electronics

MDPI

Article

Anime Character Style Classification Based on Frequency-Domain Decoupling and Multi-Scale Feature Fusion

Yunfeng Chen [1], Junxiang Diao [1,*], Hua Wei [1] and Zhihua Diao [2,*]

[1] School of Art and Design, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 2014011@zzuli.edu.cn (Y.C.); 2002066@zzuli.edu.cn (H.W.)
[2] College of Electrical and Information Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China
* Correspondence: 2017013@zzuli.edu.cn (J.D.); 2010018@zzuli.edu.cn (Z.D.)

Abstract

Automatic classification of anime character painting styles is of great significance to the digital cultural industry and visual content production. Existing methods are prone to shortcut learning when handling complex color rendering and cannot fully decouple high-frequency line drafts from low-frequency colors. To solve this problem, this study proposes an improved deep learning classification method based on EfficientNetV2-B0. This method introduces random amplitude scaling (RAS) at the data input terminal. It realizes effective decoupling of colors and line-draft structures through random low-frequency amplitude perturbation, and suppresses the model's excessive dependence on global color information from the source. Edge-guided coordinate attention (EG-CA) is integrated into the backbone network. It enhances the perception of line and contour features through edge weights and improves the model's ability to capture fine-grained structural features. Adaptive scale feature aggregation (ASFA) is designed in the multi-scale feature fusion stage. It achieves efficient fusion of shallow textures and deep semantics through dynamic weighting, so as to enhance the model's discriminative ability under complex painting styles. On a dataset containing 7887 images of four categories, the classification accuracy of the model reaches 95.81%. It significantly outperforms mainstream models such as MViTv2-T. Meanwhile, the number of parameters is only 7.84 M and the inference speed reaches 68.83 FPS. Ablation experiments show that the synergistic effect of the three modules improves the accuracy of the baseline model by 6.06%. It proves that the proposed method provides reliable technical support for the structured management and copyright traceability of anime images.

Keywords: anime character style classification; multi-scale feature fusion; deep learning; improved EfficientNetV2-B0

Check for updates

Academic Editor: Kefeng Ji

Received: 3 April 2026
Revised: 11 May 2026
Accepted: 15 May 2026
Published: 17 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As a key expressive form in the digital culture industry and visual content production, anime character painting styles embody the aesthetic orientations and visual symbol systems of specific creator groups, and are widely applied in scenarios such as animation production, game art, illustration creation and virtual content generation [1–3]. With the rapid development of digital media and online platforms, the systematic organization, retrieval and management of anime character painting styles have become important fundamental tasks supporting content understanding, copyright traceability and creation assistance. Among these, style classification, as the core task of anime image database

construction and content analysis, directly influences the performance of subsequent applications such as style retrieval, similarity measurement and cross-domain generation [4,5]. Especially in applications such as copyright traceability, painting style features are vulnerable to the interference of complex visual noise. The concept of robust feature learning and active detection in complex scenarios emphasized by recent studies [6] provides an effective approach to improve the anti-interference ability of features and content credibility. Therefore, investigating and optimizing classification methods for anime character painting styles possesses distinct theoretical value and practical significance.

Different from image classification with significant inter-class differences, the classification of anime character painting styles is essentially a challenging fine-grained visual recognition task. Anime character painting styles of different categories often exhibit high similarity in character pose and skeleton topology. Their core fine-grained differences are mainly reflected in the head–body proportion of characters, the exaggeration degree of facial features, the representation form of line art, and the realism of light and shadow rendering [7,8]. For example, the deformation degree of geometric contours and the difference in underlying color texture are the most basic visual cues for defining different anime character painting styles [9]. Meanwhile, high-frequency structural line art and low-frequency color rendering in anime images are often highly coupled. Personal habits of different authors further increase the complexity of visual features within the same painting style [10,11]. These unique representation characteristics make anime character painting style recognition more complicated than traditional image classification tasks.

In recent years, deep learning technology has provided solid technical support for anime character style classification with its powerful feature extraction capability, rich multi-level semantic representation, and accurate discrimination of complex visual patterns. In terms of deep learning-based feature extraction, although the study by Young-Min et al. [12] shows that standard CNNs can effectively capture visual features of comics and illustrations, networks tend to show shortcut-learning tendencies that rely on color statistics when facing modern anime images with complex color rendering. This leads to insufficient sensitivity to subtle local changes in line art. In addition, the study by Xu et al. [13] points out that high-frequency line art structures and low-frequency color rendering in anime images are often highly coupled. Basic multi-scale fusion strategies lack adaptive ability in dynamically balancing structural and color semantics when processing features with high intra-class variance, which may cause the loss of key discriminative information. As noted by Li et al. [14] when constructing a benchmark for anime style recognition, different anime character styles usually exhibit high visual similarity and feature overlap. This puts forward higher requirements for the fine-grained discrimination and anti-interference ability of classification algorithms.

To address the above problems, this study proposes a fine-grained anime character style classification method based on improved EfficientNetV2-B0 [15]. A frequency-domain perturbation strategy named random amplitude scaling (RAS) is introduced at the feature input end. By explicitly suppressing low-frequency amplitude components, it achieves effective decoupling of high-frequency and low-frequency features and enhances the model's ability to resist interference from complex color rendering. Furthermore, an edge-guided coordinate attention (EG-CA) mechanism is integrated into the backbone network. Physical edge weights are used to guide the model to accurately focus on key structures such as facial features and geometric contours, which effectively compensates for the insensitivity of the network to subtle changes in local line art. In addition, to solve the problem that basic fusion strategies easily lead to the loss of discriminative information, this study designs an adaptive scale feature aggregation (ASFA) module. By dynamically balancing shallow textures and deep semantics, it realizes deep collaboration of cross-scale

features and significantly improves the discriminative stability of the model when dealing with complex feature overlap.

2. Materials and Methods

2.1. Dataset Collection and Preprocessing

Visual styles of anime characters exhibit high diversity and distinct aesthetic differences. In view of this, the image dataset constructed in this study mainly covers four typical drawing styles of anime characters, namely chibi art style, bold art style, realistic art style and moe art style, as shown in Figure 1.

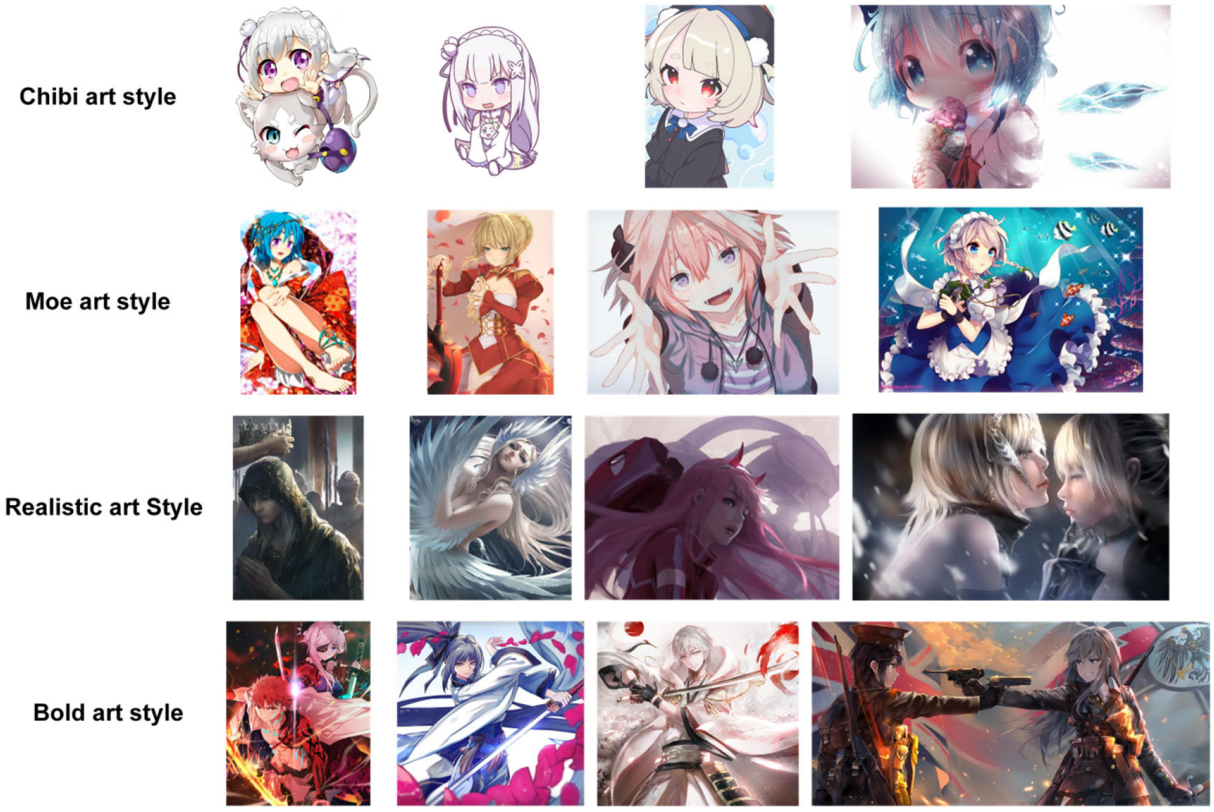


Figure 1. Examples of four anime character styles.

To determine the visual criteria for fine-grained classification, this dataset clearly defines the physical characteristics and rendering techniques of the four typical anime character painting styles. The chibi art style is characterized by an exaggerated proportion of 2 to 3 heads tall and extremely simplified facial topology, with internal textures and redundant details largely removed. The moe art style has a moderate head-to-body ratio, mostly adopts soft color transitions and rounded contour edges, and focuses on the delicate depiction of facial features. The realistic art style follows the proportion of natural human anatomy, shapes forms through smooth color transitions and natural physical light and shadow mapping, and basically abandons explicit external boundary contours. The bold art style takes clear and sharp structural line art as its core feature, adopts high contrast between light and dark colors, and presents distinct light and shadow expressions as well as strong geometric edge perception. Overall, the above samples show significant systematic differences in key dimensions such as character proportional relationship, line expression, color and light–dark organization, and texture presentation. These differences effectively improve the generalization ability of the dataset in different creative media and practical applications.

The raw data collected in this study contains a total of 7887 independent images. All images are manually screened to ensure accurate category annotations. To ensure the rationality of model training and the objectivity of evaluation, the dataset is randomly divided into a training set (80%), a validation set (10%) and a test set (10%) by category, as detailed in Table 1. This study strictly follows the best practices in the field of computer vision and organizes the dataset in the standard ImageNet format [16]. All raw images are converted to RGB format uniformly before being input into the network. According to different requirements for training and inference, dynamic data preprocessing and standardization strategies are adopted.

Table 1. The types and quantities of dataset labels.

Category	Training Set	Validation Set	Test Set
Chibi art style	1582	200	196
Bold art style	1571	197	197
Realistic art style	1584	197	199
Moe art style	1572	196	196

Notably, the anime character images used for dataset construction and illustration descriptions in this study are all derived from the public anime dataset of the PaddlePaddle AI Studio platform (https://aistudio.baidu.com/datasetdetail/237153, accessed on 10 May 2026). As an open-source computer vision sample library, all images included in this dataset strictly comply with the CC0 (Creative Commons Zero) license. The use of these images in this study is limited to academic research and illustration purposes to ensure compliance with academic standards and copyright requirements.

2.2. Proposed Method

2.2.1. EfficientNetV2 Network

EfficientNetV2 is an improved version of the EfficientNetV1 [17] series models. It is a convolutional neural network that focuses on training speed and computational efficiency while ensuring model accuracy. The model fully considers problems such as high computational cost and long training time in the actual training process, and makes targeted adjustments to the network structure. EfficientNetV2 inherits the basic idea of unified scaling of network depth, width and input resolution from the EfficientNetV1 series. Through reasonable allocation across different dimensions, the model can maintain satisfactory performance under different computational resource conditions.

Among the EfficientNetV2 series, EfficientNetV2-B0 is a basic model with a relatively small structure, which is mainly used for small- and medium-scale image classification tasks. The model has a shallow overall depth and a small number of parameters, making it suitable for training and testing with limited computational resources. EfficientNetV2-B0 introduces various convolutional modules into the network structure and selects appropriate module forms according to different network layers. In this way, it minimizes unnecessary computational overhead while ensuring feature representation capability.

In the shallow part of the network, EfficientNetV2-B0 mainly adopts standard convolutional structures. This structure can quickly extract basic image features such as edge information and local texture information at a low computational cost. Since shallow features play an important role in subsequent feature extraction, the use of computationally efficient convolution helps accelerate the overall training process. As the network depth gradually increases, the model gradually introduces inverted residual structures to enhance the ability to extract complex features. These structures mainly include Fused-MBConv and MBConv modules.

The structure of Fused-MBConv is shown in Figure 2. Input features first go through a 3×3 standard convolutional layer for feature extraction, and the feature map transformation is completed according to the convolution stride. When residual connection is enabled, DropPath is introduced at the output end of the residual branch for regularization. When the convolution stride is 1 and the spatial size and channel number of input and output features of the module remain consistent, residual connection is used to add input features and convolutional output features. When the convolution stride is 2 and the module is used in the downsampling stage, residual connection is not applied and the feature map obtained by convolution is directly output. The Fused-MBConv module can be divided into two structural forms according to different channel expansion ratios. When the channel expansion ratio is not equal to 1, the module first expands the number of channels through a 3×3 standard convolution and extracts spatial features, and then compresses the number of channels using a 1×1 convolution. When the channel expansion ratio is equal to 1, the module only uses a 3×3 standard convolution to complete feature extraction.

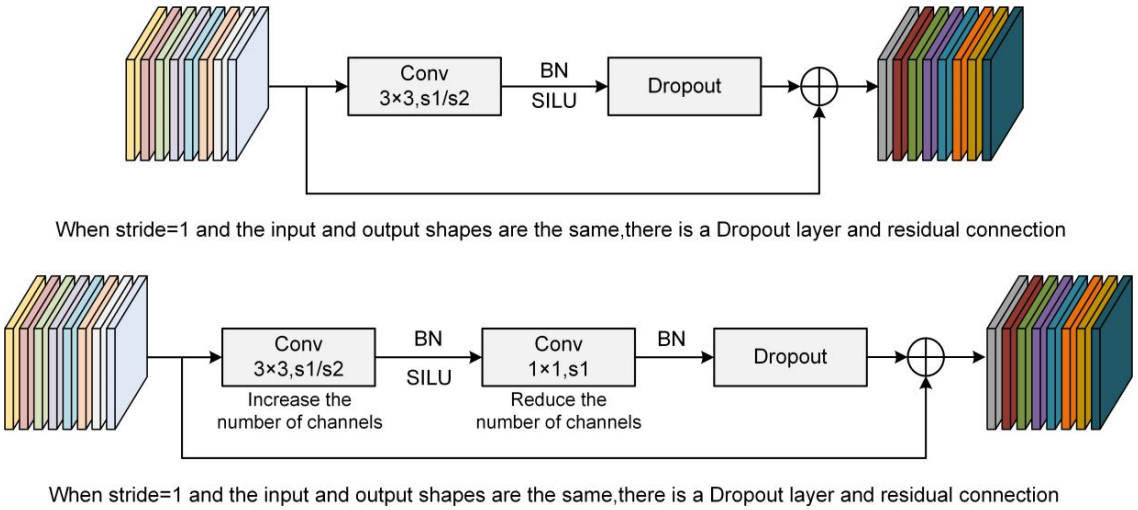


Figure 2. Fused-MBConv structure.

The structure of MBConv is shown in Figure 3. Input features first pass through a 1×1 convolutional layer to expand the number of channels and project feature maps into a high-dimensional space. Spatial features are then extracted through a depthwise convolutional layer. After depthwise convolution, the Squeeze-and-Excitation (SE) attention mechanism [18] is introduced to perform weighted adjustment on channel features. A 1×1 convolutional layer with a linear activation function is then adopted to compress the number of channels so that the output feature dimension matches the subsequent network structure. The MBConv module can be divided into two structural forms according to the stride of the depthwise convolutional layer. When the stride of the depthwise convolutional layer is 1 and the spatial size and channel number of input and output features remain consistent, DropPath is introduced at the output end of the residual branch for regularization. Input features and output features are then added through residual connection. When the stride of the depthwise convolutional layer is not 1 and the module is used in the downsampling stage, neither DropPath nor residual connection is applied. Features are directly output after dimension reduction by the 1×1 convolution.

2.2.2. RAS Frequency-Domain Perturbation Strategy

In the task of anime character style classification, when facing highly complex underlying rendering features, the model tends to rely excessively on large-area color rendering and cannot fully capture line features that represent local structures. This leads to degraded

generalization ability when encountering color interference. Existing frequency-domain generalization methods represented by PASTA [19] usually improve model robustness by significantly perturbing high-frequency amplitudes proportionally. However, in anime character style classification, high-frequency line art is exactly the core discriminative feature that must be strictly preserved. To overcome this limitation and effectively decouple style and structure, inspired by the Fourier frequency-domain characteristics of images, this study proposes a RAS frequency-domain perturbation strategy. In the spectral representation, the phase spectrum usually captures precise edge and structural semantics, while the low-frequency components of the amplitude spectrum mainly represent underlying statistics such as smooth colors and global illumination. Based on this, RAS explicitly applies random scaling perturbations to the central low-frequency region. It destroys the original rendering style while retaining high-frequency local line features to the greatest extent. The flowchart of RAS is shown in Figure 4.

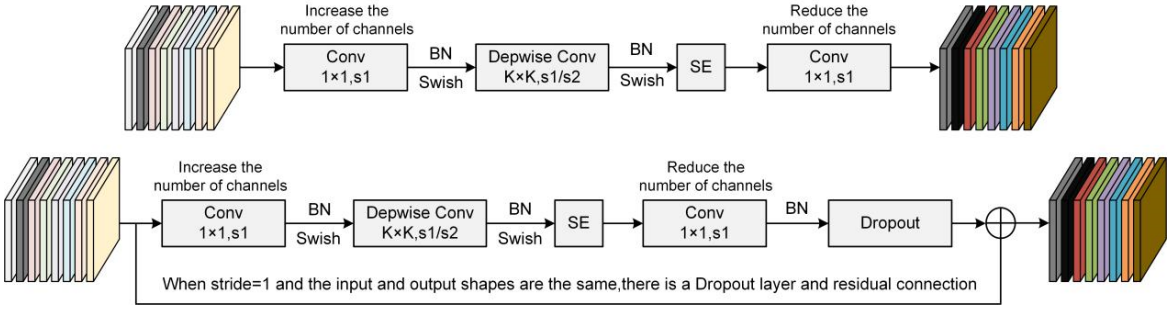


Figure 3. MBConv structure.

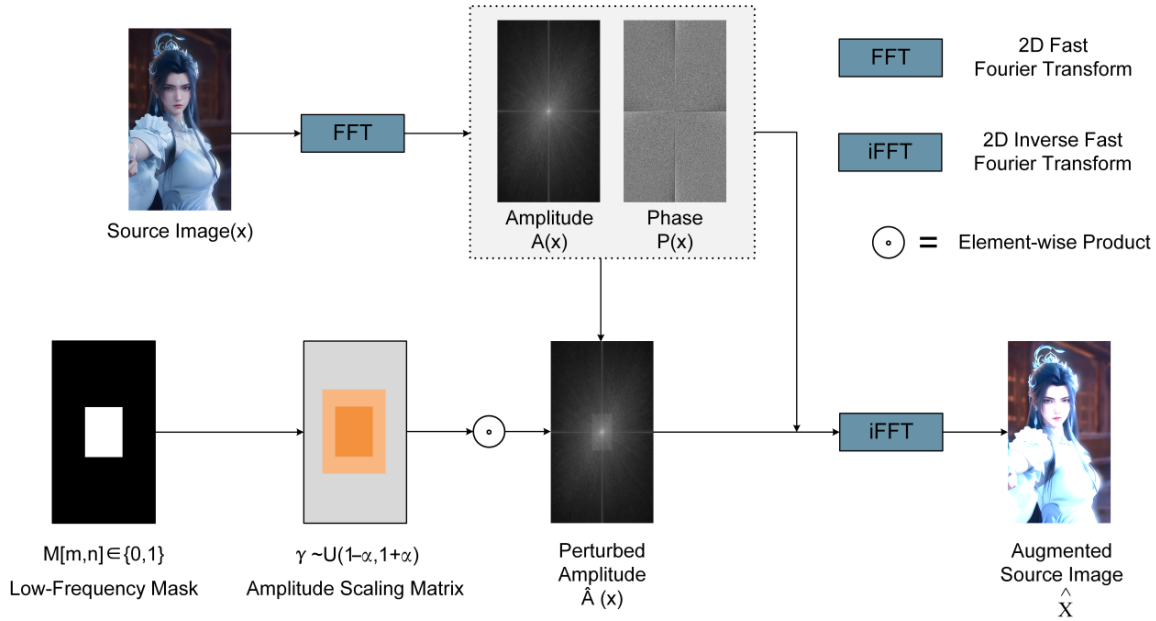


Figure 4. RAS flowchart (In the Amplitude Scaling Matrix, the orange area represents the random perturbation multiplier γ applied to the low-frequency region, while the gray area corresponds to a value of 1, keeping the high-frequency region unperturbed).

Given an input single-channel image $x \in R^{H \times W}$, RAS first converts it into the frequency domain $F(x)$ using two-dimensional Fast Fourier Transform (2D FFT), and extracts the centered amplitude spectrum $A(x)$ and phase spectrum $P(x)$. To accurately locate the low-frequency components that control the global rendering style, the module defines a central

low-frequency mask $M \in \{0,1\}^{H \times W}$. The mask is set to 1 only in a rectangular region within a preset radius ratio r from the spectrum center, and 0 in other high-frequency regions.

Using this mask, RAS applies random scalar multiplicative jitter to the low-frequency regions of the amplitude spectrum. The perturbation multiplier γ is independently sampled from the uniform distribution $\gamma \sim U(1-\alpha, 1+\alpha)$, where α is a hyperparameter that controls the perturbation intensity. The calculation formula for the perturbed new amplitude spectrum $\hat{A}(x)$ is shown in Equation (1).

$$\hat{A}(x) = A(x) \odot (1 - M) + A(x) \odot M \cdot \gamma \quad (1)$$

where $\odot$ denotes element-wise multiplication. This operation allows fine texture radiation in the peripheral part of the amplitude spectrum to remain completely unaffected by scaling. The central energy that determines color contrast and overall exposure is dynamically enlarged or reduced.

After obtaining the perturbed feature $\hat{A}(x)$, the module recombines it with the original phase spectrum $P(x)$ without any modification. The final enhanced image $\hat{x}$ is generated by reconstruction using two-dimensional inverse Fast Fourier Transform (2D iFFT) and taking the real part, as defined in Equation (2).

$$\hat{x} = \mathrm{Re}(F^{-1}(\hat{A}(x) \cdot \exp(j \cdot P(x)))) \quad (2)$$

Through this frequency-domain scaling mechanism, the reconstructed image perfectly preserves key line art contours such as hair strands and facial features at the pixel level, while the global color distribution undergoes drastic random changes. In addition, the slight Gibbs phenomenon caused by the hard truncation mask in the spatial domain acts as high-quality high-frequency structural regularization noise. Such interference naturally introduced by physical mechanisms effectively blocks the network's tendency of shortcut learning based on color statistics, thus providing robust low-level visual input for subsequent feature extraction modules to accurately capture line differences. It should be noted that RAS serves as a data augmentation and regularization mechanism and is only applied during the network training phase. In the inference phase, RAS is completely disabled. The network only performs standard spatial domain preprocessing on test images to ensure the stability and repeatability of model outputs.

2.2.3. EG-CA Module

Coordinate attention (CA) [20] decomposes two-dimensional global pooling into one-dimensional feature encoding in horizontal and vertical directions. It can capture cross-channel semantics while retaining precise positional information. However, this mechanism adopts global average pooling in spatial feature aggregation. Such undifferentiated pooling treats high-frequency line art edges and smooth colored backgrounds equally. This leads to the dilution of key discriminative clues representing different character art styles by extensive color information in deep networks. To overcome this limitation, this study proposes the EG-CA mechanism. This module explicitly injects physical edge priors into the pooling stage. It effectively guides the network to focus highly on contour and texture boundaries during feature aggregation, so as to accurately extract discriminative core topological structures. The flowchart of the EG-CA mechanism is shown in Figure 5.

Given an input feature map $X \in R^{C \times H \times W}$, EG-CA first computes the mean along the channel dimension to extract a grayscale feature representation. Then, fixed 3×3 Sobel operators [21] are used to calculate gradients in the horizontal and vertical directions respectively, and the gradient magnitude is obtained. To ensure stable feature scales and avoid additional training costs, this extraction process does not participate in gradient

backpropagation. Min–Max normalization is performed on the magnitude map in the spatial dimension to obtain an edge weight map $E \in R^{1\times H\times W}$ within the range [0, 1]. Using this edge map, we perform structure enhancement on the original input features based on the broadcasting mechanism, as shown in Equation (3).

$$X = X \odot (1 + \beta E) \tag{3}$$

where β is a hyperparameter that controls the saliency of edge features, and $\odot$ denotes element-wise multiplication. This operation enables regions with rich style line art information to obtain higher responses in the feature map.

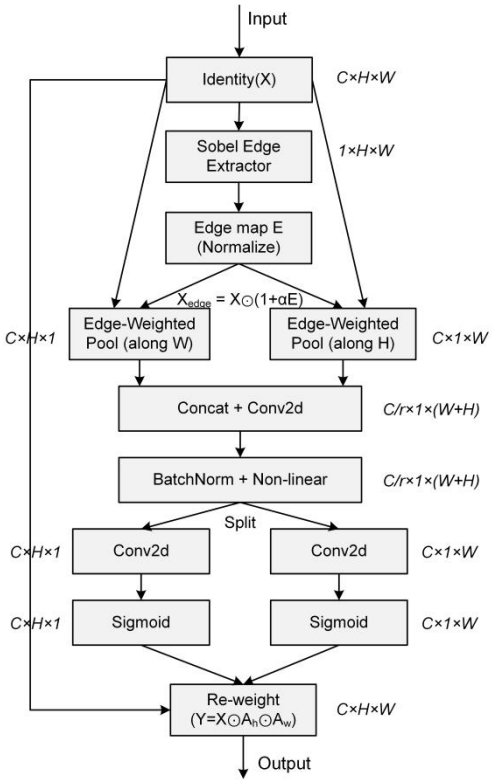


Figure 5. EG-CA network structure.

After obtaining the enhanced feature X_{edge}, EG-CA abandons the traditional one-dimensional average pooling and introduces an edge-weighted pooling strategy. For the height H and width W directions, the spatial feature aggregation is dynamically guided by the edge map E as the weight. The weighted pooling outputs $Z^h \in R^{C\times H\times 1}$ and $Z^w \in R^{C\times 1\times W}$ along the horizontal and vertical directions are defined in Equations (4) and (5).

$$Z^h = \frac{\sum_{w=1}^{W}(X_{edge} \odot E)}{\sum_{w=1}^{W} E + \varepsilon} \tag{4}$$

$$Z^w = \frac{\sum_{h=1}^{H}(X_{edge} \odot E)}{\sum_{h=1}^{H} E + \varepsilon} \tag{5}$$

where ε is a small constant used to avoid division by zero errors. Through this weighting mechanism, high-frequency line art and contour details in anime characters are significantly enhanced and preserved during the feature compression stage.

After completing edge-guided pooling, the module concatenates Z^h and Z^w along the spatial dimension and sends them into a 1×1 convolutional layer for channel dimension reduction with a compression ratio of r. It then passes through batch normalization and the

SiLU nonlinear activation function in sequence to generate an intermediate feature map $f \in R^{C/r \times 1 \times (W+H)}$ containing rich local line details. Next, f is split into two parts along the spatial dimension. The number of channels is restored to C using 1×1 convolutions respectively, and the Sigmoid activation function is applied to generate the final attention weight matrices $A_h \in R^{C \times H \times 1}$ and $A_w \in R^{C \times 1 \times W}$ for spatial coordinate recalibration. Finally, the calculation formula for the output feature map $Y \in R^{C \times H \times W}$ is shown in Equation (6).

$$Y = X \odot A_h \odot A_w \tag{6}$$

By introducing an edge-guided mechanism into the coordinate attention structure, the EG-CA module can effectively enhance the network's perception of structural boundaries and contour information while maintaining the original lightweight characteristics. This edge-enhanced attention modeling method can better capture line features and morphological differences in anime character images, thereby improving the model's ability to distinguish subtle differences between different visual styles and enhancing overall classification performance.

2.2.4. The Adaptive Scale Feature Aggregation

In the fusion stage of multi-scale features, the model needs to dynamically adjust the dependence on shallow line art features and deep semantic features in response to the drastic global color fluctuations caused by the RAS frequency-domain perturbation strategy. Traditional cross-layer feature fusion mostly uses simple concatenation or addition. As indicated in AFF [22]-related research, this static aggregation method ignores the contribution differences of different features in specific contexts and cannot perform dynamic trade-offs according to specific style attributes and rendering interference degrees. To this end, this study proposes the ASFA module, which realizes dynamically weighted fusion of cross-level multi-scale features by evaluating the global context. The structure of ASFA is shown in Figure 6.

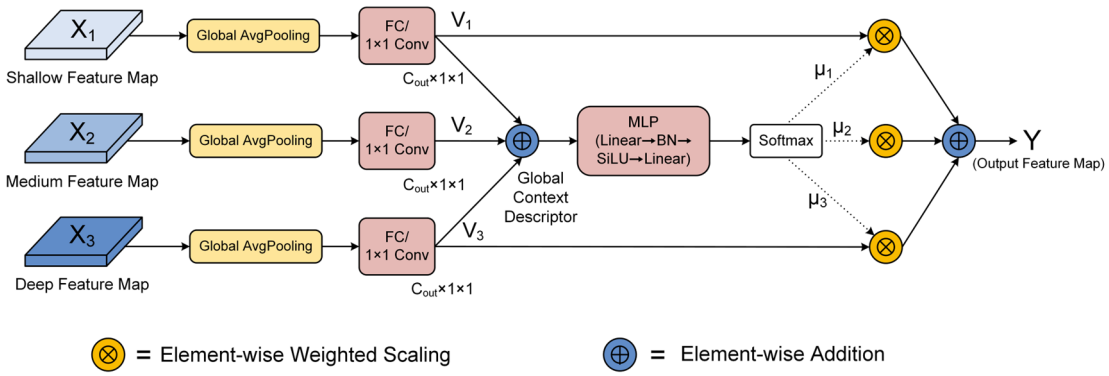


Figure 6. ASFA structure diagram.

Given the multi-scale feature set $\{X_1, X_2, X_3\}$ output by the backbone network at shallow, middle and deep layers, the ASFA module first compresses the spatial dimension of each layer feature into 1×1 channel statistics through two-dimensional global average pooling. A mapping network is then introduced to uniformly align the different channel numbers C_i to the target dimension D. The calculation formula for the aligned feature vector $V_i \in R^D$ is shown in Equation (7).

$$V_i = \delta(B(W_{A_i} \cdot GAP(X_i))) \tag{7}$$

where $W_{A_i} \in R^{D \times C_i}$ denotes the learnable weight of the fully connected layer, $B(\cdot)$ denotes the one-dimensional batch normalization operation, and $\delta(\cdot)$ denotes the SiLU nonlinear activation function.

Next, the dimension-aligned features $\{V_1, V_2, V_3\}$ are added element-wise to construct a global descriptor $G \in R^D$ containing cross-scale information. It is then fed into a lightweight multilayer perceptron with a dimensionality-reduction bottleneck structure. The probabilistic adaptive weight set $\mu = [\mu_1, \mu_2, \mu_3]$ is output via the Softmax function. The calculation processes are shown in Equation (8) and Equation (9) respectively.

$$Z = W_{S_2} \cdot \delta(B(W_{S_1} \cdot G)) \tag{8}$$

$$\mu_i = \frac{\exp(Z_i)}{\sum_{j=1}^{3} \exp(Z_j)} \tag{9}$$

where $W_{S_1} \in R^{\frac{D}{\varphi} \times D}$ is the dimensionality-reduction compression matrix, and φ is the reduction ratio aimed at decreasing the number of parameters and encouraging the network to learn core correlations. $W_{S_2} \in R^{3 \times \frac{D}{\varphi}}$ is the output projection matrix. This operation introduces an inter-scale competition mechanism to ensure $\sum \mu_i = 1$.

Finally, using the learned dynamic weights μ_i, element-wise weighted scaling and aggregation are performed on the features V_i to obtain the output feature descriptor $Y \in R^D$, as shown in Equation (10).

$$Y = \sum_{i=1}^{3} \mu_i V_i \tag{10}$$

With this dynamic aggregation strategy, the network can adaptively adjust cross-layer weights according to the representation requirements of specific art styles even when the color rendering of images is greatly disturbed. This approach breaks the model's excessive dependence on deep global colors, enabling the fused features finally sent to the classifier to achieve an optimal balance between local fine structures and macroscopic geometric shapes. The generalization ability of the model under complex art styles is significantly improved.

2.2.5. Improved EfficientNetV2-B0 Network

In this study, a multi-level improved network architecture based on EfficientNetV2-B0 is proposed to enhance the representation ability of core fine-grained attributes such as topological proportion, edge line art and color rendering in the anime character style classification task. Considering that high-frequency line art structures and low-frequency rendering colors in anime images are often highly coupled, joint optimization is performed at the data input end and the network feature extraction end. The overall architecture is shown in Figure 7.

Firstly, a frequency-domain perturbation strategy named RAS is introduced in the data preprocessing stage. The image is decoupled into magnitude spectrum and phase spectrum using Fast Fourier Transform. Random noise perturbation is applied to the low-frequency magnitude regions that control color and global rendering effects, while the high-frequency phase components that determine line art contours and local structures are preserved. This strategy weakens visual interference caused by single color rendering from the data source, effectively reduces the model's excessive dependence on large color blocks, and encourages the network to focus on geometric contours and low-level topological structures with stronger style discriminability during feature extraction. In the feature extraction stage, EfficientNetV2-B0 is used as the backbone network, and the edge-guided coordinate attention module (EG-CA) is cascaded at the end of each core module. This module jointly models feature importance along horizontal and vertical spatial directions,

accurately locates edge physical distributions across categories, and dynamically assigns higher response weights to key edge features. Thus, it significantly enhances the model's perception of local high-frequency details and spatial topological boundary features while maintaining the efficiency of the backbone network. On this basis, to solve the problem that single deep features easily lose shallow style details, this study breaks the conventional global pooling paradigm, extracts representative multi-scale features from the backbone network, and designs an adaptive scale feature aggregation neck network named ASFA. After spatial compression and dimension alignment of deep features, the ASFA module uses a lightweight multilayer perceptron to construct a global context descriptor, dynamically calculates probabilistic weights of multi-scale features and performs weighted fusion. It realizes effective cross-layer combination of shallow fine-grained geometric textures and deep high-level global semantics.

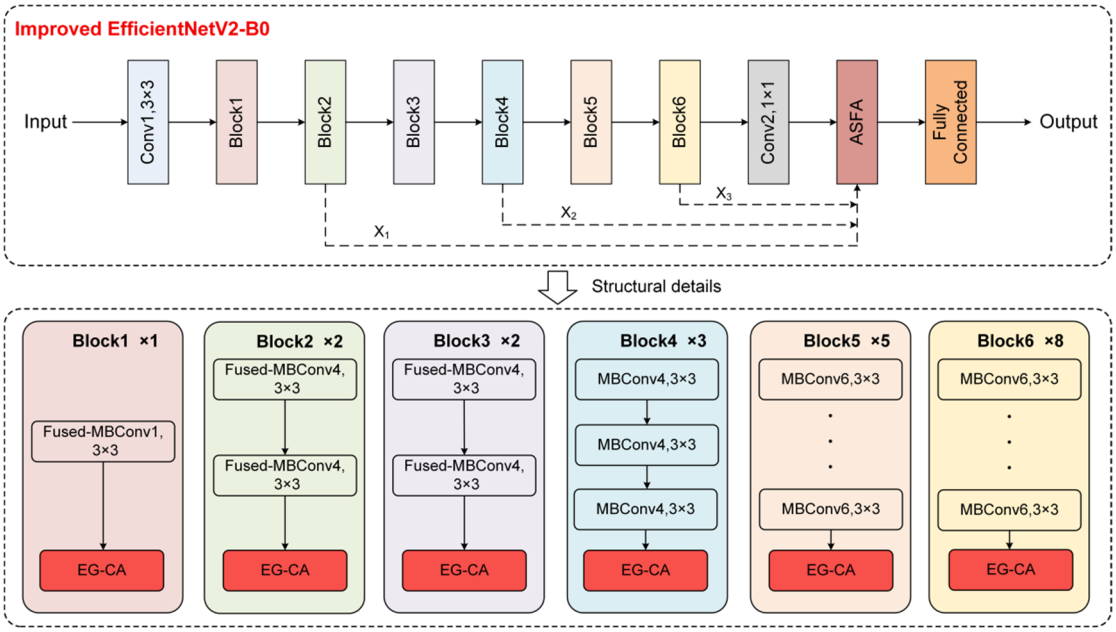


Figure 7. Animation character style classification network.

In summary, the proposed improved EfficientNetV2-B0 model achieves effective decoupling of color rendering and low-level topology through frequency-domain decomposition and perturbation at the data end. It deeply integrates the structure-aware attention mechanism and multi-scale adaptive feature aggregation strategy at the network end. This effectively improves the network's comprehensive modeling ability for fine-grained features of the four typical art styles, thereby providing highly discriminative and robust feature support for fine-grained classification of anime character art styles.

3. Experiment and Result Analysis

3.1. Experimental Environment and Parameter Settings

The experimental platform for this study utilized a 64-bit Windows 11 operating system. The hardware configuration included an Intel(R) Core(TM) i5-11260H 2.6 GHz CPU, an NVIDIA GeForce RTX 3050 12 GB GPU, and 16 GB RAM. The software environment was configured with CUDA version 11.8, Python version 3.8, and PyTorch version 2.0.0.

In terms of parameter settings, the total number of training epochs is set to 150, and the batch size is 8. Stochastic Gradient Descent with momentum is adopted as the optimizer. The initial learning rate is set to 0.01, the momentum coefficient is 0.9, and the weight decay is 1×10^{-4}. A MultiStepLR step decay strategy is used for learning rate

scheduling, with decay performed at the 30th, 60th and 90th epochs where ρ = 0.1, to achieve fine optimization in the late training stage. To enhance model generalization ability and alleviate overfitting, the training stage performs Random Resized Crop on images first, and then adopts the frequency-domain perturbation strategy RAS proposed in this study in series to deeply decouple color rendering and line features. Model parameter initialization is loaded with the weights of EfficientNetV2-B0 pre-trained on the ImageNet dataset to make full use of the general visual feature representations learned by the pre-trained model.

3.2. Evaluation Metrics

To comprehensively evaluate the performance of the proposed anime character style classification model, this study adopts several mainstream evaluation metrics, including overall accuracy, precision, recall and F1-score. Overall accuracy represents the proportion of correctly predicted images among all anime test samples. It is a core metric that intuitively reflects the overall classification performance and generalization level of the model for various art styles. Precision represents the proportion of samples that actually belong to a specific art style among those predicted as the corresponding style by the model. It is used to evaluate the accuracy of the model's judgment result for a single art style. Recall is used to measure the ability of the model to successfully identify all samples that actually belong to a specific art style, reflecting the detection rate of the model for that style. F1-score is the harmonic mean of precision and recall. It comprehensively considers precision and recall, and can more objectively evaluate the overall balance of the classification model in the prediction results of complex art styles. The calculation formula of overall accuracy is shown in Equation (11).

$$Acc = \frac{\sum_{i=1}^{C} TP_{\mathrm{i}}}{N} \tag{11}$$

where Acc denotes overall accuracy, C is the total number of classification categories, TP_i represents the number of true positives for the i-th category, and N is the total number of samples in the test set.

The specific calculation formulas of precision, recall and F1-score are shown in Equations (12)–(14).

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall} \tag{14}$$

where TP denotes the number of samples correctly predicted as positive classes; FP denotes the number of negative class samples misclassified as positive classes; FN denotes the number of positive class samples misclassified as negative classes. TN denotes the number of samples correctly predicted as negative classes.

3.3. Experimental Results

The overall classification accuracy of the proposed method on the test set of four anime character art styles reaches 95.81%, which demonstrates the effectiveness of the improved model in the fine-grained classification task of anime character art styles. As shown in Figure 8, the precision and recall of different categories vary, reflecting the inherent distinguishability of each character art style and the sensitivity of the model to specific facial features, body proportions and rendering techniques. Among them, the realistic art style performs the best, with precision and F1-score as high as 96.97% and 96.73% respectively. This is because this style follows natural anatomical proportions and

has a stable and relatively non-deformable structural basis. In contrast, the precision of the bold art style and the recall of the moe art style are relatively low. This is mainly limited by the inherent difficulties in feature decoupling of the two styles. The high-frequency sharp edges of the bold art style are prone to local topological overlap with other samples with rigid props, leading to misjudgment. As a transitional category between the chibi art style and the realistic art style, the moe art style is prone to ambiguous boundaries of head–body ratio under local close-ups or special perspectives, resulting in missed judgments. The similarity of local line features and visual ambiguity of the overall head–body ratio increase the difficulty for the model to distinguish similar character art styles.

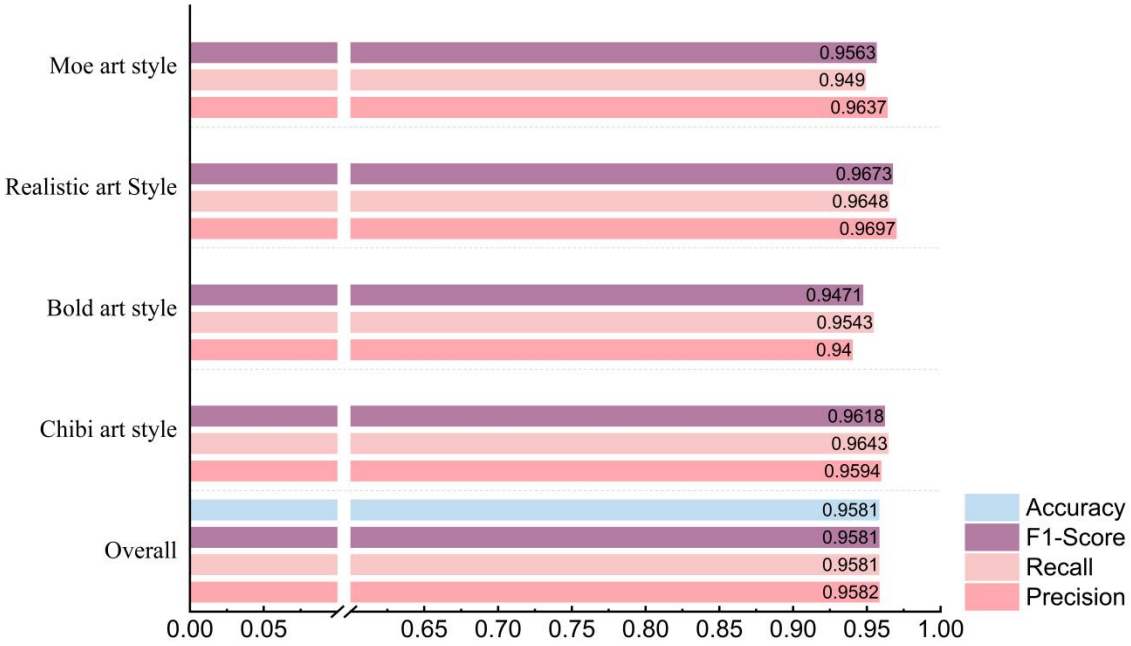


Figure 8. Classification performance metrics for four types of anime character styles.

As shown in Figure 9, the confusion matrix further reflects the classification performance of the model on the test set of four anime character art styles. The confusion matrix exhibits significant diagonal aggregation characteristics, indicating that the proposed method can accurately distinguish most test samples. This strongly proves that the method effectively suppresses superficial visual interferences such as color distribution and light and shadow techniques through frequency-domain perturbation at the data end. Combined with the collaborative modeling of edge-guided coordinate attention and multi-scale feature aggregation at the network end, the comprehensive discrimination ability of the model for the differences in head–body ratio scale span and line contour sharpness of anime characters with different art styles is significantly enhanced.

From the distribution of misclassified samples, there is a small amount of recognition overlap between some categories. For example, there is a small number of mutual misjudgments between the chibi art style and the moe art style. This is mainly attributed to the fact that both styles exhibit relatively compact body proportion characteristics in the overall structure. When the image is a close-up of the character's face or in the transition range of 3 to 4 head heights, the scale features of the two styles tend to overlap, increasing the difficulty for the model to extract discriminative features. At the same time, a small amount of confusion also occurs between the realistic art style and the bold art style. This is mainly due to their high overlap in overall human body proportions. A small number of realistic samples exhibit rigid edge features in local composition, which shortens the inter-class distance from the bold art style in the feature space, resulting in a small amount

of cross-class confusion. Notably, the realistic art style achieves the highest recognition accuracy among the four art styles in the test set. This benefits from its proportion benchmark following natural anatomy, with high intra-class consistency and structural stability in human skeletal topology and limb scales, thus forming a clear and stable distribution region in the feature space.

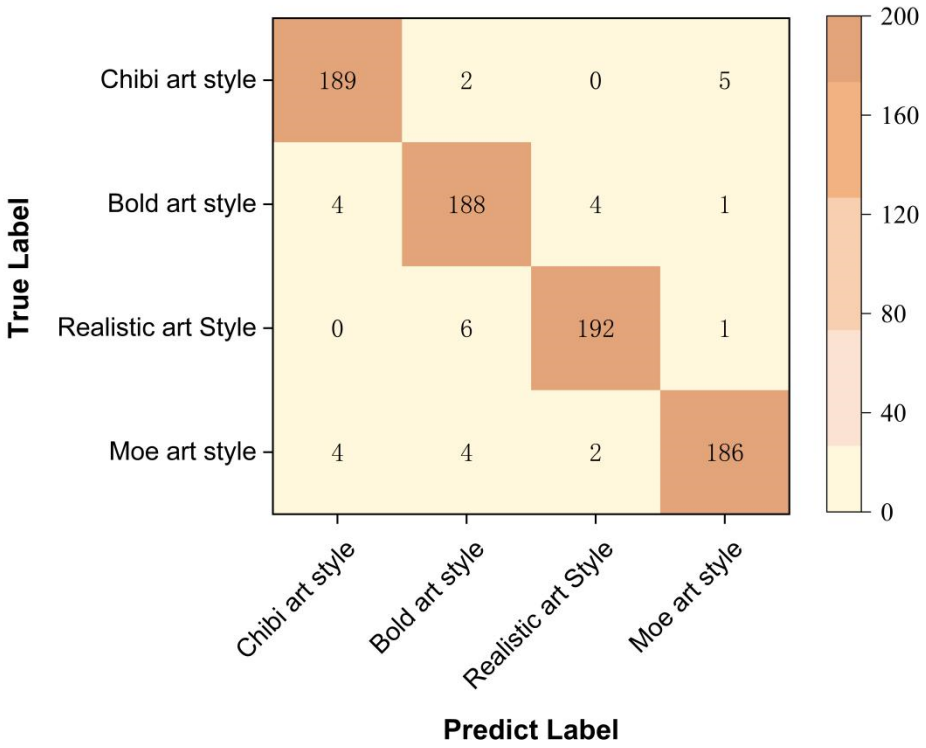


Figure 9. Confusion matrix.

Overall, the analysis results of the confusion matrix show that the proposed method achieves excellent recognition performance in the classification of the four anime character art styles. The small number of misclassifications is mainly reflected in the overlap of human body proportion boundaries and the natural similarity of local topological features between different art styles. This also provides a valuable reference direction for the subsequent introduction of more refined local feature decoupling mechanisms or anti-confusion learning for cross-scale structures.

3.4. Ablation Experiment

To verify the effectiveness of the introduced frequency-domain perturbation strategy (RAS), edge-guided coordinate attention mechanism (EG-CA), and adaptive scale feature aggregation neck network (ASFA) in the fine-grained anime character style classification task, this study conducts systematic ablation experiments on the test set. Quantitative evaluation results regarding classification performance and computational efficiency for each module combination are shown in Table 2 and Table 3, respectively.

Firstly, when the original EfficientNetV2-B0 is directly used as the baseline model, its overall classification accuracy is 89.75%. This basic performance indicates that relying only on spatial-domain deep semantic features extracted by the conventional backbone network and the single global average pooling mechanism still cannot overcome the complex color and light–shadow rendering interference in anime images, nor can it effectively decouple significant multi-scale structural differences and local topological deformation features among different character art styles. From the perspective of computational efficiency, the

parameter quantity of the baseline model is 7.14 M, the computational complexity is 0.728 G FLOPs, the inference speed reaches 85.97 FPS, and the memory consumption is 800 MB.

Table 2. Results of the ablation experiment.

Model	Accuracy	Precision	Recall	F1
EfficientNetV2-B0	0.8975	0.9023	0.8925	0.8974
+RAS	0.9015	0.8982	0.9056	0.9019
+EG-CA	0.9185	0.9131	0.9234	0.9182
+ASFA	0.9096	0.9056	0.9135	0.9095
+RAS+EG-CA	0.9355	0.9314	0.9392	0.9353
+RAS+ASFA	0.9342	0.9385	0.9301	0.9343
+EG-CA+ASFA	0.9465	0.9441	0.9492	0.9466
+RAS+EG-CA +ASFA	0.9581	0.9582	0.9581	0.9581

Table 3. Computational efficiency of the ablation experiment.

Model	Params (M)	FLOPs (G)	GPU Mem (MB)	FPS (img/s)
EfficientNetV2-B0	7.14	0.728	800	85.97
+RAS	7.22	0.749	810	83.48
+EG-CA	7.54	0.883	819	75.94
+ASFA	7.36	0.791	813	81.33
+RAS+EG-CA	7.62	0.904	830	73.45
+RAS+ASFA	7.44	0.812	821	78.83
+EG-CA+ASFA	7.76	0.946	835	71.33
+RAS+EG-CA +ASFA	7.84	0.967	840	68.83

On this basis, after introducing EG-CA, the model accuracy is significantly improved to 91.85%. This result shows that EG-CA can effectively strengthen the edge coordinate features that play a key role in style discrimination and enhance the network's ability to respond to character edge contours, line curvatures and local geometric topology, thus improving the discrimination stability of fine-grained features. In terms of computational overhead, the parameter quantity of EG-CA is 7.54 M and the computational complexity is 0.883 G FLOPs. It demonstrates that the edge-enhanced spatial attention strategy does not introduce additional computational overhead while ensuring the lightweight property of the model. In contrast, when only RAS is introduced at the data end, the accuracy increases to 90.15%. This indicates that although simple frequency-domain decoupling can suppress the interference caused by surface color rendering to a certain extent and force the network to focus on low-level high-frequency structures, it is difficult for the model to construct a highly discriminative feature decision boundary without explicit attention modeling of key structural regions inside the network. In addition, since the RAS strategy mainly performs frequency-domain perturbation at the data level on the input end, its impact on model complexity is almost negligible. The number of parameters remains at 7.22 M and achieves a real-time inference performance of 83.48 FPS. Similarly, when only ASFA is used, the accuracy is 90.96%. This shows that although relying solely on multi-scale feature fusion can capture the span information of different character art styles at the macro scale, the model's ability to depict complex style boundaries is still insufficient when the input features have not undergone edge coordinate enhancement by EG-CA and frequency-domain anti-interference processing by RAS. The introduction of ASFA increases the parameter quantity to 7.36 M and the computational complexity to 0.791 G FLOPs, while maintaining an efficient inference performance of 81.33 FPS.

Further results of combination experiments show that each module presents obvious synergistic and complementary characteristics. When EG-CA is combined with ASFA, the accuracy increases to 94.65%. At this time, the model parameter quantity is 7.76 M and the computational complexity is 0.946 G FLOPs. While maintaining an inference speed of 71.33 FPS, this combination indicates that spatial redundant information in feature maps at all levels is effectively suppressed under the guidance of EG-CA, and the response of key structures is enhanced. At this stage, the multi-scale aggregation of ASFA can more efficiently fuse shallow local geometric edges with deep global scale semantics. Similarly, the combination of RAS and EG-CA achieves an accuracy of 93.55% with a parameter quantity of 7.62 M and an inference speed of 73.45 FPS. It shows that after removing surface information through frequency-domain processing, high-frequency structural data exhibits stronger discriminative power under the further enhancement of the coordinate attention mechanism. In addition, the combination of RAS and ASFA reaches an accuracy of 93.42% with a parameter quantity of 7.44 M and a computational complexity of 0.812 G FLOPs. This indicates that frequency-domain decoupling and multi-scale fusion can indeed provide complementary information. However, in the absence of edge guidance and coordinate attention mechanism, the network still lacks the ability to focus on local geometric contours and key line-draft features.

When the three modules of RAS, EG-CA and ASFA are introduced simultaneously, the model achieves the best performance with an overall classification accuracy of 95.81%. The experimental results show that the above three strategies form closed-loop collaboration in three dimensions including frequency-domain anti-interference at the data end, spatial structure enhancement at the feature extraction end, and multi-scale semantic fusion at the neck stage. It is worth noting that although the proposed model integrates multiple functional modules, its parameter quantity is only 7.84 M with a computational cost of 0.967 G FLOPs. It still maintains an inference speed of 68.83 FPS, which fully verifies that the proposed method achieves an effective balance between high-precision classification and real-time inference performance.

Parameter Sensitivity Analysis of RAS

To further verify the stability and effectiveness of the proposed RAS module in the anime character style classification task, this study conducts a systematic sensitivity analysis on two key hyperparameters that control frequency-domain perturbation. The parameters include the low-frequency mask radius ratio r and perturbation intensity α. The parameter r is used to control the range of the low-frequency region perturbed in the Fourier spectrum. The parameter α determines the fluctuation amplitude of the random scaling factor. Taking EfficientNetV2-B0 combined with the RAS module as the basic architecture, this study independently evaluates the influence of each parameter on the overall classification performance of the model while keeping other hyperparameters fixed.

With α fixed at 0.6, the classification performance corresponding to different low-frequency mask radius ratios r is presented in Table 4. When r is set to 0.05, the low-frequency perturbation area is too small to effectively destroy the overall color distribution and light and shadow rendering features in anime images. The model accuracy is only 88.52%. When r increases to 0.10 and 0.15, the model performance improves gradually and reaches the optimal value at $r = 0.15$ with an accuracy of 90.15%. This indicates that appropriately expanding the low-frequency perturbation range can more fully weaken the interference caused by global color rendering. Meanwhile, it can completely retain high-frequency line contours and local topological structures, thereby encouraging the model to pay more attention to geometric edges and structural features with style discrimination capability. However, the classification performance decreases significantly when r increases

to 0.20 and 0.25. It shows that excessive low-frequency perturbation will weaken the overall semantic consistency and global spatial distribution stability of images, and further affect the model's ability to extract effective discriminative features.

Table 4. Classification performance under different low-frequency mask radius ratios (r) at a fixed perturbation intensity (α = 0.6).

Mask Radius Ratio (r)	Accuracy	Precision	Recall	F1
0.05	0.8852	0.8835	0.8861	0.8848
0.10	0.8965	0.8941	0.8982	0.8961
0.15 (Ours)	0.9015	0.8982	0.9056	0.9019
0.20	0.8912	0.8895	0.8931	0.8913
0.25	0.8775	0.8752	0.8794	0.8773

With r fixed at 0.15, the classification performance corresponding to different perturbation intensities α is shown in Table 5. When α is set to 0.2, the amplitude of frequency-domain perturbation is weak. The model's dependence on color rendering information is not effectively suppressed, and the classification accuracy is 88.85% at this time. As the perturbation intensity increases to 0.4 and 0.6, the model performance continues to improve and achieves the best effect at α = 0.6 with an accuracy of 90.15%. This demonstrates that moderate random amplitude scaling can enhance the randomness of low-frequency color distribution and enable the network to learn more stable and discriminative high-frequency structural features. When α further increases to 0.8 and 1.0, the model performance begins to decline. Particularly, the accuracy is only 87.42% when α is set to 1.0. The main reason is that excessively strong low-frequency perturbation will cause obvious color distortion and spectrum noise in the reconstructed image. It will affect the overall semantic consistency of images and the stability of local structure perception, and eventually lead to the degradation of model performance.

Table 5. Classification performance under different perturbation intensities (α) at a fixed mask radius ratio (r = 0.15).

Perturbation Intensity (α)	Accuracy	Precision	Recall	F1
0.2	0.8885	0.8862	0.8904	0.8883
0.4	0.8972	0.8954	0.8985	0.8969
0.6 (Ours)	0.9015	0.8982	0.9056	0.9019
0.8	0.8920	0.8891	0.8945	0.8918
1.0	0.8742	0.8718	0.8755	0.8736

In summary, the RAS module exhibits obvious parameter sensitivity characteristics. An appropriate low-frequency mask range and perturbation intensity can effectively suppress the model's shortcut learning relying on color statistical information while retaining high-frequency line structural features to the greatest extent. Either excessively weak or excessively strong perturbation will degrade model performance. The parameter combination of r = 0.15 and α = 0.6 achieves the optimal balance between color rendering interference suppression and structural semantic maintenance, and is therefore determined as the final experimental configuration in this paper.

3.5. Heat Map Analysis

To verify the focusing ability of the proposed model on key discriminative features in the anime character art style classification task, this study uses the Grad-CAM method to generate heatmaps of the four anime character art styles [23,24], as shown in Figure 10. The

heatmaps intuitively show the visual attention mechanism behind the network decisions. The red highlighted regions represent the features that the model considers most important for style discrimination, while the blue regions indicate the background or irrelevant regions with small contributions.



Figure 10. Heat map comparison of four types of anime character styles.

By comparing the heatmap visualization results of the four typical anime character art styles, the baseline model shows an obvious tendency of shortcut learning. The activation regions are easily disturbed by background noise, large soft clothing areas or high-saturation color blocks, showing a diffuse block distribution and losing feature response to the geometric contours and low-level physical structures of characters. In contrast, the proposed method exhibits strong anti-interference and feature decoupling capabilities, and can dynamically modulate feature responses according to the characteristics of different character art styles. For the chibi art style and moe art style, the model effectively suppresses global dispersion, and the highlighted regions compactly converge on high-discriminative advanced semantic points such as core facial features, head contours and specific gestures. For the realistic art tyle, the network effectively strips the interference of high-saturation color blocks and accurately focuses on details such as the nasal bridge, facial contours, hands and fan rib structures. For the bold art style, the model shows significant physical edge convergence characteristics. It not only extracts key facial topological structures, but also naturally conforms its feature responses to the geometric trends of foreground props. This fully demonstrates that the proposed method is robust to surface color rendering and background noise, and can achieve complete representation of the core structures and semantics of different art styles.

In conclusion, the heatmap analysis fully validates the effectiveness of the proposed method and reveals that the model forms a feature response pattern consistent with the core semantics of art styles during the recognition of different anime styles. The visualization results show that the model tends to focus on the lines and semantic parts with the richest art style information in anime character images, avoiding excessive response to flat colors or irrelevant backgrounds. This not only effectively verifies the strong feature decoupling ability of the model, but also enhances the interpretability of its outputs.

3.6. Comparison with Other Classification Models

To comprehensively evaluate the performance advantages of the proposed method in the anime character art style classification task, we compare the improved model

with several mainstream classification models such as ResNet50 [25], ShuffleNet [26], ConvNeXt [27], DenseNet [28], MobileNetV2 [29] and MViTv2-T [30] under the condition that the dataset and software–hardware configuration remain completely consistent. The classification performance and computational efficiency results are presented in Table 6 and Table 7, respectively.

Table 6. The results of the comparative experiment.

Model	Accuracy	Precision	Recall	F1
ResNet50	0.8741	0.8682	0.8795	0.8738
ShuffleNet	0.8375	0.8465	0.8291	0.8377
ConvNeXt	0.8492	0.8425	0.8566	0.8495
DenseNet	0.8642	0.8715	0.8562	0.8638
MobileNetV2	0.8521	0.8453	0.8598	0.8525
MViTv2-T	0.8782	0.8745	0.8821	0.8783
Ours	0.9581	0.9582	0.9581	0.9581

Table 7. Computational efficiency of the comparative experiment.

Model	Params (M)	FLOPs (G)	GPU Mem (MB)	FPS (img/s)
ResNet50	23.2	4.243	1285	85.15
ShuffleNet	2.3	0.156	420	215.3
ConvNeXt	28.6	4.51	1350	76.45
DenseNet	8.42	2.92	1180	65.2
MobileNetV2	3.5	0.319	620	168.6
MViTv2-T	24.2	4.717	1520	52.8
Ours	7.84	0.967	840	68.83

The classic deep convolutional neural networks ResNet50 and DenseNet achieve a classification accuracy of 87.41% and 86.42%, respectively. ResNet50 can reach an inference speed of 85.15 FPS, while its parameter count, computational complexity, and memory occupation are as high as 23.2 M, 4.243 G FLOPs, and 1285 MB. DenseNet has a relatively small parameter scale of 8.42 M, but its inference speed drops to 65.2 FPS, and the memory occupation still reaches 1180 MB. This indicates that traditional deep CNN models rely heavily on the stacking of deep semantic features. When dealing with highly similar facial proportion and complex color backgrounds in anime images, they lack the fine-grained modeling capability for local line-draft features. Lightweight models such as ShuffleNet and MobileNetV2 possess high computational efficiency, but their accuracy only reaches 83.75% and 85.21%. ShuffleNet achieves an inference speed of 215.3 FPS with a parameter count of only 2.3 M and a computational cost of 0.156 G FLOPs. MobileNetV2 obtains an inference speed of 168.6 FPS under the conditions of 3.5 M parameters and 0.319 G FLOPs. It demonstrates that lightweight structures often fail to fully retain fine-grained features such as the facial topological structure and light and shadow gradient of anime characters in the process of pursuing computational efficiency and parameter compression, which further affects the classification performance of the model. ConvNeXt and MViTv2-T, as representatives of modern network architectures, exhibit stronger feature modeling capability with accuracy rates of 84.92% and 87.82%. The results show that the large receptive field design of CNN and the multi-scale global attention mechanism of Transformer have certain advantages in feature capture. However, such models are usually accompanied by high computational overhead. The parameter count and computational complexity of ConvNeXt are up to 28.6 M and 4.51 G FLOPs. MViTv2-T has a parameter count of 24.2 M and a computational complexity of 4.717 G FLOPs, with the memory

occupation reaching 1520 MB and the inference speed only 52.8 FPS. Although these models have strong global modeling capability, it is still difficult to form clear class decision boundaries when facing highly similar visual features in the anime style classification task.

The proposed method in this study achieves the optimal balance between classification performance and computational efficiency. Its accuracy, precision, recall, and F1-score reach 95.81%, 95.82%, 95.81%, and 95.81% respectively, which are all superior to the comparison models. Meanwhile, the proposed model maintains an inference speed of 68.83 FPS with a parameter count of 7.84 M, a computational complexity of 0.967 G FLOPs, and a GPU memory occupation of 840 MB. Compared with MViTv2-T which has the best performance among existing baseline models, the proposed method not only achieves a significant improvement in classification accuracy but also greatly reduces the parameter count and computational complexity. This result fully verifies the effectiveness of the customized modules in this study for character style recognition.

In addition, to further evaluate the convergence stability and robustness of the model, Figure 11 shows the distribution of mean values and standard deviations (error bars) of each classification model under five independent repeated experiments. It can be intuitively observed from the dispersion degree of the error bars that the proposed method not only outperforms other mainstream classification networks significantly in performance, but also has the tightest error bar range representing metric fluctuation. This strongly proves that the network effectively suppresses the uncertainty of the decision boundary under the deep collaboration of frequency-domain decoupling, edge coordinate prior and adaptive multi-scale fusion. This result fully verifies the stability of the model and further demonstrates its strong adaptability and practical application potential in complex fine-grained classification tasks such as anime character art style classification.

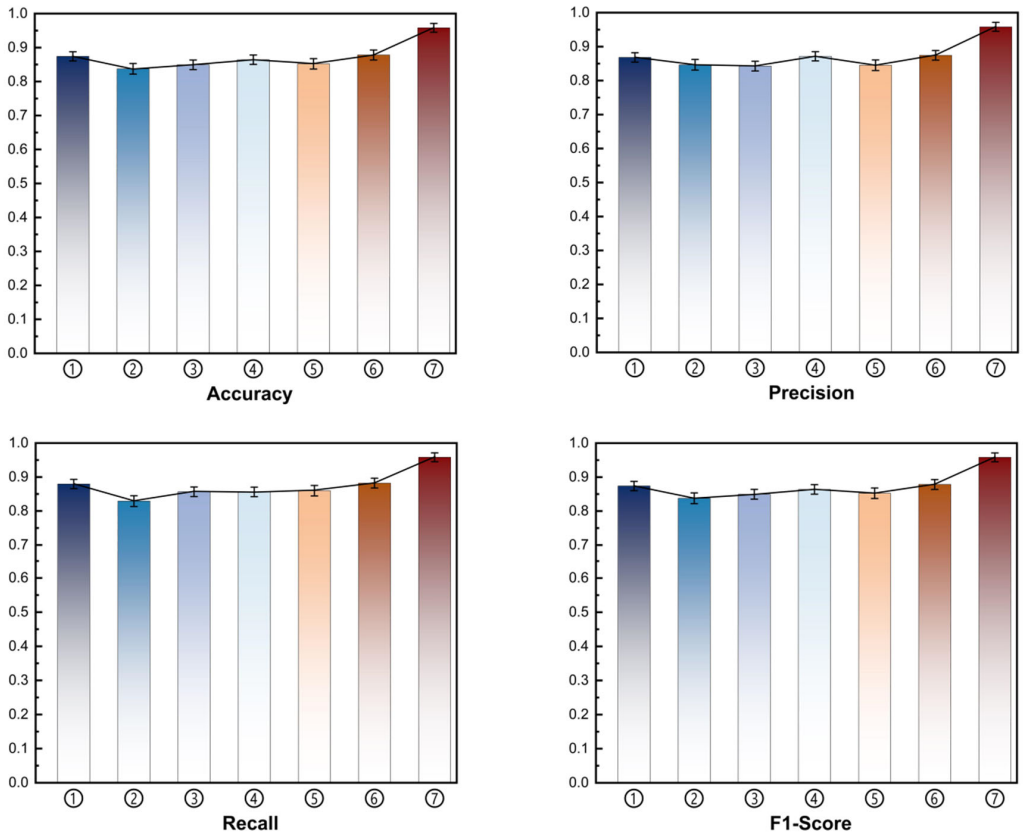


Figure 11. Performance comparison of different algorithms (numbers ① to ⑦ represent ResNet50, ShuffleNet, ConvNeXt, DenseNet, MobileNetV2, MViTv2-T, and the proposed model, respectively.).

4. Discussion

From the comprehensive analysis results of various evaluation metrics and heatmaps, it can be found that the improved EfficientNetV2-B0 model proposed in this study not only exhibits strong discrimination stability in the fine-grained anime character art style classification task, but also significantly enhances the interpretability of network decisions. The visualization results of heatmaps show that the model successfully overcomes the defect of traditional networks relying on low-frequency color blocks for shortcut learning, and forms a discriminative feature response pattern focusing on the key art style contours of anime characters. Specifically, at the data input end, RAS decouples high and low-frequency information, suppresses redundant interference of light, shadow and color from the source, and effectively reduces the network's excessive dependence on large-area color rendering features. At the feature extraction end, EG-CA uses the Sobel operator to extract physical edge weights, realizes explicit modeling of highly discriminative local coordinate priors, optimizes the model's attention allocation on edge topology and line-draft structures, and significantly enhances the model's feature representation ability for local spatial structures and key line-drawing contours. At the multi-scale fusion end, ASFA uses MLP to dynamically calculate scale-adaptive weights and realizes efficient cross-layer fusion of shallow local line-draft features and deep global art style semantics. This multi-dimensional feature modeling strategy enables the model to obtain stable and consistent discrimination basis when dealing with four types of anime character art styles with highly overlapping visual features.

Although the experimental results verify the effectiveness and robustness of the proposed method, it still has certain limitations in feature modeling and application scope. Firstly, the dataset constructed in this experiment is relatively limited in scale and only contains four representative artistic styles of anime characters. Although these four styles cover core structural features such as head-to-body ratio, line abstraction degree and rendering realism degree, they do not include niche styles, mixed styles and highly personalized creative styles of illustrators. The generalization ability of the model in cross-media and mixed style scenarios remains to be further verified. Secondly, the RAS module realizes the explicit decoupling of structural features by suppressing low-frequency color information. This processing method weakens the representation ability of the model for painting styles dominated by color distribution and complex light and shadow such as cyberpunk or gradient style. For such painting styles, color itself is the key discriminative feature. The suppression of low-frequency amplitude components may destroy critical visual clues and further limit the general applicability of this strategy. In addition, EG-CA exhibits strong representation capability for human facial feature proportions, line art contours and geometric boundaries by enhancing spatial structure prior. However, this mechanism focusing on extracting structural features essentially limits its application boundary. When dealing with non-structural style features that rely highly on light and shadow rendering, texture features and color layers, the modeling capability of this module is still insufficient. Finally, while maintaining lightweight characteristics, the ASFA module achieves multi-scale feature balance by fusing shallow, middle and deep layer features. Nevertheless, the existing multi-scale fusion mechanism still has limitations when coping with extremely subtle painting style differences. Without fine-grained layer-by-layer feature sampling, the model may face the risk of local detail loss when processing highly coupled complex details.

In view of the above shortcomings, future research can be further carried out in multiple dimensions. On the one hand, we will continuously expand the anime character style dataset covering niche, hybrid, and cross-media anime styles to enhance the model's representative capacity, while integrating domain generalization and style adaptation techniques to improve robustness against feature shifts in highly personalized artworks.

On the other hand, for the confusion of similar categories such as chibi art style and moe art style in local facial features, subsequent research can further explore the construction of a dual-branch feature extraction architecture based on structural priors and color semantics to dynamically adapt to the weight requirements of different art styles for spatial geometry and color rendering. In addition, for the possible feature overlap between bold art style and realistic art style, the introduction of a cross-domain style adversarial learning mechanism can be attempted to further clarify the classification decision boundary between highly similar art styles by optimizing the inter-class feature distance. Finally, on the premise of maintaining high classification accuracy, combined with technologies such as model pruning and knowledge distillation, lightweight deployment optimization is performed on the improved EfficientNetV2-B0 network, so as to exert greater technical value in practical application scenarios such as structured management of anime images, intelligent art style retrieval and copyright traceability.

5. Conclusions

Using the improved EfficientNetV2-B0 network, this study proposes an anime character art style classification method based on multi-dimensional feature collaborative modeling. This method introduces RAS in the data preprocessing stage to decouple high- and low-frequency information through random mask scaling on the low-frequency regions of the magnitude spectrum, suppresses redundant interference of global color rendering from the source, and thus effectively retains high-frequency line art features. In the feature extraction stage, the network cascades EG-CA, which uses physical edge weights to strengthen the network's ability to respond to and explicitly model local spatial structures and contour lines. Meanwhile, ASFA is designed in the neck network to dynamically weigh and fuse shallow fine-grained textures and deep global semantics using a lightweight multilayer perceptron. Finally, an overall accuracy of 95.81% is achieved on the dataset containing four types of anime character art styles, while maintaining a highly efficient inference speed of 68.83 FPS and a lightweight parameter scale of 7.84 M.

Experimental results show that the proposed method outperforms mainstream image classification models such as ResNet50, ShuffleNet, ConvNeXt, DenseNet, MobileNetV2, and MViTv2-T in comprehensive evaluation metrics including overall accuracy, precision, recall, and F1-score. Especially when dealing with art style categories with highly overlapping inter-class features, the model exhibits stronger discrimination stability and inter-class feature decoupling ability. Ablation experiments further verify the complementary effects of RAS, EG-CA, and ASFA at different levels including data anti-interference, spatial structure enhancement, and dynamic multi-scale feature fusion. Their combined application can effectively alleviate the cross-misjudgment between easily confused art style categories.

In summary, the proposed method effectively strips the redundant interference of complex color rendering and significantly enhances the network's ability to collaboratively model local edge priors, high-frequency line art details, and cross-scale global semantics. This research provides feasible and efficient technical support for structured management of anime images, intelligent art style retrieval, and content understanding. It has certain theoretical reference and promotion value in practical application scenarios such as the digital cultural industry, virtual content generation, and copyright traceability.

Author Contributions: Conceptualization, Y.C., J.D. and H.W.; Methodology, Y.C., J.D. and Z.D.; Software, Y.C. and Z.D.; Validation, H.W.; Resources, Z.D.; Data curation, Y.C.; Writing–original draft, J.D.; Supervision, J.D., H.W. and Z.D.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Henan Provincial Finance Department and Communist Party of China Publicity Department of Henan Province (No. 2024BYS019). And The APC was funded by (No. 2024BYS019).

Data Availability Statement: The anime character dataset used in this study is publicly available at https://aistudio.baidu.com/datasetdetail/237153 (accessed on 10 May 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Li, R.; Oji, R.; Fujishiro, I. Controllable automatic generation of non-player characters in 3D anime style. *Comput. Animat. Virtual Worlds* **2023**, *34*, e2047. [CrossRef]
2. Li, B.; Zhu, Y.; Wang, Y.; Lin, C.-W.; Ghanem, B.; Shen, L. AniGAN: Style-guided generative adversarial networks for unsupervised anime face generation. *IEEE Trans. Multimed.* **2021**, *24*, 4077–4091. [CrossRef]
3. Kathiravan, M.; Ramesh, S.; Buvanesvari, R.; Sreesubha, S.; Irumporai, A.; Venkataramana, R. Generative Adversarial Network Based Advanced Framework for High Resolution Multi-Style Cartoon Image Synthesis in Animation and Game Development. In Proceedings of the 2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET), Pune, India, 16–17 January 2025; pp. 1–5. [CrossRef]
4. Lan, Z.; Maeda, K.; Ogawa, T.; Haseyama, M. Multi-label classification in anime illustrations based on hierarchical attribute relationships. *Sensors* **2023**, *23*, 4798. [CrossRef] [PubMed]
5. Men, Y.; Yao, Y.; Cui, M.; Lian, Z.; Xie, X.; Hua, X.-S. Unpaired Cartoon Image Synthesis via Gated Cycle Mapping. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LO, USA, 19–24 June 2022; pp. 3501–3510. [CrossRef]
6. Wang, C.; Ma, W.; Zhang, S.; Gui, J.; Li, Q.; Liu, Y.; Xia, Z. Focus on finding deepfakes: A robust proactive detection method based on orthogonal moment watermarking. *IEEE Trans. Image Process.* **2026**, *35*, 3507–3521. [CrossRef]
7. Wei, X.-S.; Song, Y.-Z.; Mac Aodha, O.; Wu, J.; Peng, Y.; Tang, J.; Yang, J.; Belongie, S. Fine-grained image analysis with deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8927–8948. [CrossRef] [PubMed]
8. Wei, X.S.; Wu, J.; Cui, Q. Deep learning for fine-grained image analysis: A survey. *arXiv* **2019**, arXiv:1907.03069. [CrossRef] [PubMed]
9. Ruta, D.; Gilbert, A.; Aggarwal, P.; Marri, N.; Kale, A.; Briggs, J.; Speed, C.; Jin, H.; Faieta, B.; Filipkowski, A.; et al. StyleBabel: Artistic Style Tagging and Captioning. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2022; pp. 219–236. [CrossRef]
10. Lan, Z.; Maeda, K.; Ogawa, T.; Haseyama, M. GCN-Based Multi-Modal Multi-Label Attribute Classification in Anime Illustration Using Domain-Specific Semantic Features. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 2021–2025. [CrossRef]
11. Xiang, S.; Li, H. Disentangling style and content in anime illustrations. *arXiv* **2019**, arXiv:1905.10742. [CrossRef]
12. Young-Min, K. Feature visualization in comic artist classification using deep neural networks. *J. Big Data* **2019**, *6*, 56. [CrossRef]
13. Xu, C.; Xu, Y.; Zhang, H.; Xu, X.; He, S. DreamAnime: Learning style-identity textual disentanglement for anime and beyond. *IEEE Trans. Vis. Comput. Graph.* **2024**, *31*, 4198–4209. [CrossRef] [PubMed]
14. Li, H.; Guo, S.; Lyu, K.; Yang, X.; Chen, T.; Zhu, J.; Zeng, H. A Challenging Benchmark of Anime Style Recognition. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LO, USA, 19–24 June 2022; pp. 4721–4730. [CrossRef]
15. Tan, M.; Le, Q. Efficientnetv2: Smaller Models and Faster Training. In Proceedings of the International Conference on Machine Learning (PMLR), Online, 18–24 July 2021; pp. 10096–10106. [CrossRef]
16. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
17. Tan, M.; Le, Q. Efficientnet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the International Conference on Machine Learning (PMLR), Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114. [CrossRef]
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141. [CrossRef]
19. Chattopadhyay, P.; Sarangmath, K.; Vijaykumar, V.; Hoffman, J. Pasta: Proportional Amplitude Spectrum Training Augmentation for Syn-to-Real Domain Generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 19288–19300. [CrossRef]
20. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Online, 19–25 June 2021; pp. 13713–13722. [CrossRef]

21. Han, L.; Tian, Y.; Qi, Q. Research on edge detection algorithm based on improved sobel operator. *MATEC Web Conf.* **2020**, *309*, 03031. [CrossRef]
22. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional Feature Fusion. In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), Online, 5–9 January 2021; pp. 3560–3569. [CrossRef]
23. Chattopadhay, A.; Sarkar, A.; Howlader, P.; Balasubramanian, V.N. Grad-cam++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 839–847. [CrossRef]
24. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE on compute Rvision and Pattern Recognition, Denver, CO, USA, 3–7 June 2016; pp. 770–778. [CrossRef]
26. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6848–6856. [CrossRef]
27. Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986. [CrossRef]
28. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708. [CrossRef]
29. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L. MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520. [CrossRef]
30. Fan, Q.; Xiong, Y.; Mangalam, K.; Li, J.; Yan, S.; Malik, J.; Feichtenhofer, C. MViTv2: Improved multiscale vision transformers for classification and detection. *arXiv* **2021**, arXiv:2112.01526. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.