CAAI Transactions on Intelligence Technology

WILEY

The Institution of Engineering and Technology

ORIGINAL RESEARCH OPEN ACCESS

Edge-Channel Aggregation Network and Two-Stage Fine Tuning Scheme for Handwritten Dongba Character Recognition

Xiali Li[1,2] | Yang Xiao[1,2] | Xiaojun Bi[1,2] | Yanlong Luo[3]

[1]Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing, China | [2]School of Information Engineering, Minzu University of China, Beijing, China | [3]College of Information and Communication Engineering, Harbin Engineering University, Harbin, China

Correspondence: Xiali Li (lixiali@muc.edu.cn; xiaer_li@163.com)

Received: 4 June 2025 | **Revised:** 14 June 2026 | **Accepted:** 17 June 2026

Keywords: channel attention mechanism | Dongba character recognition | edge attention mechanism | exponential moving average | fine tuning | image recognition

ABSTRACT

Handwritten Dongba Character Recognition (HDCR) contains a large number of visually similar characters with subtle and fragile edge cues, posing severe challenges to feature learning. To address this issue, an Edge Channel Aggregation Network (EdgeCANet) model is proposed. The core of EdgeCANet is the Edge Channel Attention Aggregation (ECAA) block, which integrates the Edge-Guided Enhancement Module (EGEM) for edge-aware spatial refinement and the Channel Texture Recalibration Module (CTRM) for channel-wise texture recalibration. Besides, the two-stage misclassified sample weight redistribution fine-tuning scheme based on Exponential Moving Average (EMA) algorithm is proposed to stabilise parameter updates, and to emphasise difficult and highly similar classes. The recognition accuracy of EdgeCANet significantly outperforms state-of-the-art models across the constructed high-similarity datasets HS-DB and HS-S20 K, as well as the public DB1404, OBC306, Sketch-20 K, and ImageNet-Sketch datasets. The results validate that EdgeCANet effectively enhances recognition accuracy for highly similar handwritten characters and ancient minority-script characters, offering more promising applicability for related recognition tasks.

1 | Introduction

Dongba character is one of the oldest pictographic writing systems in China, created by the ancestors of the Naxi ethnic group. As a pictographic language, it has been passed down through generations and is also referred to as the Naxi script. Dongba character chronicles the historical and cultural development of the Naxi people and provides valuable material for the study of Chinese history and culture. These characters are considered not only a significant legacy of historical and linguistic importance but also a treasure of world civilization. In recognition of their cultural value, Dongba characters have been inscribed in United Nations Educational, Scientific and Cultural Organization's Memory of the World Register.

Figure 1 shows a fragment of an ancient manuscript containing multiple Dongba characters [1]. The sentence in this image can be interpreted as "they even killed each other for more land, food, and slaves". Due to the complexity and visual similarity of the characters, non-specialists often face great difficulty in accurately recognising them. Since the handwritten Dongba

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

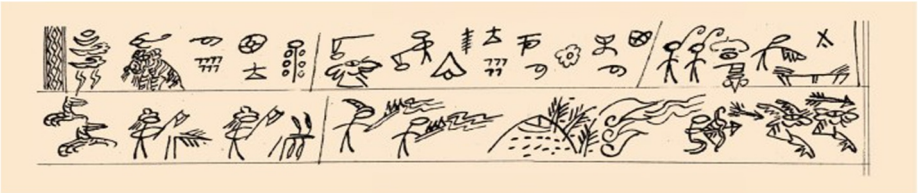


FIGURE 1 | A piece of ancient book paper with a sentence composed of multiple Dongba characters written on it.

character recognition (HDCR) task was first introduced in 2010, the field has made significant strides. For instance, a handwritten recognition model for Dongba characters based on a Multi-Attention Aggregation Network (MAAN) [2] effectively addresses the fine-grained feature recognition challenges posed by these pictographic characters. By employing hybrid attention mapping units and spatial attention aggregation modules, the model achieved a recognition accuracy of 99.34% on the DB1404 dataset, including 97.10% accuracy on a 32-class dataset of highly similar handwritten Dongba characters.

Although significant progress has been made in HDCR, several challenges remain. Many Dongba characters exhibit high visual similarity, differing only in subtle details, which imposes stringent demands on the extraction of fine-grained edge features.

To address these challenges, we propose an Edge Channel Aggregation Network (EdgeCANet) with a reduced parameter scale compared with the previous HDCR SOTA model and the two-stage misclassified sample weight redistribution fine-tuning scheme. EdgeCANet model captures subtle edge feature cues by calibrating channel responses to adapt to learnable edge-enhancement dominant patterns. The training scheme aims to enhance the model's generalisation and robustness. To thoroughly validate the model's performance, we constructed two high-similarity datasets. One is derived from DB1404 [2] and called HS-DB, the other is derived from Sketch-20 K [3] and called HS-S20 K. EdgeCANet attains the highest Top-1 accuracy and Top-5 accuracy on HS-DB dataset, and achieves the highest Top-1 accuracy on DB1404 dataset with 26.88 M parameters. EdgeCANet also outperforms state-of-the-art models on OBC306, Sketch-20 K, ImageNet-Sketch and HS-S20 K datasets. EdgeCANet effectively enhances recognition accuracy for highly similar handwritten characters and ancient minority-script characters, offering broader applicability for related recognition tasks. The main contributions of our work are as follows.

1. An EdgeCANet model is proposed to recognise highly similar Dongba characters with high accuracy. Different from the models with hybrid attention mapping units and spatial attention aggregation module, the core of our model is Edge Channel Attention Aggregation (ECAA) module. ECAA incorporates the Edge-Guided Enhancement Module (EGEM) and the Channel Texture Recalibration Module (CTRM). The EGEM can greatly enhance the extraction capability of critical edge details while suppressing redundant features. The CTRM recalibrates channel weights to enhance the extraction of edge information across different channels. This improves learning and discrimination capabilities in edge-sensitive regions, thereby increasing recognition accuracy for highly similar characters.
2. Two-stage fine-tuning training scheme based on exponential moving average (EMA) algorithm is proposed. Different from other methods freezing certain model parameters to restrict the update range, our two-stage misclassified sample weight redistribution fine-tuning scheme simplifies the workflow and enhance practicality. In the first training stage, the number of misclassified samples is obtained by the best model. During the second fine-tuning phase, the classification error-weighted fine-tuning strategy is applied to focus on difficult-to-classify samples. This fine-tuning scheme maintains initial high recognition accuracy, prioritises difficult-to-classify samples, and stabilises parameter updates.
3. Two high similarity datasets HS-DB and HS-S20 K are constructed to validate the performance of EdgeCANet. HS-DB dataset originates from DB1404, comprising 4414 images across 32 categories. HS-S20 K dataset originates from Sketch-20 K [3], comprising 1600 images across 20 categories. The two high similarity datasets are available at https://github.com/xy1369/high-similarity-datasets.git.

Our proposed model outperforms state-of-the-art models in recognising handwritten Dongba script and other sketch characters with similar structures such as Oracle Bone Script [4]. Besides, our work can advance the digital dissemination of minority cultures, provide support for preserving and transmitting intangible cultural heritage, and accelerate the collation and research of ancient texts from minority ethnic groups.

2 | Related Work

2.1 | Research on Dongba Character Recognition

Convolutional neural networks (CNNs) have continued to evolve rapidly and achieved highly competitive performance on large-scale image classification through innovative architectural designs [5, 6]. Deep learning techniques based on CNNs have become the mainstream approach for solving the HDCR task [7–9].

In the field of handwritten Dongba character recognition, researchers initially employed convolutional neural networks (CNNs) for single-scale feature recognition of Dongba characters [7], and subsequently expanded this to multi-scale feature classification, such as the Dongba character recognition method based on multi-scale feature fusion [9]. Recently, researchers have shifted from handcrafted features to end-to-end models to address challenges in specific scenarios. For example, DB2RNet [10] proposed a framework that integrates multi-scale restoration and recognition to handle incomplete handwritten Dongba characters. MAAN [2] employs a Multi-Attention Aggregation

Network (MAAN), which is currently the best-performing model in the field of handwritten Dongba character recognition. By combining hybrid attention maps with spatial attention mechanisms to process complex Dongba characters, this model achieved a recognition rate of 99.34% on the DB1404 dataset.

At the same time, this paper reviews research on handwritten character recognition in related handwritten character recognition tasks, with the aim of providing a more modern technical framework and methodological references for Dongba character recognition. Li et al. conducted a systematic analysis of issues such as the abundance of script variants, limited sample sizes, and poor image quality in the recognition of oracle bone script and ancient Chinese characters, and objectively summarised the effectiveness and limitations of various models currently used in these recognition tasks [11]. Meanwhile, in the field of Urdu handwritten character recognition task, KO and Poruran proposed the OCR-Nets model, which utilises variants of the pre-trained AlexNet and GoogLeNet to recognise Urdu handwritten characters through transfer learning [12]. This demonstrates that models based on convolutional neural networks (CNNs) can effectively adapt to handwritten character recognition tasks across different writing systems, providing valuable methodological references for related recognition scenarios involving limited character sets. Another study on Urdu handwritten character recognition proposed a transfer-learning-based deep CNN architecture that integrates feature representations extracted from AlexNet, GoogLeNet, and ResNet18 [13]. By concatenating features from multiple pre-trained CNNs and using fully connected layers for classification, this method improved the recognition performance of cursive Urdu characters. This study indicates that effective feature representation and feature fusion are important for handwritten character recognition across different scripts.

In the task of recognising handwritten Dongba characters, some characters are visually similar, with subtle edge features that are prone to distortion, posing a significant challenge for feature learning. While the MAAN method has achieved high recognition accuracy on the DB1404 dataset, there is an urgent need for research aimed at improving recognition accuracy for highly similar Dongba characters to improve the overall performance of handwritten Dongba character recognition tasks.

2.2 | Exponential Moving Average Algorithms

EMA is a weighted moving average algorithm commonly used in deep learning as a gradient update strategy to enhance the stability and generalisation capabilities of model training. In the field of image classification, researchers have proposed a strategy that combines EMA with a weighted gradient optimiser for tomato leaf disease recognition, which effectively improved the classification accuracy [14]. In medical image segmentation tasks, some studies employ the Residual Exponential Moving Average (R-EMA) strategy to construct a robust teacher model to stabilise pseudo-label updates and smooth teacher model weights [15]. In the field of unsupervised visual representation learning, the Momentum Contrastive Learning (MoCo) framework [16] employs the EMA mechanism to update the weights w_k of the Momentum Encoder. In semi-supervised learning, Mean Teacher maintains an exponential moving average of model weights to construct more stable teacher targets, thereby improving training stability and generalisation [17].

EMA maintains the exponential moving average of model parameters, thereby effectively suppressing parameter fluctuations caused by stochastic gradient updates during training and reducing training noise. Its exponentially decayed weighting mechanism gradually weakens the influence of historical parameters on the current model over time, thereby mitigating the instantaneous interference of abnormal updates or extreme samples on model performance metrics and preventing pronounced oscillations during training. Therefore, in this study, the EMA algorithm is introduced into the model training process for the HDCR task, aiming to impose a smoothing constraint on parameter updates and further enhance the robustness and stability of model training.

2.3 | Fine-Tuning Strategy

Researchers in computer vision have proposed various strategies to enhance model performance inspired by fine-tuning methods for large models such as Parameter-efficient fine-tuning (PEFT) [18]. In the field of handwritten character recognition, KO and Poruran proposed OCR-Nets as a transfer learning-based method for fine-tuning Urdu handwritten character recognition [12]. By modifying task-specific layers and fine-tuning the transfer network, they adapted the pre-trained AlexNet and GoogLeNet models to the target Urdu handwritten character recognition task. This work effectively demonstrated that pre-trained CNN models can provide effective feature representations for handwritten character recognition tasks, particularly when task-specific data is limited. Some scholars have introduced two-stage fine-tuning methods for addressing category imbalance tasks [19–21]. These approaches typically rely on existing pre-trained models and introduce additional training phases, often freezing certain model parameters to restrict the update range, similar to strategies used in large model fine-tuning.

In the field of target detection, Focal Loss was proposed to address the problem of extreme imbalance between foreground and background categories [22]. Label-Distribution-Aware Margin (LDAM) loss enhances classification performance and generalisation by assigning category-specific decision margins and incorporating a class-aware boundary adjustment strategy [23]. Optimising Samples After Positive Ones & AP loss (OSAP-Loss) method was introduced to tackle severe inter-class image imbalance, achieving superior results in remote sensing image retrieval tasks [24].

The above methods is not suitable for HDCR characteristics: Some categories contain a large number of samples due to the richness of writing styles, yet they remain inherently difficult to classify. This atypical category imbalance—where a large number of samples does not imply easy classification—requires a more targeted optimization mechanism. In our work, the weight reassignment fine-tuning strategy based on the number of misclassified samples is proposed.

2.4 | Data Augmentation

In visual recognition tasks, data augmentation is a key method for improving model generalisation. Given the characteristics of the datasets studied in this paper, geometric data augmentation strategies offer distinct advantages in terms of task adaptability and practical application value. In character recognition and text analysis scenarios, geometric transformations such as rotation, scaling, and cropping can effectively simulate variations in real-world writing habits and imaging conditions. Current image augmentation methods in deep learning encompass a variety of technical principles and applications, including geometric transformations, colour space operations, and Generative Adversarial Networks (GANs). Meanwhile, geometric transformation strategies have played a pivotal role in enhancing the generalisation performance of Convolutional Neural Networks (CNNs) in image recognition tasks [25]. Furthermore, the Agent Augmentation method innovatively introduces a learnable training augmentation mechanism for text recognition. By utilising an agent network approach to achieve joint optimization between the augmentation network and the recognition network, it significantly enhances the robustness and generalisation capabilities of recognition models in scenarios involving complex handwriting deformations and irregular text [26].

In recent years, data augmentation methods based on generative models have also garnered significant attention. Generative adversarial networks (GANs) and diffusion models have demonstrated strong data generation capabilities in tasks such as medical image processing and few-shot learning [27, 28]. At the same time, researchers have further explored and improved generative augmentation mechanisms [29, 30]. However, compared to geometric transformations, these methods typically require an additional generative model training process, resulting in more complex implementation workflows and higher computational costs. For the handwritten character recognition task addressed in this paper—which involves low-resolution images with distinct foreground-background contrast and relatively simple image structures—such complex generative augmentation may not necessarily yield practical benefits commensurate with their costs.

Therefore, we prioritise the use of geometric data augmentation strategies. On the one hand, geometric transformations can directly simulate natural perturbations in the orientation, scale, and position of handwritten characters, better aligning with the actual distribution characteristics of Dongba characters and related sketch-like character images, on the other hand, this strategy offers advantages such as low implementation cost, stable training processes, and ease of reproducibility. It not only better meets practical application needs but also effectively enhances sample diversity and model generalisation capabilities without introducing additional complex models. For the datasets used in this paper, HS-DB, HS-S20 K, DB1404, OBC306, Sketch-20 K, and ImageNet-Sketch, the application of efficient geometric augmentation strategies can more effectively improve the model's recognition performance in complex real-world scenarios.

3 | Method

Existing Dongba character datasets contain a certain number of characters with highly similar visual morphologies. Such characters are typically characterised by subtle edge structures and local differences that are prone to being weakened, posing a severe challenge to the learning of discriminative features by models. To address this issue, this paper proposes the EdgeCANet model and further introduces a two-stage fine-tuning scheme based on the exponential moving average strategy, aiming to improve the recognition accuracy of the model for both ordinary Dongba characters and highly similar Dongba character classes with a smaller parameter scale than the Dongba character recognition model that currently achieves the highest recognition accuracy. Unlike related models that employ hybrid attention mapping units and spatial attention aggregation modules, the proposed EdgeCANet is centred on the Edge-Channel Attention Aggregation Block. This block integrates EGEM for edge-guided enhancement and CTRM for channel texture recalibration, thereby enhancing the model's ability to represent fine-grained edge structures and texture differences. Meanwhile, unlike fine-tuning methods that restrict the update range by freezing specific model parameters, the proposed EMA-based two-stage misclassified-sample weight redistribution fine-tuning scheme simplifies the fine-tuning workflow while improving the training stability and practical applicability of the method.

3.1 | Edge Channel Aggregation Network (EdgeCANet)

Edge attention performs element-wise weighting on backbone features effectively by explicitly constructing an edge attention map within the feature space, thus enhancing edge regions while suppressing non-edge responses. The boundary-aware architectures that explicitly model shape and boundary cues have demonstrated strong capability in preserving subtle structural boundaries [31]. Edge attention has achieved significant results in the field of medical image processing [32]. Inspired by the above research, we propose EdgeCANet to accurately distinguish subtle and fragile edge cues in handwritten Dongba characters.

Edge attention refers to explicitly constructing an edge attention map within the feature space to perform element-wise weighting on backbone features, thereby enhancing edge regions while suppressing non-edge responses. Our proposed EdgeCANet is an edge channel aggregation architecture, as shown in Figure 2.

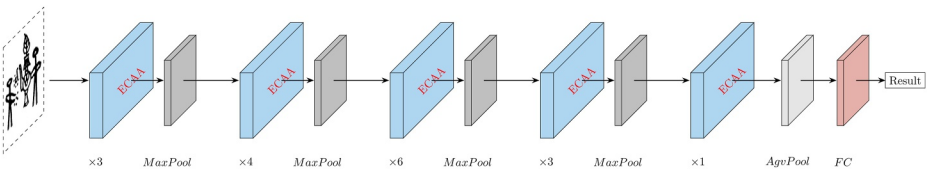


FIGURE 2 | Architecture of EdgeCANet model.

This architecture aims to leverage complementary feature information across different depth layers. Shallow layers emphasise edge and texture features, while deep layers encode global shape-related information. On one hand, a learnable edge attention mechanism is designed to effectively extract target edge features, providing clearer edge structure information for subsequent recognition. On the other hand, a Squeeze-Excitation attention module is integrated to dynamically learn dependencies between channels. This mechanism boosts the weights of important feature channels while weakening the influence of redundant or noisy channels. This complementary enhancement across both edge and channel dimensions significantly improves the model's feature representation quality.

As shown in Figure 2, the backbone network of EdgeCANet stacks ECAA via residual connections to integrate edge features and channel features. Given the pronounced foreground-background contrast in handwritten/sketch-like image inputs, max pooling is employed during downsampling to preserve salient structural responses. The input image is processed through four layers of ECAA and MaxPool, feeding the output to the final ECAA layer for further aggregation of edge feature information and prediction output. The max pooling layer following each ECAA uses a kernel size of 2 × 2 and a stride of 2. The average pooling layer employs a kernel size of 1 × 1 with a stride of 1.

3.2 | Edge-Channel Attention Aggregation Block

ECAA performs edge and channel feature extraction on the input image, fuses the edge and channel feature information, and passes it to the next layer. The EGEM and CTRM units as the core feature modelling components for ECAA, as shown in Figure 3. The EGEM performs edge-aware spatial refinement, while the CTRM executes channel recalibration. This architecture aims to preserve the representational capabilities of both branches by decoupling the channel-recalibration branch from the edge-aware spatial-refinement branch.

3.2.1 | *Edge-Detail Synergistic Attention (EDSA)*

The EDSA, as illustrated in Figure 4, consists of the following main sub-modules, which guide the model to more effectively focus on key edge regions in the image.

Foreground Attention: This module generates an attention map of the foreground, namely the target region by applying a Sigmoid activation on the input feature map through Equation (1), thereby highlighting the key regional features and suppressing background noise in the image.

$$\text{foreground}_x = x \otimes \sigma(\text{pred}) \tag{1}$$

x is the feature map output from the backbone network, $\otimes$ denotes element-wise (pixel-by-pixel) multiplication. Pred is the prediction mask, and $\sigma(\cdot)$ denotes the Sigmoid activation function. Pred is obtained by Equation (2), where k denotes the kernel size, s denotes the stride and p denotes the padding.

$$\text{Pred} = \text{Conv}_{k=1}\left(\text{ReLU}\left(\text{BN}\left(\text{Conv}(x)_{k=3,s=1,p=1}\right)\right)\right) \tag{2}$$

Learnable Edge Detector: This module uses a residual structure with depthwise separable 3 × 3 and 1 × 1 convolutions and ReLU function to extract edge features by Equation (3).

$$\begin{cases} \text{LED_feature} = x \otimes E(x) \\ E(x) = x - \text{ReLU}\left(\text{Conv}_{k=3,p=1,g=c}\left(\text{Conv}(x)_{k=1}\right)\right) \end{cases} \tag{3}$$

$x \in R^{C\times H\times W}$ is the input feature map, $E(x)$ represents the edge response map, k denotes the kernel size, p denotes the padding, and g denotes the number of groups. R denotes the set of real numbers, indicating that each value in the input feature map x is a real number. C represents the number of channels in the image. H denotes the height of the feature map. W denotes the width of the feature map.

High-Frequency Feature: This module is designed to enhance the representation capability of detail-sensitive regions. The detailed edge feature map is obtained from the Laplacian

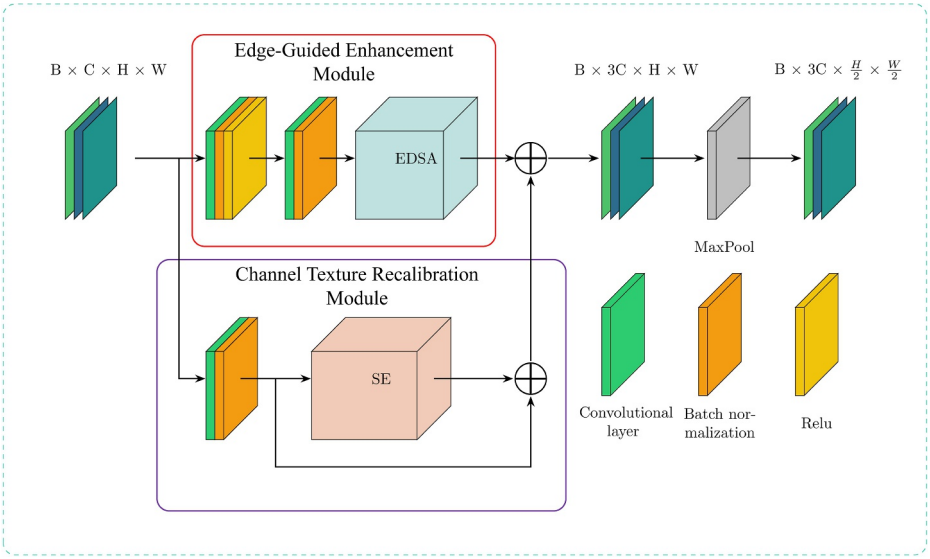


FIGURE 3 | ECAA block of EdgeCANet network model.

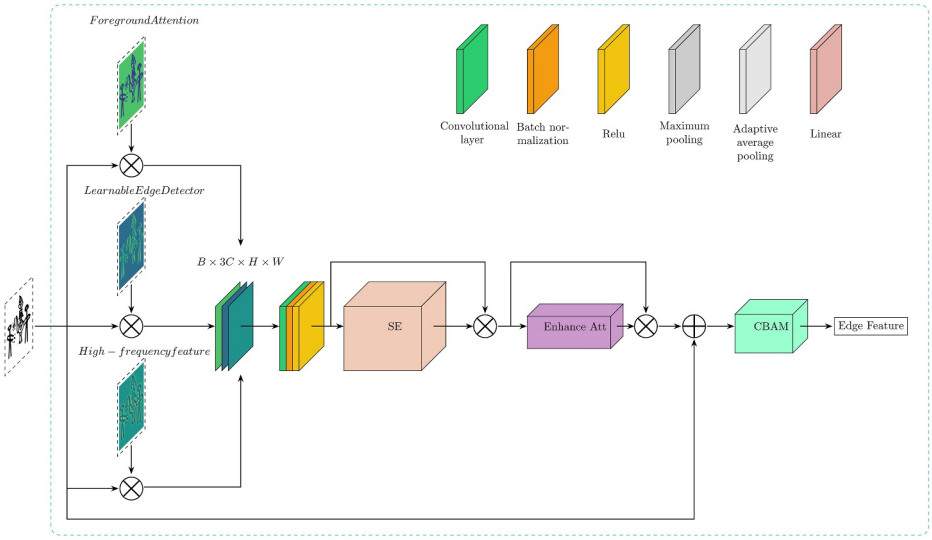


FIGURE 4 | Details of the Edge-Detail Synergistic Attention. ⊗ denotes element-wise multiplication, ⊕ denotes element-wise addition. CBAM, Convolutional Block Attention Module; Enhance Att, Enhance Attention; SE, Squeeze-Excitation Attention.

pyramid and upsampled to match the spatial resolution of the input feature map through Equation (4).

$$\begin{cases} I_{k+1} = \downarrow_2(\mathrm{G} * I_k) \\ L_k = I_k - u(I_{k+1}) \end{cases} \tag{4}$$

where L_k denotes the k_{th}-level Laplacian component, I_k and I_{k+1} denote adjacent levels of the Gaussian pyramid respectively, $\downarrow_2(\cdot)$ represents double downsampling of the smoothed image, $\mathrm{G} * I_k$ denotes Gaussian convolution applied to I_k, and $u(\cdot)$ signifies upsampling followed by Gaussian convolution to reconstruct the feature information of I_k.

After the above computations to preserve the target edge detail features within L_k at this scale, the selected edge feature L_k is then resized using bilinear interpolation shown in Equation (5).

$$\text{edge_input} = \text{Interp}_{\text{bilinear}}(L_k, \text{size} = H \times W) \tag{5}$$

where H and W represent the spatial dimensions of x. The edge-guided input features are ultimately computed through element-wise multiplication shown in Equation (6),

$$\text{high-frequency_feature} = x \odot \text{edge_input} \tag{6}$$

where the symbol ⊙ denotes the Hadamard product. This operation aims to enhance the network's ability to distinguish fine-scale structural differences by amplifying the response intensity in edge regions while suppressing it in non-edge areas.

Squeeze-Excitation Attention (SE): This module is designed for channel attention adjustment to yield the channel gate vector based on the SE mechanism [33]. After concatenating the Foreground Attention, Learnable Edge Detector and High-Frequency Feature components, the fused feature F_{SE} is obtained by Equation (7),

$$F_{\text{SE}} = F \odot \text{SE}(F) \in R^{B \times C \times H \times W} \tag{7}$$

where F represents fused features, SE(F) represents Squeeze-Excitation (SE) processing for channel attention adjustment, and B represents the batch size during a single training or inference process. F is obtained by Equation (8),

$$F = \text{ReLU}(\text{BN}(\text{Conv}_{1\times1}(X_{\text{cat}}))) \in R^{B \times C \times H \times W} \tag{8}$$

where X_{cat} denotes the fused feature map obtained by concatenating three distinct feature branches and is defined in Equation (9),

$$X_{\text{cat}} = [\text{foreground}_x;\ \text{LED_feature};\ \text{high-frequency_feature}] \in R^{B \times 3C \times H \times W} \tag{9}$$

The feature maps from the three branches are all of size B × C × H × W. These features are concatenated along the channel dimension to form $X_{\text{cat}} \in R^{B \times 3C \times H \times W}$.

A 3 × 3 convolution, batch normalisation, and ReLU activation are then applied to fuse the concatenated representations and reduce their dimension back to $B \times C \times H \times W$, ensuring compatibility with subsequent SE attention, enhanced attention, residual connections, and CBAM optimization.

It should be noted that the SE described here does not constitute the entire definition of EDSA, rather, it serves as an auxiliary channel re-calibration mechanism for the edge enhancement process. Its function is to adaptively increase the channel weights associated with edge detection based on the fused statistical features. This differs from the SE function within CTRM and should not be conflated with it.

Enhance Attention: This module comprises three 3 × 3 convolutional branches with different dilation rates (1, 3, and 5, respectively) to extract contextual information under a multi-scale field of view. The outputs are fused using a 1 × 1 convolution and activated by Sigmoid function to generate the

final attention map. The computational process is shown as Equation (10),

$$A(F_{SE}) = \sigma\left(\text{Conv}_{k=1}\left(\left[\text{Conv}_{k=3,p=1,d=1}x, \text{Conv}_{k=3,p=3,d=3}x, \text{Conv}_{k=3,p=5,d=5}x\right]\right)\right) \quad (10)$$

where d denotes a convolution operation with dilation d. $[\cdot]$ denotes channel-level concatenation, and $\text{Conv}_{k=1}$ is a 1×1 convolutional fusion operation. This module enhances the model's ability to focus on critical edge feature regions via multi-scale receptive fields.

Convolutional Block Attention Module (CBAM): Convolutional Block Attention Module (CBAM) is designed for the final refinement. CBAM first computes channel attention on the input features, followed by spatial attention, with the aim of highlighting important channel and spatial edge features to obtain the final output EdgeFeature, as shown in Equation (11). This ensures that both global context and local details are fully encoded. This global-local collaborative representation strategy has also been proven effective in recent scalable multi-task contrastive learning frameworks [34].

$$\text{Edge Feature} = \text{CBAM}(U) \quad (11)$$

U represents the final output vector after processing through the complete attention mechanism and applying residual connections. It is gotten by Equation (12),

$$U = F_{EA} + x \quad (12)$$

where F_{EA} represents the enhanced attention feature, obtained by element-wise multiplication of the channel-wise attention-adjusted feature F_{SE} with the augmented attention map $A(F_{SE})$, which is achieved by Equation (13),

$$F_{EA} = F_{SE} \odot A(F_{SE}) \quad (13)$$

3.2.2 | *Channel Texture Recalibration Module (CTRM)*

The Channel Texture Recalibration Module (CTRM) runs in parallel with EGEM within the ECAA module. CTRM is designed to enhance channel features related to edge textures while suppressing irrelevant channels, as shown in Figure 5, where $\otimes$ denotes element-wise multiplication. The basic workflow of CTRM is as follows: first, preprocessing of the input image features is performed using convolution and standard normalisation, then, the SE attention mechanism is used to perform global channel recalibration on the input features, thereby achieving stable mapping and channel integration. Its output can be expressed as Equation (14), where F_{CTRM} denotes the final feature map obtained by the Channel Texture Recalibration Module.

$$F_{CTFE} = \text{SE}(\text{BN}(\text{Conv}_{1\times1}(x))) \quad (14)$$

It should be noted that while the SE used here shares the same basic code framework as the SE within the Edge-Detail Synergistic Attention (EDSA) mechanism in the EGEM module, their functional roles differ. The SE in CTRM operates on the entire channel texture branch and plays a central role in global channel recalibration, whereas the SE in EDSA operates on local features after edge fusion and serves a supplementary role in channel weighting. Therefore, although both share the same attention primitive, they operate at different module levels.

3.3 | Two-Stage Misclassified Sample Weight Redistribution Fine-Tuning Training Based on Exponential Moving Average

The two-stage weight redistribution fine-tuning training scheme for recognising handwritten Dongba script and sketch images introduces EMA to stabilise parameter updates. The second-stage reweighting strategy redistributes misclassified sample weights to intensify focus on challenging categories without freezing parameters. After first-stage convergence, the second stage maintains initial recognition accuracy while prioritising features from difficult-to-classify samples.

3.3.1 | *Exponential Moving Average Strategy*

Although previous studies have explored the application of the EMA algorithm in image recognition tasks [14–17], it is the first time that EMA is used in stabilising the training process in handwritten Dongba script image recognition tasks. In the experimental section, we combined it with fine-tuning strategies and conducted experimental analysis.

During training, the EMA in EdgeCANet is updated according to Equation (15),

$$\text{EMA}_{\text{new}} = \text{decay} \times \text{EMA}_{\text{old}} + (1 - \text{decay}) \times \text{EgeCANet}_{\text{current}} \quad (15)$$

where EMA_{new} denotes the updated EMA parameters, EMA_{old} represents the EMA parameters from the previous step, $\text{EdgeCANet}_{\text{current}}$ refers to the model parameters at the current training step. The decay factor for parameter update is set to 0.999.

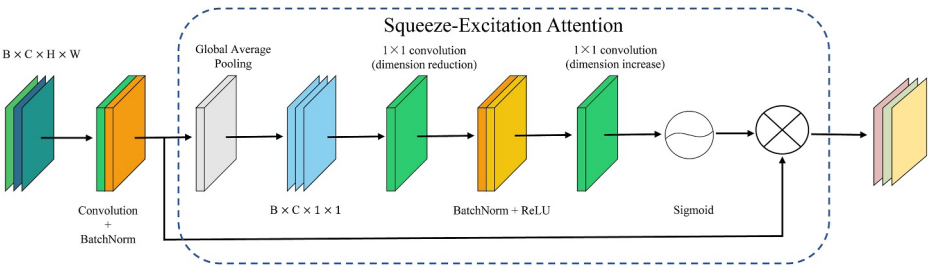


FIGURE 5 | Block diagram of the channel texture recalibration module.

Through this strategy, the EMA model continuously maintains the weighted historical average of its parameters during training, thereby providing more stable and reliable parameter estimates during evaluation and fine-tuning phases. This approach effectively smooths out sharp fluctuations in parameters, enhances training stability, and accelerates the model convergence process.

During training, we update the EMA parameters with a decay factor of 0.999 at each iteration. Unless otherwise specified, predictions on the validation and test sets are generated based on EMA weights. During fine-tuning in the second phase, the EMA state from the end of the first phase is restored.

3.3.2 | Fine-Tuning Approach Based on Weighting Misclassified Samples

Existing fine-tuning strategies primarily focus on addressing class imbalance by leveraging transfer learning or adjusting the weights of large models. Related research has also shown that modulating the optimization signal helps balance the convergence dynamics and improves the classification performance of the model [35]. However, to date, few studies have explored applying fine-tuning strategies to enhance model classification accuracy in HDCR tasks. Therefore, we propose a novel fine-tuning strategy based on the weights of misclassified samples. This strategy eliminates the need to freeze model parameters during fine-tuning, simplifies the workflow, and enhances practicality, as shown in Figure 6. Our strategy demonstrates robust generalisation capabilities. It effectively handles datasets with atypical distributions, such as DB1404, in which samples are unevenly distributed across categories and contain pairs of highly similar categories. It also significantly improves model performance even in conventional scenarios like the ImageNet-Sketch dataset.

Our strategy simplifies the fine-tuning process by employing a full-model parameter retraining approach. By shifting the fine-tuning objective from the traditional goal of reducing training resource consumption to improving target recognition accuracy, we have achieved an improvement in model performance.

In the first training stage, EdgeCANet employs a uniform sample weighting strategy, enabling the model to treat all samples equally and fully learn the overall feature distribution of the dataset. After the model converges in the first phase, for samples misclassified on the validation set, we combine cross-entropy loss with label smoothing techniques—drawing on insights from modern methods such as LS + [36] and confidence-based CALS [37]—to stabilise the optimization process and improve calibration for difficult classes. Specific weighting rules and implementation details are provided in Section 5.1. This fine-tuning strategy is designed to enhance the EdgeCANet model's ability to adaptively shift its learning focus toward challenging samples, thereby improving recognition performance for highly similar categories.

3.4 | Data Augmentation Strategies for Handwritten Character

To enhance EdgeCANet's generalisation capability and robustness in recognising sketch-like images, we designed a specialised data augmentation strategy. This approach combines geometric transformations with normalisation. These operations expand the diversity of training samples, simulating various handwritten scenarios.

Before normalisation, we compute the global mean μ and standard deviation σ of the dataset across all training images. Let N denote the number of images, H and W denote their height and width respectively, and C represent the number of channels. $I_{n,h,w,c} \in [0,1]$ represents the pixel intensity at row h, column w, and channel c for sample n. The mean and standard deviation are calculated by Equations (16) and (17) respectively.

$$\mu = \frac{1}{N \times H \times W \times C} \sum_{n=1}^{N} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} I_{n,h,w,c} \tag{16}$$

$$\sigma = \sqrt{\frac{1}{N \times H \times W \times C} \sum_{n=1}^{N} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} \left(I_{n,h,w,c} - \mu\right)^2} \tag{17}$$

For DB1404 dataset, $\mu = 0.9467$ and $\sigma = 0.1892$.

Each image is rotated by an angle θ, where $\theta \sim U(-15^\circ, 15^\circ)$ about its center. This augmentation simulates natural writing inclination, preventing the model from overfitting to a single orientation and promoting rotational invariance.

We randomly select an image coverage ratio s $\sim U(0.9, 1.0)$ and resize it to 64×64 pixels. This operation introduces minor translation and scaling, making the network robust to off-centre or slightly scaled characters and encouraging better positioning tolerance.

Finally, the intensity of each pixel in the n_{th} image is normalised by Equation (18),

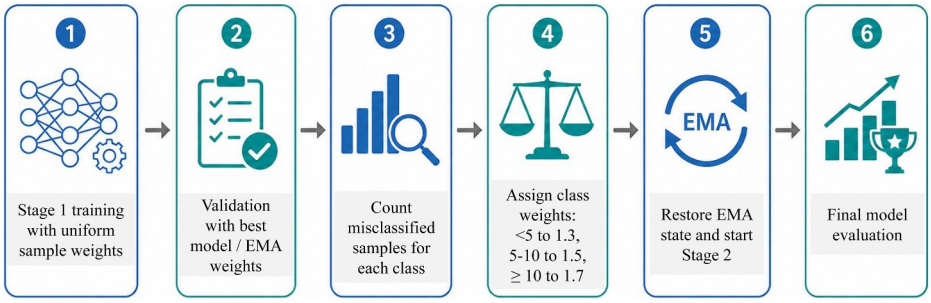


FIGURE 6 | Workflow of the two-stage fine-tuning strategy.

$$I'_{n,h,w,c} = \frac{I(n,h,w,c) - \mu}{\sigma} \tag{18}$$

where $I_{(n,h,w,c)}$ denotes the pixel intensity of image n at position (h, w, c), μ and σ represent the global mean and standard deviation respectively, and $I'_{n,h,w,c}$ is the normalised value used as network input.

Our proposed data augmentation technique diversifies the training distribution while preserving transformation fidelity, enabling EdgeCANet to achieve stronger generalisation capabilities on unseen handwriting variants. The data augmentation strategy is applied to all the datasets in our work.

4 | Construction of Highly Similar Image Dataset: HS-DB and HS-S20 K

We constructed two novel high-similarity image datasets to validate the performance of our model on recognising high similarity images. One is the high similarity Dongba script handwritten image dataset and is called HS-DB. The other is the high similarity human sketch object image dataset and called HS-S20 K. The construction details are as follows.

4.1 | Dataset Collection

Despite the abundance of existing handwritten Dongba script and sketch datasets, specialised datasets tailored for high similarity scenarios remain underdeveloped. To address this gap, we extracted and constructed high-similarity sample collections HS-DB and HS-S20 K from two authoritative datasets. HS-DB dataset originates from DB1404, comprising 4414 images across 32 categories. HS-S20 K dataset originates from Sketch-20 K [3], comprising 1600 images across 20 categories. Sample images from HS-DB and HS-S20 K are shown in Figures 7 and 8 respectively.

4.2 | Quality Control and Annotations

To ensure the high-similarity characteristics and data quality of the HS-DB and HS-S20 K samples, we manually screened all images, primarily focusing on removing dissimilar samples and selecting highly similar samples. When constructing the HS-DB dataset, we fully considered the uniqueness of Dongba script. Some characters exhibit 2 to 10 distinct writing styles, but only one style meets the high similarity requirement. Therefore, during HS-DB construction, we prioritised excluding style categories with significant morphological differences, retaining



(a)The picture on the left means dead, and the right means sleeping.



(b)The picture on the left means bite, and the right means spit.

FIGURE 7 | Examples of HS-DB dataset.



(a)The picture on the left means flying bird, and the picture on the right means seagull.



(b)The picture on the left means suv, and the picture on the right means van.

FIGURE 8 | Examples of HS-S20 K dataset.

only samples with highly consistent stroke structures to ensure strict uniformity in similarity characteristics.

For both HS-DB and HS-S20 K, we uniformly adopt core morphological consistency as the basis for determining high similarity. However, due to the richer category samples in HS-S20 K, the differences between some of its categories are slightly higher. Therefore, we differentiate the specific determination criteria for the two as follows.

HS-DB: If a sample differs only in minor stroke positions or lengths at 1–2 locations while maintaining identical core structure, it is deemed highly similar.

HS-S20 K: Minor variations in multiple locations are permissible, provided that they do not affect the overall character structure, stroke topology, or semantic consistency. Such characters are still considered highly similar.

5 | Experiments

5.1 | Experimental Setup

The proposed EdgeCANet is implemented using PyTorch and trained across HS-DB, HS-S20 K, DB1404, OBC306, Sketch-20 K, and ImageNet-Sketch datasets with a two-stage misclassified sample weight redistribution fine-tuning strategy. We partitioned the dataset into training, validation, and test sets at a ratio of 7:1:2. All experiments employ the AdamW optimiser and cross-entropy loss function (CELoss).

Before entering the second-stage fine-tuning phase, the number of misclassified samples per category on the validation set is obtained by using the best model from the first stage. Then specific weights are assigned to challenging categories according to the following rules.

If the number of misclassifications is fewer than 5, we assign 1.3 to the weight.

If misclassifications are between 5 and 10, including 5, we assign 1.5 to the weight.

If the number of misclassifications is greater than or equal to 10, including 10, we assign 1.7 to the weight.

Label smoothing (ϵ value explained below) [36] is applied as a regularisation technique to mitigate potential overfitting issues caused by excessive category weights in the second stage. To ensure fair comparison, all baseline models were trained for the same total number of epochs across all datasets as the two-stage EdgeCANet.

Baseline models: MAAN [2], the state-of-the-art models in HDCR tasks. ResNet18 [38], ResNet34 [38], ResNet50 [38], DenseNet169 [39], ResNeXt50 [40], ShuffleNetV2 [41], EfficientNetB0 [42], ResNeSt50 [43], and RepVGG [44], which are several representative backbone architectures and competitive baselines for visual recognition.

5.2 | Performance of EdgeCANet on Highly Similar Dongba Character Dataset HS-DB

To validate the effectiveness and superiority of the proposed EdgeCANet network, we conducted extensive comparative experiments on the HS-DB dataset by comparing EdgeCANet with state-of-the-art models in the field of image recognition and handwritten Dongba character recognition. Batch size is 80. The number of training epochs is set to 120. The learning rate is 0.001 in the first stage, and learning rate is 0.001 in the second stage. Label smoothing parameter ϵ is 0.1.

5.2.1 | Superiority Studies on HS-DB Dataset

Due to the high similarity of Dongba characters, which often appear in visually similar pairs, even when the correct category is not ranked first, it often still appears among the Top-5 predictions. Therefore, we adopt Top-1 and Top-5 accuracy as primary evaluation metrics, while measuring computational efficiency through model parameter scale. Given that HS-DB dataset is a small-sample dataset, the differences in inference time among models during the testing phase are negligible and do not reflect substantive performance variations. Consequently, inference speed is not considered a key evaluation metric in this experiment. Detailed experimental results are presented in Table 1.

It can be seen that our method achieves superior performance compared to all other SOTA models in terms of Top-1 recognition rate for high-similarity handwritten Dongba character recognition tasks from Table 1.

EdgeCANet surpasses MAAN [2] model by 1.80% in Top-1 accuracy while reducing parameters by 42.11%. EdgeCANet achieves an accuracy improvement of 28.89% over EfficientNetB0 [42]. The results confirm that EdgeCANet enhances edge detail representation through an edge attention mechanism, enabling the model to maintain significant discriminative power even under conditions of extremely high intraclass similarity.

TABLE 1 | Test results of EdgeCANet compared with other competitive models on HS-DB dataset, with the best metrics highlighted in bold.

Models	Top-1	Top-5	Params	From
ResNet18 [38]	91.64%	99.33%	11.19 M	CVPR
ResNet34 [38]	91.64%	99.33%	21.29 M	CVPR
ResNet50 [38]	89.63%	98.66%	23.57 M	CVPR
DenseNet169 [39]	93.53%	99.55%	12.53 M	CVPR
ResNext50 [40]	90.08%	99.00%	23.04 M	CVPR
ShuffleNetV2 [41]	75.81%	97.21%	**2.51 M**	ECCV
EfficientNetB0 [42]	70.01%	96.10%	4.05 M	ICML
ResNest50 [43]	94.87%	99.67%	25.50 M	arXiv
RepVGG [44]	83.72%	98.77%	7.87 M	CVPR
MAAN [2]	97.10%	99.55%	45.23 M	ESWA
EdgeCANet (ours)	**98.90%**	**100.00%**	26.18 M	—

In terms of Top-5 accuracy, EdgeCANet model achieves a perfect recognition rate of 100%, setting a new performance benchmark. These results fully validate EdgeCANet's dual advantages in accuracy and efficiency, as well as its outstanding robustness and practical value in recognising highly similar handwritten Dongba characters.

5.2.2 | *Ablation Studies on HS-DB Dataset*

We conducted comprehensive ablation experiments on the HS-DB dataset to analyse the effectiveness of EdgeCANet's various components, including CTRM, EGEM, EMA, data augmentation, and fine-tuning. Table 2 presents the experimental results, using Top-1 accuracy as the evaluation metric.

As shown in Table 2, the full model achieved a maximum Top-1 accuracy of 98.90%. Removing the two-stage fine-tuning strategy reduces the Top-1 accuracy to 98.05%, indicating that reweighting misclassified samples helps the model focus on difficult classes with subtle inter-class differences. Removing all data augmentation further reduces the accuracy to 97.80%, while removing rotation and cropping/scaling separately reduces the accuracy to 97.90% and 98.19%, respectively. These results indicate that geometric transformations are particularly important for enhancing sample diversity and generalisation ability on the HS-DB.

The results of ablation experiments on the various modules of EdgeCANet further demonstrate that EGEM is the dominant component in the task of recognising highly similar characters. Removing EGEM causes the Top-1 accuracy to drop from 98.90% to 98.00%, whereas removing CTRM causes only a slight decrease to 98.79%. This indicates that EGEM plays a crucial role in recognising similar characters, while CTRM provides complementary channel-level recalibration. Therefore, the EGEM branch is more critical for distinguishing highly similar Dongba characters, while the channel recalibration branch further enhances feature representations in a complementary manner.

Furthermore, removing EMA reduces the Top-1 accuracy to 98.36%, and during our experiments, we observed that the model using EMA exhibited a slow-start phenomenon. To further investigate the performance of the EdgeCANet model with the EMA strategy on the HS-DB dataset, we compared the curves of Top-1 accuracy versus training epochs under settings with EMA enabled (w/EMA) and EMA disabled (w/o EMA), as shown in Figure 9. The experiments used the same five random

TABLE 2 | Ablation study results of EdgeCANet model on HS-DB dataset.

			Data augmentation			
CTRM	**EGEM**	**EMA**	**Rotation**	**Crop + Resize**	**Fine-tuning**	**Top-1**
✓	✓	✓	✓	✓	✓	98.90%
✓	✓	✓	✓		✓	98.19%
✓	✓	✓		✓	✓	97.90%
✓	✓	✓	✓	✓		98.05%
✓	✓	✓			✓	97.80%
✓	✓		✓	✓	✓	98.36%
✓		✓	✓	✓	✓	98.00%
	✓	✓	✓	✓	✓	98.79%

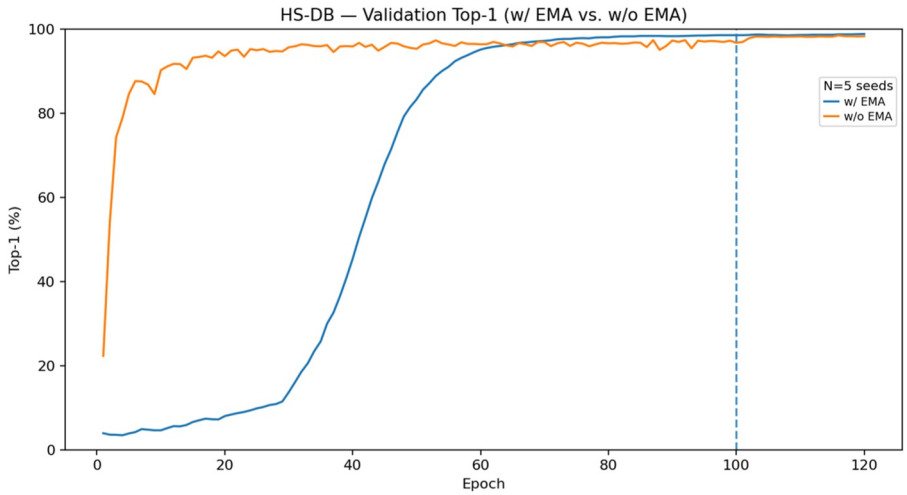


FIGURE 9 | Comparison of Top-1 accuracy on the validation set with and without EMA enabled based on 5 random seeds, the dashed line indicating the stage boundary at epoch 100.

seeds, and the results were averaged. The average sample accuracy $\bar{y}$ corresponding to the training epoch was calculated using Equation (19), where x_i denotes the Top-1 average recognition accuracy achieved by the M seeds in that epoch, specifically y_1, ... y_M.

$$\bar{y} = \frac{1}{M} \sum_{i=1}^{M} y_i \tag{19}$$

As shown in Figure 9, the EMA curve rises gradually from epoch 0 to 35, then stabilises around 99% and remains at that peak level. In contrast, the non-EMA curve surges rapidly to approximately 90% within the first 5 to 10 epochs, but its accuracy fluctuates significantly. This indicates that EMA primarily contributes to training stability rather than short-term improvements in accuracy. Overall, the ablation experiment results for HS-DB indicate that fine-tuning, geometric augmentation, and EGEM are the primary factors in high-similarity few-shot recognition tasks, while EMA and CTRM provide stability and achieve complementary feature optimization.

TABLE 3 | Test results of EdgeCANet compared with other competitive models on HS-S20 K dataset, with the best metrics highlighted in bold.

Models	Top-1	Top-5	Params	From
ResNet18 [38]	53.75%	90.93%	11.18 M	CVPR
ResNet34 [38]	55.94%	89.68%	21.29 M	CVPR
ResNet50 [38]	53.44%	88.12%	23.55 M	CVPR
DenseNet169 [39]	58.12%	92.12%	12.51 M	CVPR
ResNext50 [40]	49.69%	87.18%	23.02 M	CVPR
ShuffleNetV2 [41]	50.31%	90.62%	**2.50 M**	ECCV
EfficientNetB0 [42]	49.38%	85.31%	4.03 M	ICML
ResNest50 [43]	59.69%	91.37%	25.47 M	arXiv
RepVGG [44]	51.56%	89.68%	7.85 M	CVPR
EdgeCANet (ours)	**63.44%**	**93.43%**	26.17 M	—

5.3 | Performance of EdgeCANet on Highly Similar Human Sketch Object Dataset HS-S20 K

Experiments were conducted on HS-S20 K dataset to demonstrate the robustness and generalisation capabilities of our proposed EdgeCANet. The hyperparameter settings for EdgeCANet model are detailed as follows. Batch size is 32. The model was trained for 120 epochs. The learning rate is 0.001 in the first stage, and learning rate is 0.001 in the second stage. Label smoothing parameter ϵ is 0.1.

From Table 3, it can be seen that our EdgeCANet model achieves the best performance in both Top-1 and Top-5 accuracy, reaching 63.44% and 93.43% respectively. Although lightweight networks ShuffleNetV2 and EfficientNetB0 achieve reasonable recognition performance with fewer parameters, their overall recognition accuracy remains relatively low compared with EdgeCANet model. Deep residual networks ResNet50, ResNet34, and DenseNet169 demonstrate relatively superior Top-1 accuracy, but lower than that of EdgeCANet.

From Figure 10, it can be seen that EdgeCANet's attention is more concentrated on positions with subtle differences. The results further demonstrate that EdgeCANet can enhance recognition accuracy for high-similarity categories although EdgeCANet's parameter count (26.17 M) is slightly higher than some mainstream networks.

5.4 | Performance of EdgeCANet Compared With SOTA Models on Dongba Characters DB1404

To validate the effectiveness and superiority of EdgeCANet, we conducted comparative experiments between EdgeCANet and other state-of-the-art Dongba script recognition and image recognition models on DB1404 dataset. DB1404 is the publicly available handwritten Dongba script dataset. It comprises 1404 categories totaling 445,273 images. These experiments were designed to evaluate EdgeCANet's performance in recognising ordinary handwritten Dongba scripts. The batch size is 180 and

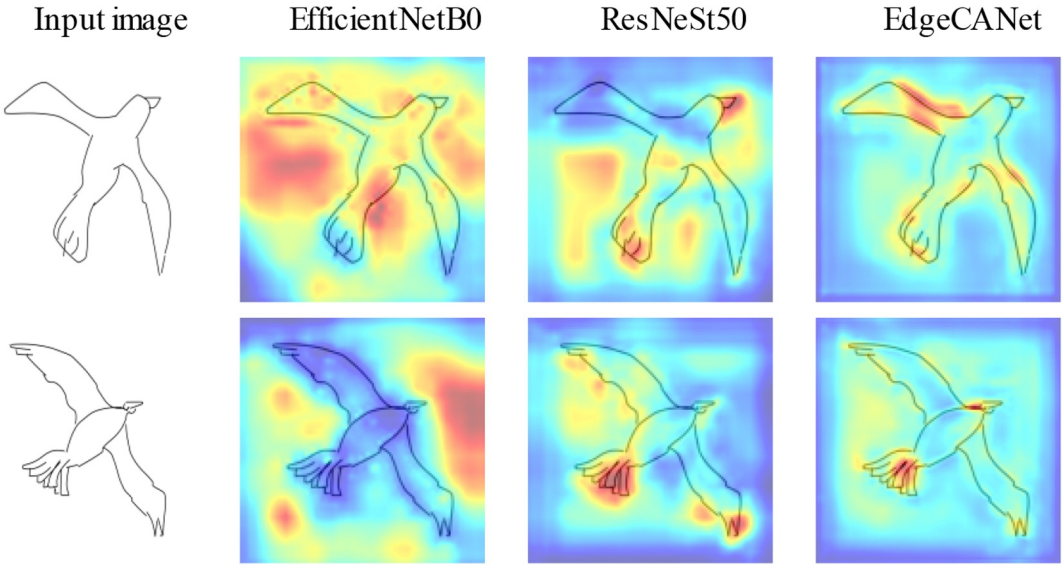


FIGURE 10 | Grad-CAM visualisation results for HS-S20 K. The first row's image category is flying bird. The second row's image category is seagull.

the number of training epochs is set to 60. In the first stage, gradient descent learning rate is 0.001 when the epoch number increases from 1 to 14, the rate is 0.0003 when the epoch number increases from 15 to 28, and the learning rate is 0.0001 when the epoch number increases from 29 to 40. In the second stage, the learning rate is 0.0005. The label smoothing coefficient ϵ is 0.15.

5.4.1 | *Superiority Studies of EdgeCANet on DB1404*

To validate the superiority of the EdgeCANet model, we conducted a systematic comparison between EdgeCANet and other SOTA models on DB1404 dataset. Evaluation metrics included Top-1 accuracy, model parameter size, and inference time.

Table 4 clearly illustrates that EdgeCANet achieves the highest Top-1 accuracy of 99.51%. EdgeCANet exhibits higher accuracy compared to models with similar or even larger parameter scales such as ResNet50 [38], ResNeXt50 [40], ResNeSt50 [43] and MAAN [2].

Specifically, EdgeCANet achieves a slight 0.17% improvement in Top-1 accuracy while reducing the number of parameters by 42.39% compared to the SOTA MAAN model which employs the hybrid attention mechanism. Furthermore, EdgeCANet achieves a 0.91% accuracy gain while reducing inference time by 36.28% compared to the architecturally mature ResNet50 [38]. This demonstrates that EdgeCANet significantly enhances sensitive feature extraction while maintaining structural efficiency and computational friendliness. This stems from the edge attention mechanism introduced in EdgeCANet, which focuses attention on character boundaries and local geometric variations, and effectively capture the most critical distinguishing features between lines.

TABLE 4 | Test results of EdgeCANet and other competitive models on DB1404 dataset.

Models	Top-1	Params	Infer(s)	From
ResNet18 [38]	98.75%	11.68 M	398	CVPR
ResNet50 [38]	98.60%	25.53 M	419	CVPR
DenseNet169 [39]	98.97%	14.82 M	471	CVPR
ResNext50 [40]	98.74%	24.99 M	428	CVPR
ShuffleNetV2 [41]	97.15%	**3.5 M**	399	ECCV
EfficientNetB0 [42]	98.89%	5.26 M	407	ICML
ResNest50 [43]	98.88%	27.5 M	470	arXiv
RepVGG [44]	98.25%	8.3 M	400	CVPR
MAAN [2]	99.34%	46.66 M	498	ESWA
EdgeCANet (ours)	**99.51%**	26.88 M	**267**	—

Note: The best metrics are shown in bold. Infer(s) denotes the inference time for 89,617 images.

5.4.2 | *Ablation Studies of EdgeCANet on DB1404*

We further conducted ablation experiments on the DB1404 dataset to evaluate whether the components validated on HS-DB could still deliver consistent performance improvements on a large-scale handwritten Dongba character dataset, as shown in Table 5.

As shown in Table 5, the full EdgeCANet achieved a top-1 accuracy of 99.51%. Among the training-related settings, removing fine-tuning resulted in the most significant performance drop, with accuracy falling to 99.32%. After removing EGEM, data augmentation, EMA, and CTRM, the accuracy dropped to 99.35%, 99.39%, 99.41%, and 99.48%, respectively. Compared to HS-DB, the performance drop on DB1404 was smaller, indicating that the large-scale data distribution reduces sensitivity to individual components. Nevertheless, each component still provides a positive and stable contribution.

The results of data augmentation further confirm the effectiveness of the geometric augmentation strategy. Without augmentation, the Top-1 accuracy was 99.39%. Rotation alone increased the accuracy to 99.46%, while cropping and scaling alone increased it to 99.42%; the combination of the two achieved the best result of 99.51%. This indicates that rotation and cropping-scaling provide complementary benefits by modelling variations in the orientation, translation, and scale of handwritten Dongba characters. Consequently, geometric augmentation not only facilitates few-shot generalisation on the HS-DB but also delivers consistent performance improvements on the large-scale DB1404 dataset.

At the same time, we further compared the impact of different EMA decay coefficients on the Top-1 accuracy, as shown in Table 6. The best Top-1 accuracy is achieved when the decay coefficient is set to 0.999. Although the differences between

TABLE 5 | Results of ablation studies of EdgeCANet model on DB1404 dataset.

			Data augmentation			
CTRM	EGEM	EMA	Rotation	Crop + Resize	Fine-tuning	Top-1
✓	✓	✓	✓	✓	✓	99.51%
✓	✓	✓	✓		✓	99.46%
✓	✓	✓		✓	✓	99.42%
✓	✓	✓			✓	99.39%
✓	✓	✓	✓	✓		99.32%
✓	✓		✓	✓	✓	99.41%
✓		✓	✓	✓	✓	99.35%
	✓	✓	✓	✓	✓	99.48%

0.95, 0.97, 0.99, and 0.999 are relatively small, the 0.999 setting consistently yields the highest accuracy, indicating that slower EMA updates are more suitable for stabilising the parameter trajectory in this task. Overall, the ablation study on DB1404 demonstrates that the proposed fine-tuning scheme, EGEM, CTRM, EMA, and geometric enhancement techniques are not only effective for recognising highly similar small-scale samples but also deliver stable performance improvements in large-scale handwritten Dongba character recognition tasks.

5.5 | Visualisation With Grad-CAM

5.5.1 | *Grad-CAM Heatmaps for Each Layer of EdgeCANet*

We extracted one pair of highly similar images from each of our self-built high-similarity datasets, HS-DB and HS-S20 K, to demonstrate the differences between attention patterns at different network depths in EdgeCANet. In Grad-CAM heatmaps [45], highly activated regions appear red and yellow, while low-activation areas are blue or transparent. Table 7 shows the differences in attention regions across various network depths in EdgeCANet. Here, Layer_1 denotes the Grad-CAM output from the first layer, Layer_3 represents the output from the first three layers, and Layer_all indicates the Grad-CAM output from the entire network.

Taking the "flying bird" category in Table 7 as an example, we can clearly observe that the Grad-CAM heatmap of the shallow Layer_1 exhibits a delicate, continuous contour pattern with minimal internal filling, highlighting only the edge contours of the targets in the image. The Grad-CAM of the intermediate Layer_3 still primarily responds to the edge features of the target, but the response bandwidth has significantly widened. Furthermore, minor hotspots begin to appear within local edge regions, indicating that the network has started extracting feature information from these areas. Meanwhile, the Grad-CAM activation regions in the deep Layer_all have covered the entire contour of the target, with distinct responses observed in the regions corresponding to the target's primary edge features.

By comparing Grad-CAM heatmaps across different layers of EdgeCANet, we verified that EdgeCANet's shallow-layer features capture local edge details, while its deep-layer features capture overall morphology and semantic information.

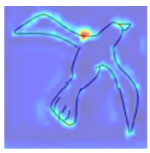
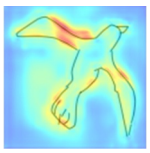
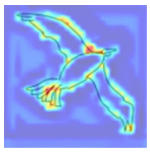
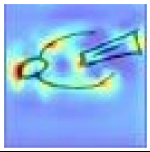
5.5.2 | *Comparison With the Optimal Model: Grad-CAM Heatmaps*

We employ Grad-CAM [45] to visualise the attention distribution patterns of ShuffleNetV2, DenseNet169, MAAN, and

TABLE 6 | Effect of decay coefficients for different EMAs on Top-1 accuracy on the DB1404 dataset.

			Data augmentation			
CTRM	**EGEM**	**Fine-tuning**	**Rotation**	**Crop + Resize**	**EMA**	**Top-1**
✓	✓	✓	✓	✓	0.999	99.51%
✓	✓	✓	✓	✓	0.99	99.49%
✓	✓	✓	✓	✓	0.97	99.48%
✓	✓	✓	✓	✓	0.95	99.47%

TABLE 7 | Layer-wise Grad-CAM visualizations of EdgeCANet on HS-DB and HS-S20 K.

		EdgeCANet		
Dataset	**Category**	**Layer_1**	**Layer_3**	**Layer_all**
HS-S20 K	flying bird			
	seagull			
HS-DB	bite			
	spit			

EdgeCANet on the highly similar HS-DB dataset, as shown in Figure 11.

From Figure 11, it can be seen that the Grad-CAM masks generated by EdgeCANet cover the target object regions much better than those of MAAN, ShuffleNetV2, and DenseNet169 for highly similar Dongba characters. EdgeCANet more accurately and efficiently focuses on regions with the most intense red, avoiding ineffective attention to blank areas of the image seen in other models.

This primarily stems from the fundamental difference in the attention mechanisms of EdgeCANet. When dealing with highly similar Dongba characters, EdgeCANet is able to accurately attend to the detailed features that distinguish between characters, rather than adopting a generalised focus over broad areas. This demonstrates a more fine-grained discrimination pattern.

Combining the above quantitative and qualitative experimental results, it can be concluded that our proposed EdgeCANet model significantly outperforms current state-of-the-art recognition methods in both ordinary and high-similarity Dongba character recognition tasks, while also reducing the parameter scale compared with MAAN on DB1404. This fully demonstrates the excellent performance and practical application value of EdgeCANet.

5.6 | Performance of EdgeCANet on Open Datasets OBC306, Sketch-20 K, and ImageNet-Sketch

To validate the credibility, robustness, and universality of the proposed EdgeCANet model, this section conducts experiments and analyzes the performance of EdgeCANet on the publicly available datasets OBC306 (Oracle Bone Script Dataset) [4], Sketch-20 K [3], and ImageNet-Sketch. EdgeCANet follows the conventional 7:1:2 training scheme which is described in Section 5.1. Notably, these three datasets share common characteristics with DB1404, HS-DB, and HS-S20 K. Images across all datasets exhibit pronounced foreground-background contrast. Their visual representations resemble sketch-style drawings. Most critically, OBC306, DB1404, HS-DB, and HS-S20 K are pictographic datasets containing numerous highly similar characters, while Sketch-20 K, ImageNet-Sketch, and HS-S20 K are sketch-based datasets encompassing thousands of object categories, including many highly similar item classes. Therefore, selecting OBC306, Sketch-20 K, and ImageNet-Sketch to validate the reliability, robustness, and universality of EdgeCANet is reasonable. Detailed experimental results for each dataset are presented in Table 8.

OBC306 is a large-scale dataset for Oracle Bone Script recognition, with sample examples shown in Figure 12. Following the data preprocessing method used in MAAN [2], we excluded

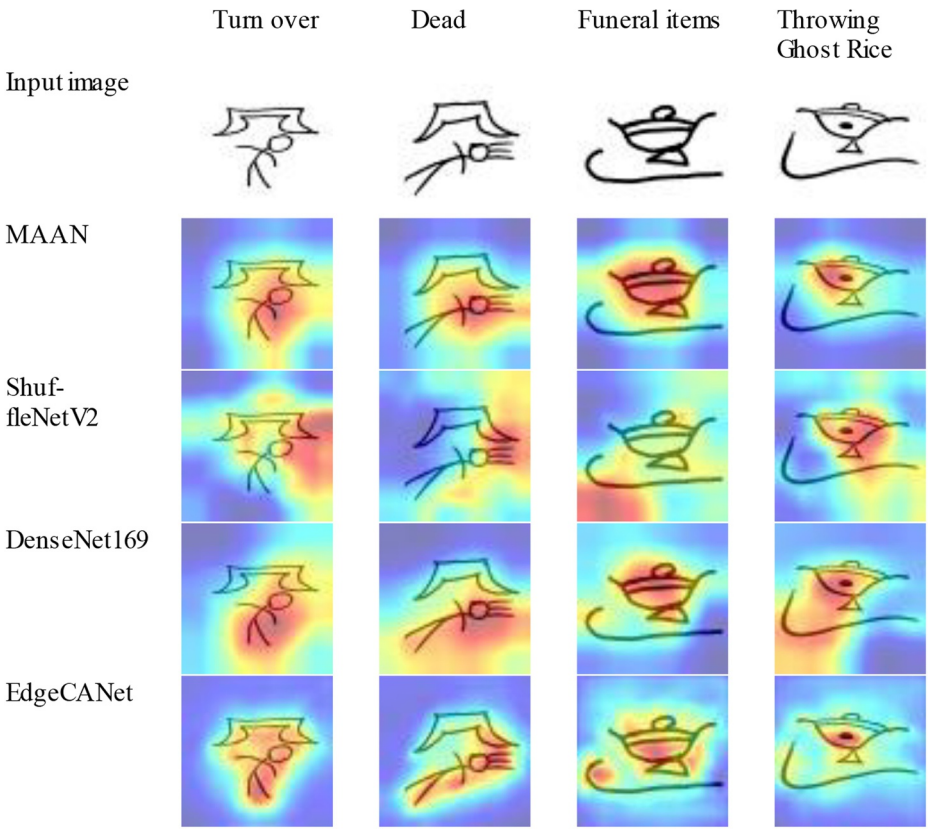


FIGURE 11 | Grad-CAM visualisation results.

TABLE 8 | Recognition performance comparison of EdgeCANet with other SOTA methods on OBC306, Sketch-20 K, and ImageNet-Sketch datasets.

Open Dataset	Method	Top-1	Data augmentation
OBC306	DenseNet+cross-modal [46]	86.70%	—
	Random polygon cover [47]	91.53%	✓
	Mix-up aug+Triplet loss [48]	91.74%	✓
	MAAN [2]	92.83%	—
	EdgeCANet (ours)	**92.98%**	—
Sketch-20 K	MKL [49]	65.81%	—
	$F_{P-V_7} - F_{X-V_9} - F_{R-V10}$ [50]	72.93%	✓
	Double-channel CNN [51]	73.24%	✓
	Hierarchical residual CNN [52]	76.14%	✓
	MAAN [2]	67.84%	—
	MAAN [2]	73.93%	✓
	EdgeCANet (ours)	**70.05%**	—
	EdgeCANet (ours)	**76.35%**	✓
ImageNet-Sketch	Pix2Pix-Anisotropic [53]	24.49%	—
	Vit-B/16+PyramidAT [54]	41.04%	—
	Vit-B/16+PyramidAT [54]	46.03%	✓
	RVT-B [55]	36.00%	—
	MGTTA [56]	53.30%	—
	LP-FT [57]	48.40%	—
	RandConv [58]	18.09%	✓
	EdgeCANet (ours)	**66.31%**	—
	EdgeCANet (ours)	**69.15%**	✓

Note: The best results are shown in bold.

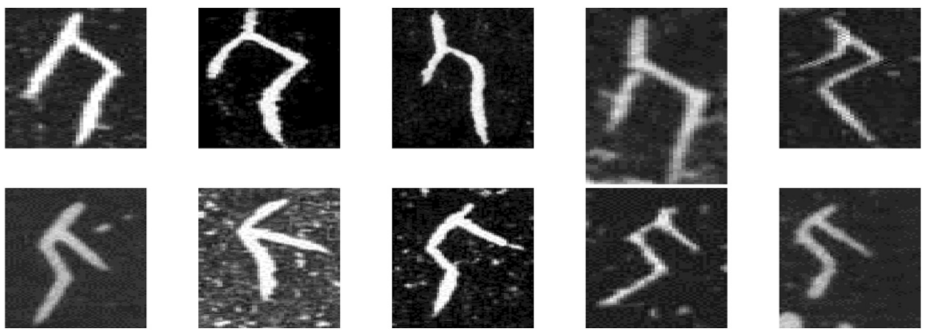


FIGURE 12 | Example of the OBC306 dataset. Each row corresponds to a category.

categories containing only a single image. After filtering, 277 categories were retained for training and evaluation. EdgeCANet achieved a Top-1 accuracy of 92.98%, surpassing the best MAAN by 0.15%. MAAN [2] aggregates large-scale local detail information with small-scale global semantic information to enhance fine-grained recognition capabilities. As a recognition system tailored for handwritten Dongba script, it still faces the key challenge in high-similarity image recognition when identifying OBC306 images with subtle edge differences in certain categories. EdgeCANet employs a dual-branch attention architecture combining edge and channel components, enabling it to better adapt to and identify images where the foreground and background differ significantly. Combined with EMA training stability and second-stage fine-tuning focused on hard-to-classify samples, it continuously enhances boundary/micro-structure separability, achieving a moderate yet stable 0.155% net gain on OBC306.

The Sketch-20 K [3] dataset is commonly used for sketch image recognition, comprising 250 categories with sample examples shown in Figure 13. On this dataset, EdgeCANet achieves a Top-1 recognition accuracy of 76.35%. Compared to state-of-the-art methods without data augmentation, EdgeCANet surpasses the best-performing model MAAN [2] by 2.21% in Top-1 accuracy. Compared to other state-of-the-art models employing data augmentation techniques, EdgeCANet still outperforms the

Hierarchical Residual CNN [52] by 0.21%, achieving the highest known recognition accuracy on this dataset. The Hierarchical Residual CNN designs a multi-scale residual block and stacks it hierarchically, paired with a compact triplet centre loss to widen the interclass gap in sketches. While the hierarchical design captures multi-scale shapes, it does not explicitly encode edge saliency or stroke topology, interval-based metric learning optimises global separability but may underestimate short strokes and stroke-end features dominant in hand-drawn sketches. EdgeCANet's edge attention mechanism prioritises target contours. Combining CTRM units to reweight information channels for better alignment with edge features in target contours within sketch data, alongside EMA to enhance training stability and second-stage fine-tuning to guide the model toward focusing on features of hard-to-classify samples, this approach significantly improves sensitivity to fine lines. Under data augmentation settings, this solution achieves a 0.21% performance gain.

ImageNet-Sketch is a sketch image dataset comprising 1000 categories, with examples shown in Figure 14. On this dataset, EdgeCANet achieved the highest recognition accuracy of 69.15% among non-large-model approaches, surpassing the current state-of-the-art method MGTTA by 15.85%, representing a breakthrough advancement. MGTTA is the Test-Time Adaptation framework that learns a Meta Gradient Generator and a Gradient Meta Learner to generate optimised adaptation gradients for out-of-distribution data, demonstrating broad gains across ImageNet-C/R/Sketch/A. Its advantage lies in achieving instant robustness through gradient generation without cumbersome retraining. Test-Time Adaptation enhances robustness but relies on adaptation dynamics during inference and is not tailored for sketch edge statistics, it also introduces adaptation overhead and may fail to fully leverage category-discriminative stroke geometry when the backbone lacks explicit edge priors. EdgeCANet is architecturally biased toward sketch/contour data: EGEM directs attention to contours, corners, and stroke terminations, while CTRM units emphasise channels responsive to edge-dominant features. Combined with EMA and fine-tuning strategies, this yields substantial gains on 1000-class, sparse-stroke benchmarks without requiring test-time optimization.

EdgeCANet demonstrates consistent advantages over the best competing models across three public datasets. These gains align with the model's design objective of explicit edge-channel aggregation for highly similar characters and sketch images.

In summary, there is no doubt that EdgeCANet demonstrates robust performance across multiple datasets, validating its robustness and generalisation capabilities across diverse tasks and data distribution scenarios.

6 | Conclusions

Handwritten Dongba Character Recognition contains many visually similar characters with subtle and fragile edge cues, making it challenging to distinguish highly similar handwritten Dongba characters. To address this, we propose an Edge Channel Aggregation Network, which uses the Edge Channel Attention Aggregation block and its Edge-Detail Synergistic Attention mechanism, to enhance critical edge perception and discriminative representation capabilities. An EMA-based training scheme with an error-weighted fine-tuning strategy is also proposed to efficiently extract target edge and shape information and improve recognition accuracy for difficult-to-classify samples. We also construct two high-similarity datasets HS-DB and HS-S20 K to validate the performance of our proposed model. HS-DB contains 4414 handwritten Dongba

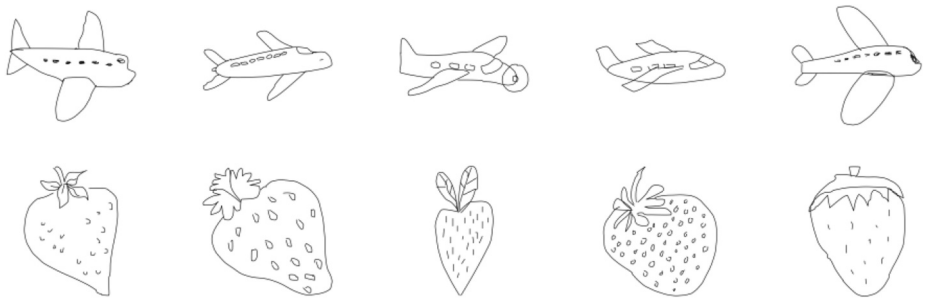


FIGURE 13 | Examples from Sketch-20 K dataset. Each row corresponds to a category. The first row shows sample sketches from the aeroplane category, while the second row presents samples from the strawberry category.

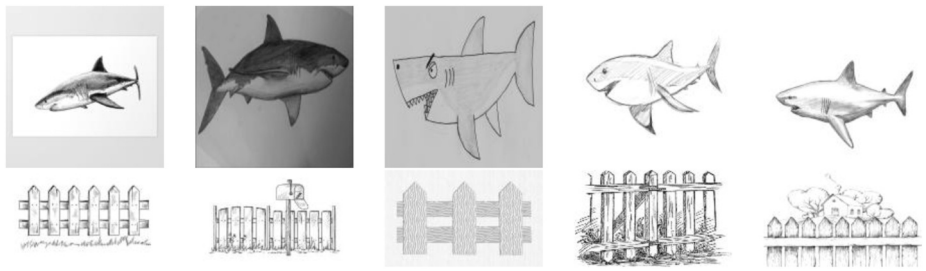


FIGURE 14 | Examples from ImageNet-Sketch dataset. Each row corresponds to a specific category. The first row illustrates sketches from shark category, while the second row presents samples from fence category.

character images covering a broader range of categories, while HS-S20 K features contains 1600 highly similar sketches of sketch objects spanning diverse categories.

Experiments conducted on HS-DB and HS-S20 K datasets demonstrate that EdgeCANet achieves superior results compared to SOTA methods in high-similarity scenarios. Evaluation on public datasets DB1404, OBC306, Sketch-20 K, and ImageNet-Sketch highlights our model's reliability, robustness, and generalisation capabilities compared to SOTA models. EdgeCANet effectively enhances recognition accuracy for highly similar handwritten characters and ancient minority-script characters, offering more promising applicability for related recognition tasks.

However, our work still has certain limitations. The scale of the model parameters remains relatively large compared to lightweight architectures that have emerged in recent years. This still imposes certain constraints when deployed on mobile devices or in resource-constrained scenarios. Future work will focus on exploring methods such as parameter sharing, structural compression, or knowledge distillation, and investigating whether hierarchical neural structures can further improve efficiency under limited computational resources [59]. The goal is to further enhance model compactness and inference efficiency while maintaining high accuracy, thereby improving recognition efficiency and accuracy for Dongba script and sketch images.

Funding

This work is supported by the National Natural Science Foundation of China under Grant No. 62276285, the key Projects of National Natural Science Foundation of China under Grant No. 62236011, the Major Project of National Social Sciences Foundation of China under Grant No. 20&ZD279.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Our self built two high similarity datasets can be gotten at https://github.com/xy1369/high-similarity-datasets.git. The code of our model can be gotten by the email xiaer_li@163.com.

References

1. P. He, W. Zhang, H. Yang, and M. Zeng, *Yunnan Minority Literature Picture Book Collection Genesis Chinese Naxi Hieroglyphics in English* (Yunnan People's Publishing House, 2021).

2. Y. Luo, Y. Sun, and X. Bi, "Multiple Attentional Aggregation Network for Handwritten Dongba Character Recognition," *Expert Systems with Applications* 213 (2023): 118865, https://doi.org/10.1016/j.eswa.2022.118865.

3. M. Eitz, J. Hays, and M. Alexa, "How do Humans Sketch Objects?," *ACM Transactions on Graphics* 31, no. 4 (2012): 1–10, https://doi.org/10.1145/2185520.2335395.

4. S. Huang, H. Wang, Y. Liu, et al., "Obc306: A Large-Scale Oracle Bone Character Recognition Dataset," in *2019 International Conference on Document Analysis and Recognition (ICDAR)* (IEEE, 2019), 681–688.

5. S. Woo, S. Debnath, R. Hu, et al., "Convnext v2: Co-Designing and Scaling Convnets With Masked Autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 16133–16142.

6. W. Wang, J. Dai, Z. Chen, et al., "Internimage: Exploring Large-Scale Vision Foundation Models With Deformable Convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), 14408–14419.

7. R. Hua and X. Xu, "Intelligent Classification on Images of Dongba Ancient Books," *Journal of Engineering* 2019, no. 23 (2019): 9039–9042, https://doi.org/10.1049/joe.2018.9177.

8. Y. Luo, X. Bi, L. Wu, et al., "Recognition of Dongba Pictographs Based on Improved Residual Learning," *CAAI Transactions on Intelligent Systems* 17, no. 1 (2021): 79–87, https://doi.org/10.11992/tis.202112009.

9. H. Luo, D. Xu, B. Yang, et al., "Multi-Scale Feature Fusion Based Dongba Character Recognition," in *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)* (IEEE, 2020), 1571–1575.

10. X. Bi and Y. Luo, "Incomplete Handwritten Dongba Character Image Recognition by Multiscale Feature Restoration," *Heritage Science* 12, no. 1 (2024): 1–19, https://doi.org/10.1186/s40494-024-01329-8.

11. J. Li, X. Chi, Q. Wang, et al., "A Comprehensive Survey of Oracle Character Recognition: Challenges, Datasets, Methodology, and Beyond," *Pattern Recognition* 169 (2026): 111824, https://doi.org/10.1016/j.patcog.2025.111824.

12. M. A. Ko and S. Poruran, "OCR-nets: Variants of Pre-Trained CNN for Urdu Handwritten Character Recognition via Transfer Learning," *Procedia Computer Science* 171 (2020): 2294–2301, https://doi.org/10.1016/j.procs.2020.04.248.

13. M. A. Kilvisharam Oziuddeen, S. Poruran, and M. Y. Caffiyar, "A Novel Deep Convolutional Neural Network Architecture Based on Transfer Learning for Handwritten Urdu Character Recognition," *Tehnički vjesnik* 27, no. 4 (2020): 1160–1165, https://doi.org/10.17559/TV-20190319095323.

14. A. M. Kumar, J. I. R. Praveen, S. Venkatraman, et al., "Improved Tomato Leaf Disease Classification Through Adaptive Ensemble Models With Exponential Moving Average Fusion and Enhanced Weighted Gradient Optimization," *Frontiers in Plant Science* 15 (2024): 1382416, https://doi.org/10.3389/fpls.2024.1382416.

15. H. Wu, Y. Geng, D. Gai, et al., "Dual-Student Adversarial Framework With Discriminator and Consistency-Driven Learning for Semi-Supervised Medical Image Segmentation," *IEEE Journal of Biomedical and Health Informatics* 30, no. 2 (2026): 1492–1505, https://doi.org/10.1109/jbhi.2025.3597469.

16. He K., Fan H., Wu Y., et al., "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 9729–9738.

17. A. Tarvainen, "Mean Teachers are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results," *Advances in Neural Information Processing Systems* 30 (2017): 1196, https://dl.acm.org/doi/10.5555/3294771.3294885.

18. L. Wang, S. Chen, L. Jiang, et al., "Parameter-Efficient Fine-Tuning in Large Language Models: A Survey of Methodologies," *Artificial Intelligence Review* 58, no. 8 (2025): 227, https://doi.org/10.1007/s10462-025-11236-4.

19. T. ValizadehAslani, Y. Shi, J. Wang, et al., "Two-Stage Fine-Tuning With Chatgpt Data Augmentation for Learning Class-Imbalanced Data," *Neurocomputing* 592 (2024): 127801, https://doi.org/10.1016/j.neucom.2024.127801.

20. S. Bondugula, G. Qian, and A. Beach, "Two-Stage Training for Improved Classification of Poorly Localized Object Images," in *European Conference on Computer Vision* (Springer International Publishing, 2020), 245–260, https://link.springer.com/chapter/10.1007/978-3-030-68238-5_18.

21. M. Alsaidi, A. S. Altaher, M. T. Jan, et al., "Covid-19 Classification Using Deep Learning Two-Stage Approach," arXiv preprint, arXiv: 2211.15817, November 28, 2022, https://doi.org/10.48550/arXiv.2211.15817.

22. T. Y. Lin, P. Goyal, R. Girshick, et al., "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision* (2017), 2980–2988.

23. K. Cao, C. Wei, A. Gaidon, et al., "Learning Imbalanced Datasets With Label-Distribution-Aware Margin Loss," *Advances in Neural Information Processing Systems* (2019): 32.

24. X. Yuan, X. Xu, X. Wang, et al., "Osap-Loss: Efficient Optimization of Average Precision via Involving Samples After Positive Ones Towards Remote Sensing Image Retrieval," *CAAI Transactions on Intelligence Technology* 8, no. 4 (2023): 1191–1212, https://doi.org/10.1049/cit2.12151.

25. C. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data* 6, no. 1 (2019): 1–48, https://doi.org/10.1186/s40537-019-0197-0.

26. C. Luo, Y. Zhu, L. Jin, et al., "Learn to Augment: Joint Data Augmentation and Network Optimization for Text Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 13746–13755.

27. P. Alimisis, I. Mademlis, P. Radoglou-Grammatikis, P. Sarigiannidis, and G. T. Papadopoulos, "Advances in Diffusion Models for Image Data Augmentation: A Review of Methods, Models, Evaluation Metrics and Future Research Directions," *Artificial Intelligence Review* 58, no. 4 (2025): 112, https://doi.org/10.1007/s10462-025-11116-x.

28. X. Xiong, Y. Sun, X. Liu, et al., "Distance Guided Generative Adversarial Network for Explainable Medical Image Classifications," *Computerized Medical Imaging and Graphics* 118 (2024): 102444, https://doi.org/10.1016/j.compmedimag.2024.102444.

29. C. Zheng, G. Wu, and C. Li, "Toward Understanding Generative Data Augmentation," *Advances in Neural Information Processing Systems* 36 (2023): 54046–54060, https://doi.org/10.52202/075280-2352.

30. K. Islam, M. Z. Zaheer, A. Mahmood, et al., "Diffusemix: Label-Preserving Data Augmentation With Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), 27621–27630.

31. T. Takikawa, D. Acuna, V. Jampani, et al., "Gated-Scnn: Gated Shape Cnns for Semantic Segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), 5229–5238.

32. N. T. Bui, D. H. Hoang, Q. T. Nguyen, et al., "Meganet: Multi-Scale Edge-Guided Attention Network for Weak Boundary Polyp Segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2024), 7985–7994.

33. J. Hu, L. Shen, and G. Sun, "Squeeze-And-Excitation Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 7132–7141.

34. Y. Huang, F. Liu, and A. G. S. M. C. Zhou, "A Global-Local Scalable Multi-Task Contrastive Learning Framework," in *Computer Graphics International Conference* (Springer Nature Switzerland, 2024), 297–308.

35. F. Liu, Z. Fu, and Y. Wang, "Reward-Based Gradient Modulation for Multimodal Emotion Recognition With Lora," *IEEE Transactions on Computational Social Systems* 12, no. 5 (2025): 1–11, https://doi.org/10.1109/tcss.2025.3566373.

36. A. S. Sambyal, U. Niyaz, S. Shrivastava, et al., "Ls+: Informed Label Smoothing for Improving Calibration in Medical Image Classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer Nature Switzerland, 2024), 513–523.

37. B. Huang, H. Sharma, and Y. Mao, "Enhancing Language Model Alignment: A Confidence-Based Approach to Label Smoothing," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (2024), 21341–21352.

38. K. He, X. Zhang, S. Ren, et al., "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 770–778.

39. G. Huang, Z. Liu, L. Van Der Maaten, et al., "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), 4700–4708.

40. S. Xie, R. Girshick, P. Dollár, et al., "Aggregated Residual Transformations for Deep Neural Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), 1492–1500.

41. N. Ma, X. Zhang, H. T. Zheng, et al., "Shufflenet v2: Practical Guidelines for Efficient Cnn Architecture Design," in *Proceedings of the European Conference on Computer Vision* (ECCV, 2018), 116–131.

42. M. Tan and Q. Le, "Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning* (PMLR, 2019), 6105–6114.

43. H. Zhang, C. Wu, Z. Zhang, et al., "Resnest: Split-Attention Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 2736–2746.

44. X. Ding, X. Zhang, N. Ma, et al., "Repvgg: Making vgg-style Convnets Great Again," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), 13733–13742.

45. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision* 128, no. 2 (2020): 336–359, https://doi.org/10.1007/s11263-019-01228-7.

46. Z. Yi-Kang, Z. Heng, L. Yong-Ge, et al., "Oracle Character Recognition Based on Cross-Modal Deep Metric Learning," *Acta Automatica Sinica* 47, no. 4 (2021): 791–800, https://doi.org/10.16383/j.aas.c200443.

47. L. Dazheng, "Random Polygon Cover for Oracle Bone Character Recognition," in *Proceedings of the 2021 5th International Conference on Computer Science and Artificial Intelligence* (2021), 138–142.

48. J. Li, Q. F. Wang, R. Zhang, et al., "Mix-up Augmentation for Oracle Character Recognition With Imbalanced Data Distribution," in*International Conference on Document Analysis and Recognition* (Springer International Publishing, 2021), 237–251.

49. Y. Li, T. M. Hospedales, Y. Z. Song, and S. Gong, "Free-Hand Sketch Recognition by Multi-Kernel Feature Learning," *Computer Vision and Image Understanding* 137 (2015): 1–11, https://doi.org/10.1016/j.cviu.2015.02.003.

50. S. Hayat, S. Kun, S. Shahzad, P. Suwansrikham, M. Mateen, and Y. Yu, "Entropy Information-Based Heterogeneous Deep Selective Fused Features Using Deep Convolutional Neural Network for Sketch Recognition," *IET Computer Vision* 15, no. 3 (2021): 165–180, https://doi.org/10.1049/cvi2.12019.

51. L. Zhang, "Hand-Drawn Sketch Recognition With a Double-Channel Convolutional Neural Network," *EURASIP Journal on Applied Signal Processing* 2021, no. 1 (2021): 1–12, https://doi.org/10.1186/s13634-021-00752-4.

52. L. Wang, S. Zhang, H. He, X. Zhang, and Y. Sang, "A Hierarchical Residual Network With Compact Triplet-Center Loss for Sketch

Recognition," *Multimedia Tools and Applications* 81, no. 11 (2022): 15879–15899, https://doi.org/10.1007/s11042-022-12431-z.

53. S. Mishra, A. Shah, A. Bansal, et al., "Learning Visual Representations for Transfer Learning by Suppressing Texture," in *33rd British Machine Vision Conference Proceedings* (BMVC, 2022).

54. C. Herrmann, K. Sargent, L. Jiang, et al., "Pyramid Adversarial Training Improves Vit Performance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 13419–13429.

55. X. Mao, G. Qi, Y. Chen, et al., "Towards Robust Vision Transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), 12042–12051.

56. Q. Deng, S. Niu, R. Zhang, et al., "Learning to Generate Gradients for Test-Time Adaptation via Test-Time Training Layers," *Proceedings of the AAAI Conference on Artificial Intelligence* 39, no. 15 (2025): 16235–16243, https://doi.org/10.1609/aaai.v39i15.33783.

57. A. Kumar, A. Raghunathan, J. Robbie, et al., "Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution," in *10th International Conference on Learning Representations* (ICLR, 2022).

58. Z. Xu, D. Liu, J. Yang, et al., "Robust and Generalizable Visual Representation Learning via Random Convolutions," arXiv preprint, arXiv:2007.13003, July 25, 2020, https://doi.org/10.48550/arXiv.2007.13003.

59. X. Li, X. Fan, J. Yu, Z. Dong, X. Cairang, and P. Lan, "Jiu Fusion Artificial Intelligence (JFA): A Two-Stage Reinforcement Learning Model With Hierarchical Neural Networks and Human Knowledge for Tibetan Jiu Chess," *Frontiers of Information Technology & Electronic Engineering* 26, no. 10 (2025): 1969–1983, https://doi.org/10.1631/fitee.2500287.