features while suppressing irrelevant channels, improving the model’s performance.

The SE module helps the network capture key information more effectively by reducing background noise interference, especially in complex backgrounds or fine-grained classification. It enhances the response to important features, improving classification accuracy and robustness. Unlike traditional convolutional layers, the SE module makes the model pay more attention to key features. HardFused IB module that integrates SE improves feature extraction capabilities without significantly increasing computational burden, enhancing classification performance.

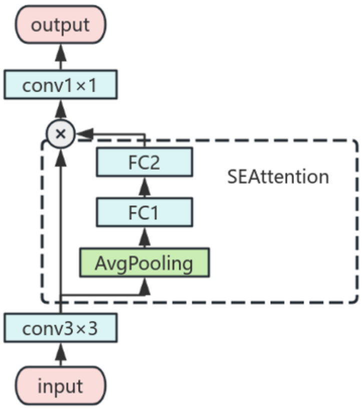

Fig. 2. Structure Diagram of HardFused IB.

In addition, when building this module, this paper replaces the ReLU and Sigmoid activation functions in the original SE module with HardSwish and HardSigmoid, respectively, which offers several advantages. The formulas for these four activation functions are as follows:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{1}$$

$$\mathrm{Sigmoid}(x) = \frac{1}{1+e^{-x}} \tag{2}$$

$$\mathrm{HardSwish}(x) = x \cdot (\frac{\min(\max(x+3, 0), 6)}{6}) \tag{3}$$

$$\mathrm{HardSigmoid}(x) = \frac{\min(\max(x+3, 0), 6)}{6} \tag{4}$$

The comparison chart of the activation function is shown in Fig. 3.

As shown in the figure, HardSwish has a smoother transition in the negative region than ReLU, which avoids the hard switching of ReLU, mitigating the problem of gradient vanishing or explosion and maintaining better nonlinear characteristics. This helps improve the model’s representation ability. At the same time, HardSigmoid approximates Sigmoid through linear transformation, avoiding the complex exponential operation in Sigmoid and significantly reducing computational overhead. This mainly benefits deep networks, improving inference speed and reducing power consumption. The improved SE module pseudocode is shown in Algorithm 1.

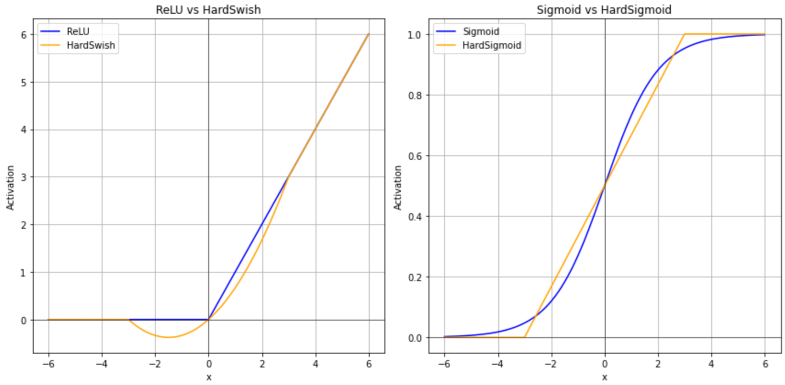

Fig. 3. Comparison of Activation Functions.

Algorithm 1. Improved SE Module

```
Input: Input feature map x, compression ratio se_ratio, input channels input_c, squeezed
    channels squeeze_c, expanded channels expand_c
Output: Output feature map output
squeeze_c = input_c * se_ratio    // Calculate the squeezed channel number
scale = mean(x, dim=(2, 3), keepdim=True)    // Global average pooling to obtain the average value of each channel
scale = Conv2d(scale, kernel_size=1, out_channels=squeeze_c)    // Compress the channel number
scale = Hardswish(scale)    // Apply the H-Swish activation function
scale = Conv2d(scale, kernel_size=1, out_channels=expand_c)    // Restore the channel number
scale = Hardsigmoid(scale)    // Apply the H-Sigmoid activation function
output = scale * x    // Output the result by element-wise multiplication of the input and the scaling factor
return output
```

3.2 Static Dynamic Context Fusion Module

Traditional attention mechanisms [1,10,16], mainly depending on the input, are capable of efficiently initiating feature interactions at various spatial positions. However, in the traditional self-attention mechanism, query-key relationships are learned independently, focusing only on individual pairs. This approach does not fully explore the abundant contextual information among them. Consequently, during the process of visual representation learning on 2D feature maps, the learning capacity of self-attention is significantly restricted.To solve this problem, the Contextual Transformer (CoT) module [6] is introduced in this paper, as shown in Fig. 4.

This module integrates contextual information mining and self-attention learning into a unified framework. The core idea is to fully utilize the contextual information between adjacent keys to enhance self-attention learning efficiently, thereby improving the expressiveness and representativeness of output features. The pseudocode of the CoTAttention module is shown in Algorithm 2.

```
Algorithm 2. CoTAttention
Input: Input feature map x, batch size B, input channels C, feature map height H, feature
    map width W
Output: Output feature map output
k1 = key_embed(x)    // Encode static context information
v = value_embed(x).view(B, C, -1)    // Encode the value matrix
y = concatenate(k1, x, dim=1)    // Concatenate k1 and x to obtain the input for attention
att = attention_embed(y)    // Compute the attention matrix
att = reshape(att, B, C, kernel_size * kernel_size, H, W)    // Reshape the tensor
att = mean(att, dim=2)    // Apply average pooling to get the attention weights for the co-
    ordinates
k2 = softmax(att, dim=-1) * v    // Compute the softmax weights and multiply with the
    value matrix
k2 = reshape(k2, B, C, H, W)    // Restore to the original shape
output = k1 + k2    // Generate the final output
return output
```

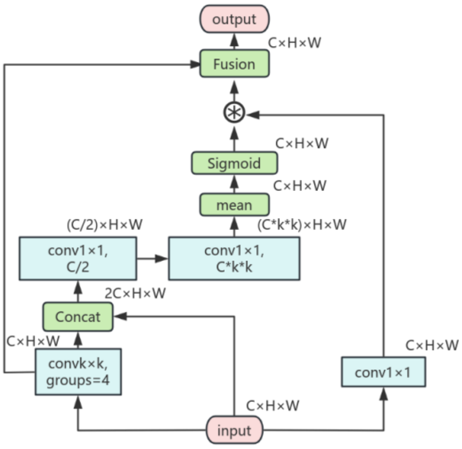

Fig. 4. Structure Diagram of CoTAttention.

3.3 Bilinear Multi-Scale Feature Fusion Module

This module improves the Multi-scale Selective Fusion Module (MSFblock) in the existing SHISRCNet model [13]. Although the module performs well in image classification tasks, it still has certain limitations when dealing with complex Chinese character image retrieval tasks, particularly feature extraction and multi-scale information fusion. To address this, this paper proposes an improved BMSFblock, which extracts features at different levels through multi-scale feature fusion and further enhances the model's expressive ability using an adaptive weight adjustment mechanism. The BMSFblock structure is shown in Fig. 5.

Through the operation of dimensionality reduction, interpolation, pooling, and weighted fusion of multi-scale features, the module effectively fuses feature information from different levels. It enhances the ability to capture Chinese characters' complex local and global features. The 1×1 convolution for dimensionality reduction and bilinear interpolation are used to unify the feature map size. This ensures that the features of different scales can be fused at the same spatial scale and maintain the contextual information. The global average pooling and channel weighting mechanism adaptively selects the key features, which improves the model's attention to the important features, especially when dealing with the writing characteristics of Chinese characters, such as stroke thickness and structural symmetry. Finally, through weighted fusion and optimization, the generated feature map not only comprehensively reflects the multi-scale information but also improves the accuracy and robustness of the model. The pseudocode of the BMSFblock is shown in Algorithm 3.

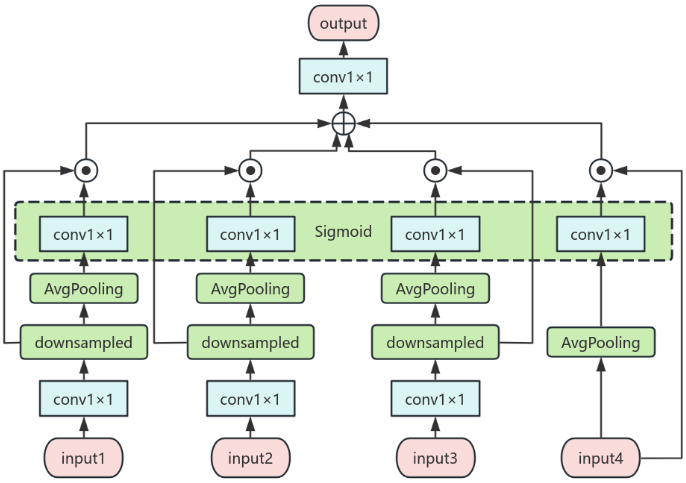

Fig. 5. BMSFblock Structure Diagram.

Algorithm 3. BMSFblock

Input: Input feature maps *x0*, *x1*, *x2*, *x3*, height *H3* of *x3*, width *W3* of *x3*
Output: Output feature map *output*

```
x0_downsampled = Interpolate(Conv1x1(x0), size=(H3, W3))    // Downsample
x1_downsampled = Interpolate(Conv1x1(x1), size=(H3, W3))
x2_downsampled = Interpolate(Conv1x1(x2), size=(H3, W3))
y0_weight = Conv2d(AdaptiveAvgPool(x0_downsampled))    // Global pooling and feature processing
y1_weight = Conv2d(AdaptiveAvgPool(x1_downsampled))
y2_weight = Conv2d(AdaptiveAvgPool(x2_downsampled))
y3_weight = Conv2d(AdaptiveAvgPool(x3))
weight = Concatenate([y0_weight, y1_weight, y2_weight, y3_weight], dim=2)
    // Compute channel description weights
```

```
weight = Softmax(Sigmoid(weight))    // Scale weighting
y0_weight = Unsqueeze(weight[:, :, 0], dim=2)    // Compute the weighting coefficient for
    each scale and apply element-wise multiplication to weight each scale's features
y1_weight = Unsqueeze(weight[:, :, 1], dim=2)
y2_weight = Unsqueeze(weight[:, :, 2], dim=2)
y3_weight = Unsqueeze(weight[:, :, 3], dim=2)
x_att = y0_weight * x0 + y1_weight * x1 + y2_weight * x2 + y3_weight * x3
    // Fuse features from multiple scales
output = Project(x_att)    // Apply projection to the weighted features to obtain the output
return output
```

4 Experiments

4.1 Dataset and Evaluation Metrics

In this study, numerous samples were collected from the Siku Quanshu, a seminal ancient Chinese text. To enhance sample diversity and prevent overfitting, ancient Chinese characters were annotated. Data augmentation methods, including affine transformation, rotation, and Gaussian noise, were employed to create a dataset with 633 image categories. Finally, the dataset was divided into a retrieval set, a test set, and a training set, with sizes of 33,989, 3,508, and 29,767, respectively. An example of a dataset is shown in Fig. 6. Mean Average Precision (mAP), average retrieval time (ART), precision, and recall rate were used to evaluate the retrieval performance.

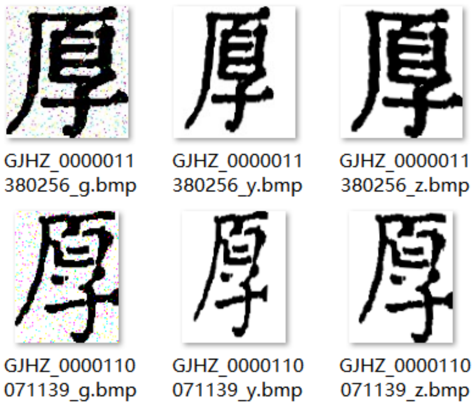

Fig. 6. Sample Dataset.

4.2 Performance Comparison

In this part, we validate the proposed model's effectiveness for the ancient Chinese character image retrieval task by comparing it with several classic Chinese character image retrieval models, including ResNeXt50 [14], RegNetX_12GF [15], DenseNet121 [5], and DPN [3]. We trained these four models using the same methodology to ensure a fair comparison. The performance comparison and retrieval effect comparison of the models are shown in Table 1 and Fig. 7.

Table 1. Performance Comparison of Different Models.

Method	Precision@60	Recall@60	ART(s)	mAP@-1
ResNeXt50	0.7443	0.8047	0.331	0.854
RegNetX_12GF	0.7090	0.7672	0.404	0.814
DenseNet121	0.7547	0.8153	**0.322**	0.865
DPN	0.7415	0.7996	0.491	0.844
Ours(SAMSFF)	**0.8175**	**0.8845**	0.336	**0.932**

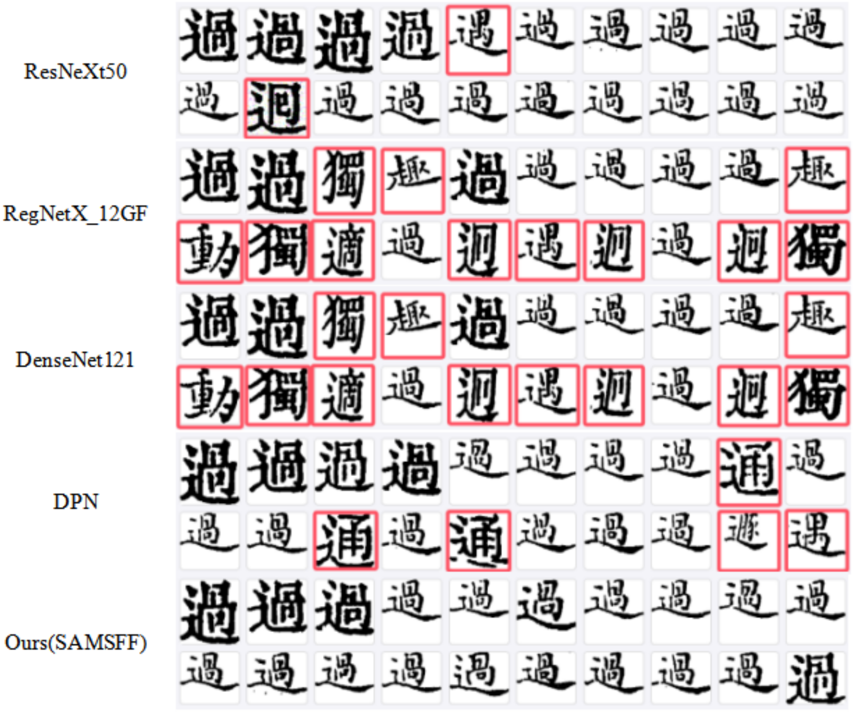

Fig. 7. Retrieval Performance of Different Models.

The results in the table show that the four single-network models—ResNeXt50, RegNetX_12GF, DenseNet121, and DPN—perform poorly when directly applied to the ancient Chinese character image dataset, which is characterized by diverse and complex stroke styles. By intelligently combining the HardFused IB, CoTAttention, and BMSFblock modules, the proposed model demonstrates significant improvements in the accuracy and robustness of ancient Chinese character image retrieval, outperforming classical deep learning models and showing clear advantages in precision, recall, and mAP. Therefore, the proposed model has strong potential for application in ancient Chinese character image retrieval tasks.

4.3 Ablation Experiments

To confirm the validity of the proposed enhanced method, ablation experiments were carried out. The specific experimental data are shown in Table 2, where baseline refers to MobileNetV4_medium, A represents the use of HardFused IB, B represents the use of CoTAttention, and C represents the use of BMSFblock. All experiments were performed on the same training and test datasets.

Table 2. Ablation Experiments.

Method	Precision@60	Recall@60	ART(s)	mAP@-1
(a) baseline	0.7703	0.8315	**0.315**	0.877
(b) baseline+A	0.7999	0.8637	0.321	0.910
(c) baseline+B	0.7841	0.8463	0.319	0.892
(d) baseline+C	0.8087	0.8739	0.323	0.922
(e) baseline+A+B	0.8058	0.8700	0.327	0.917
(f) baseline+A+C	0.8123	0.8782	0.331	0.927
(g) baseline+B+C	0.8125	0.8783	0.329	0.925
(h) SAMSFF(baseline+A+B+C)	**0.8175**	**0.8845**	0.336	**0.932**

As shown in the results in Table 2:

The HardFused IB module significantly improves model performance. By incorporating the HardFused IB module, the model's mAP increased by 3.3%, and Recall@60 increased by 3.22%. The CoTAttention module provides a relatively minor performance improvement, with mAP and Precision@60 improving by 1.5% and 1.38%, respectively. The BMSFblock module demonstrates a notable improvement in model performance, with Recall@60 increasing by 4.24% and an even more significant improvement in mAP. Combining multiple modules (e.g., model (e), model (f), and model (g)) further enhances model performance. In particular, the SAMSFF model (h) proposed in this paper achieves the best performance, with a 5.5% improvement in mAP compared to the baseline.

Integrating the HardFused IB, CoTAttention, and BMSFblock modules into MobileNetV4_medium significantly improves the retrieval performance for ancient Chinese character images. HardFused IB uses depthwise separable convolution to enhance computational efficiency and ensure the stability of deep network training. CoTAttention improves feature representation by dynamically modeling contextual information. BMSFblock effectively enhances the ability to capture features at different levels through multi-scale feature fusion and channel-dependent modeling. Combining these three modules makes the model more robust in handling challenges such as variations in Chinese character font styles, calligraphy strokes, and image quality, enabling it to adapt to deformations of different characters, complex backgrounds, and multi-scale feature information. The HardFused IB module also reduces computational overhead, accelerating the inference process. Combined with the fine-grained information capture capabilities of CoTAttention and BMSFblock, the model can more accurately extract stroke details and improve retrieval accuracy. This modular integration effectively enhances the accuracy and robustness of ancient Chinese character image retrieval.

5 Conclusion

In this paper, we propose an image retrieval network architecture based on the improved Inverted Residual Module (HardFused IB) and the Bilinear Multi-scale Feature Fusion Module (BMSFblock), combined with the Static-Dynamic Context Fusion Module (CoTAttention). First, the HardFused IB module, constructed using the optimized SE attention mechanism, effectively enhances key feature information and improves the network's attention to image details and important regions. Then, CoTAttention enhances the richness and representativeness of feature expressions by fully exploiting the contextual information between adjacent keys, thereby improving the accuracy and robustness of image retrieval. Finally, through the adaptive fusion of multi-scale features, the BMSFblock module enables the network to comprehensively consider feature information at different levels, further improving the integration capability of multi-scale information. Future research can focus on more efficient feature fusion methods and network optimization and explore deeper contextual information integration to enhance the model's robustness for hard-to-identify images. At the same time, improving the real-time performance and accuracy of the network without significantly increasing computational costs is also a key problem that urgently needs to be solved and deserves further in-depth investigation.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Cai, R., Tian, X.: Hierarchical retrieval of ancient chinese character images based on region saliency and skeleton matching. In: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data. pp. 268–282. Springer (2023)
3. Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks. Advances in neural information processing systems **30** (2017)
4. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
5. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
6. Li, Y., Yao, T., Pan, Y., Mei, T.: Contextual transformer networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **45**(2), 1489–1500 (2022)

7. Lu, H., Luo, M.: Survey on new progresses of deep learning-based computer vision. Journal of Data Acquisition & Processing/Shu Ju Cai Ji Yu Chu Li **37**(2) (2022)
8. Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., et al.: Mobilenetv4-universal models for the mobile ecosystem. arxiv 2024. arXiv preprint arXiv:2404.10518 (2024)
9. Qu, X., Xu, N., Wang, W., Lu, K.: Similar handwritten chinese character recognition based on adaptive discriminative locality alignment. In: 2015 14th IAPR International Conference on Machine Vision Applications (MVA). pp. 130–133. IEEE (2015)
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
11. Wang, X., Tian, X.: Ancient chinese character image retrieval based on fusing deep features and skeleton features. In: 2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS). pp. 2752–2755. IEEE (2023)
12. Wei, X., Luo, J., Wu, J., Zhou, Z.: Selective convolutional descriptor aggregation for fine-grained image retrieval. IEEE transactions on image processing **26**(6), 2868–2881 (2017)
13. Xie, L., Li, C., Wang, Z., Zhang, X., Chen, B., Shen, Q., Wu, Z.: Shisrcnet: super-resolution and classification network for low-resolution breast cancer histopathology image. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 23–32. Springer (2023)
14. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
15. Xu, J., Pan, Y., Pan, X., Hoi, S., Yi, Z., Xu, Z.: Regnet: self-regulated network for image classification. IEEE Transactions on Neural Networks and Learning Systems **34**(11), 9562–9567 (2022)
16. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
17. Zhang, X., Zhang, L., Han, D., Bi, K.: Adaptive matching and retrieval for calligraphic character. Journal of Zhejiang University (Engineering Science) **50**(04), 766–776 (2016)
18. Zhang, X., Zhuang, Y., Lu, W., Wu, F.: Chinese calligraphic character retrieval based on shape similarity. Journal of Computer-Aided Design Computer Graphics (11), 185–189 (2005)