

Application and Optimization Strategies of the Improved ResNet50 Model in Image Aesthetic Scoring

Peiling Wu *

School of Engineering, University of Connecticut, Storrs, Connecticut, CT06269, The United States of America

* Corresponding Author Email: peiling.wu@uconn.edu

Abstract. This research introduces an image aesthetics evaluation model designed to enhance the accuracy and robustness of image aesthetics scoring by modifying the pre-trained ResNet50 model. The experimental findings indicate that this model surpasses other prevalent convolutional neural network architectures and optimizer combinations. Additionally, this paper examines the optimal configuration of ResNet50 combined with the Adam optimization algorithm with a learning rate set at 0.001. A thorough analysis reveals that the residual block design of ResNet50 is particularly effective in managing deep features, while the Adam optimizer offers substantial benefits in maintaining stable convergence and efficient learning. The outcomes suggest that this model has extensive application potential in various scenarios, including social media platforms, image retrieval and search engines, photography training and automated design, advertising creation and content generation, smart albums and image management, and artistic creation and review. In these applications, the model can substantially improve the precision and performance of image recommendation, retrieval, classification, and evaluation, resulting in a superior user experience.

Keywords: Deep learning; aesthetics evaluation; convolutional neural network.

1. Introduction

Image aesthetics evaluation refers to the quantitative or qualitative evaluation of the aesthetic quality of images. The core of aesthetics evaluation lies in analyzing the visual elements and aesthetic features of an image to provide an aesthetic score. This research field has garnered extensive attention in recent years, not only because of its significance in academic research but also because of its broad potential in practical applications. Applications of image aesthetics evaluation span various domains, including image retrieval, image editing, and content recommendation on social media. For instance, platforms like Flickr and Instagram can use aesthetics evaluation models to recommend high-quality images, thereby enhancing user experience. Additionally, aesthetics evaluation can be applied in photography training, automated design, and advertising creation, helping users to create more visually appealing content.

Although traditional methods for image aesthetics evaluation have achieved some success, these methods typically rely on manually designed feature extraction and rule-based approaches, which suffer from poor adaptability and weak generalization capabilities. The swift advancement of artificial intelligence technology has introduced deep learning as a novel approach for evaluating image aesthetics. Deep learning, particularly with the advancements in convolutional neural networks (CNNs), has shown great potential. This paper will further explore the characteristics, advantages, and applications of deep learning in aesthetics evaluation.

Deep learning, especially CNNs, has demonstrated tremendous potential in image processing and analysis [1]. Firstly, deep learning models have the capability of automatic feature extraction, enabling them to learn high-level features of images through multi-layer neural network structures. This makes them perform excellently in tasks including image recognition, classification, and generation. Secondly, deep learning models possess efficient nonlinear mapping abilities, capturing complex data patterns through multiple layers of nonlinear transformations, and representing various details and global features in images. This results in outstanding efficacy in activities like image categorization, detecting objects, and creating images. Additionally, deep learning models have the

ability to handle large-scale data. Modern deep learning frameworks and hardware make it possible to train large neural network models. Training on large datasets allows these models to learn more generalized features, thereby improving robustness and accuracy. The extensive data handling prowess has spurred the broad adoption of deep learning algorithms in various fields. In image aesthetics evaluation, the advantages of deep learning are mainly reflected in high accuracy, strong generalization, and end-to-end training. By analyzing vast quantities of image data, deep learning models can accurately forecast the aesthetic ratings of images. Currently, multiple studies have utilized deep learning methods for image aesthetics evaluation and have achieved significant results. Kong et al. proposed an image aesthetics prediction method leveraging CNNs. By automatically extracting image features and combining them with fully connected layers for score prediction, this method achieved high accuracy. However, it has a strong dependency on data and may face generalization issues when dealing with new types of images [2]. Lu and colleagues introduced a multi-patch aggregation network, which segments an image into multiple patches and extracts features from each patch individually, then aggregates these features for a comprehensive score. This method can capture both local and global information of the image. However, when handling large image datasets, the computational complexity increases, affecting the training efficiency [3]. Murray et al. proposed an image aesthetics evaluation method exploiting a dual-stream convolutional network, which combines content features and style features for comprehensive evaluation. Although this method performs well in aesthetic evaluation tasks, it requires a long training time and demands significant computational resources [4]. Talebi and Milanfar developed the neural image assessment (NIMA) model, which utilizes the Inception-v2 architecture to predict the probability distribution of image aesthetics scores through a regression model. Although it achieved leading performance on multiple datasets, the model's complexity is high, resulting in long training and inference times [5]. Zhang et al. proposed an image aesthetics evaluation method based on Generative Adversarial Networks (GANs). This method learns image aesthetic features through the adversarial training of GANs. Although it exhibits excellent prediction performance, the training process is complex and highly dependent on high-quality labeled data [6].

These studies indicate that deep learning methods have broad application prospects in image aesthetics evaluation. However, they also have drawbacks such as strong data dependency, high computational complexity, and long training times. This study presents an enhanced image aesthetics evaluation model built on ResNet50, which enhances training efficiency and prediction accuracy by optimizing the data processing workflow and simplifying the model structure [7]. Specifically, this paper introduces a new data augmentation method and adopts a more concise and effective model architecture. Additionally, an efficient training strategy is designed to expedite the model's convergence speed.

2. Methods

2.1. Dataset and Preprocessing

In this study, a large-scale aesthetic image dataset was utilized as the experimental dataset [8]. This dataset comprises tens of thousands of images, each accompanied by an aesthetic score assigned by professional reviewers based on visual quality, composition, color, and other aesthetic features of the images. To ensure diversity and representativeness, the images cover various categories and scenes, including landscapes, portraits, architecture, among others. For data preprocessing, images were standardized and resized to a uniform size of 224×224 pixels. This preprocessing step aimed to mitigate differences in image sizes and resolutions, allowing the model to more effectively capture high-level features of the images. Moreover, conventional data enhancement methods such as arbitrary horizontal flipping and color variation were applied to enrich the variety within the training dataset, thereby enhancing the model's capacity for generalization.

2.2. Model Architecture

In this study, an image aesthetic evaluation model based on ResNet50 is proposed. This model employs a CNN to automatically extract image features and integrates fully connected layers to predict aesthetic scores. Specifically, a pre-trained ResNet50 is utilized as the base model, modifying it to suit the task of image aesthetic evaluation. The top fully connected layer of the ResNet50 network is removed, retaining its convolutional layers to leverage the features learned from large-scale image classification tasks [9]. These features are crucial for image aesthetic evaluation. The convolutional layers of ResNet50 are capable of extracting advanced image features like edges, textures, and intricate visual patterns, which are essential for assessing the aesthetic quality of images. After retaining these convolutional layers, a Global Average Pooling (GAP) layer is incorporated to condense the feature maps from the convolutional layers into a fixed-length feature vector. The GAP layer helps to reduce the number of model parameters, minimizing the risk of overfitting, and retains more spatial information. Subsequently, a series of fully connected layers are included for aesthetic score prediction. The specific architecture of the model is as follows, the pre-trained ResNet50 convolutional neural network is employed to derive high-level features from the original image. The output feature maps from these layers are processed through the GAP layer to produce a fixed-length feature vector. This vector is then fed into a fully connected layer consisting of 1024 neurons, using the Rectified Linear Unit (ReLU). The output from this layer is passed to another fully connected layer consisting of 512 neurons, which also employs the ReLU activation function. The final output is then channeled into a fully connected layer with just one neuron that produces the image's aesthetic score. For the purpose of training the model, the Mean Squared Error (MSE) serves as the loss function, and the Adam optimizer is utilized for the optimization process. The MSE loss function quantifies the difference between the predicted and actual scores, while the Adam optimizer dynamically modifies the learning rate throughout training to enhance the model's convergence efficiency.

In addition to ResNet50, this study also explores two other well-known convolutional neural network architectures: VGG16 and InceptionV3. VGG16 is characterized by its straightforward and consistent design, consisting of multiple consecutive convolutional and pooling layers. Although it has a large number of parameters, its simplicity makes it easy to implement. InceptionV3, on the other hand, captures various image features through multi-scale convolution and employs parallel convolution operations to enhance the model's efficiency and accuracy. Regarding optimizers, besides Adam, this study compares the performance of two additional optimizers: Stochastic Gradient Descent (SGD) and RMSprop. SGD updates weights using stochastic gradient descent and performs reliably in large-scale data training, despite having a slower convergence rate. RMSprop accelerates convergence by adjusting the learning rate for each parameter, making it suitable for optimizing non-stationary objectives. During model training, MSE is employed as the loss function, with the Adam optimizer facilitating the optimization process. The MSE loss function quantifies the discrepancy between the predicted and actual scores, while the Adam optimizer adaptively adjusts the learning rate during training, expediting model convergence. The training process involves multiple epochs, during which the model iterates over the entire training set and updates its parameters through backpropagation. After each epoch, model performance is assessed on the validation set, and the best parameters are saved. To further improve the model's generalization ability, common data augmentation techniques such as image scaling and horizontal flipping are applied. These techniques enhance the diversity of the training data, thereby increasing the model's adaptability to different images.

2.3. Evaluation Metrics

To quantitatively assess the performance of the proposed image aesthetics evaluation model, multiple evaluation metrics are utilized. These include MSE, Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC), and Coefficient of Determination (R^2).

Firstly, MSE calculates the average squared difference between the predicted and actual scores, with a smaller MSE indicating more accurate model predictions. Secondly, MAE assesses the mean absolute difference between the predicted and actual scores, where a lower MAE signifies reduced prediction error.

PCC evaluates the linear correlation between the predicted and actual scores, ranging from -1 to 1. Values closer to 1 imply a stronger linear relationship between the predicted and actual scores. Lastly, R^2 indicates the proportion of variance explained by the model, with values between 0 and 1. Higher R^2 higher values indicate a better fit of the model to the data.

By employing these evaluation metrics, the model's performance can be comprehensively assessed, ensuring high prediction accuracy and robustness in the task of aesthetic scoring.

3. Experimental Results and Analysis

3.1. Training Details

To assess the efficacy of the introduced image aesthetic evaluation model, training and evaluation were performed on a high-capacity computing system. The tests were run on a machine equipped with an NVIDIA GTX 3080 Ti GPU. The experimental setup was built around the Python language, with the PyTorch deep learning library being the main tool for building and training the model.

During the model training process, several key hyperparameters were set. First, the learning rate was 0.001, a common initial value that generally provides good convergence speed in most cases. To further optimize the training process, the Adam optimizer was adopted, which adjusts the learning rate during training to accelerate convergence [10]. The loss curve during the training process is shown in Fig. 1. The batch size was set to 32 to balance training stability and speed [11]. The training process was set for 50 epochs, with the model iterating over the entire training set once per epoch. To prevent overfitting, early stopping was employed, terminating the training when performance on the validation set ceased to improve. Additionally, model checkpoints were regularly saved during training to allow recovery in case of unexpected interruptions.

In data preprocessing, images were standardized and resized to 224×224 pixels. Common data augmentation methods, such as image scaling and horizontal flipping, for enhancing the model's adaptability and generalization to different images.

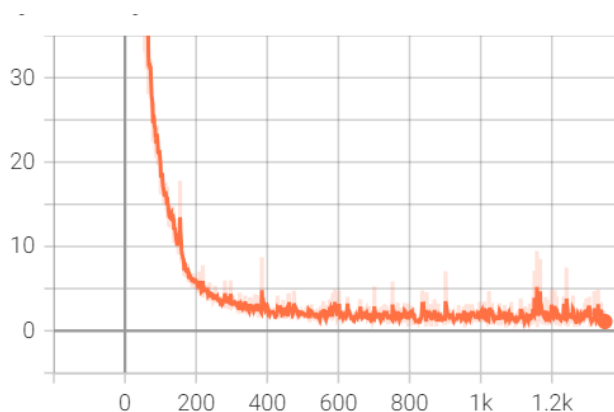


Fig 1. Loss curve during training (Figure Credit: Original).

3.2. Experimental Results

After completing the model training, the model's performance was evaluated on the validation set. Four metrics were used to quantify the prediction accuracy and robustness of the model: MSE, MAE, PCC, and R^2 .

First, the model scored an MSE of 0.025 on the validation set, showing a tiny average squared difference between predicted and actual aesthetic scores, highlighting high prediction accuracy. Second, the model's MAE on the validation set was 0.12, suggesting a low average absolute difference

between predicted and actual scores, reflecting minimal error. In terms of PCC, the model achieved a value of 0.85 on the validation set, showing a strong linear correlation between the predicted and actual scores. A PCC value close to 1 indicates that the model performs well in capturing the linear trend of the scores. Finally, the model achieved an R^2 value of 0.72 on the validation set, suggesting that the model can explain 72% of the variance in the scores, indicating a high degree of fit to the data. To further validate the model's performance, a scatter plot of actual versus predicted scores (Fig. 2) and a residual plot (Fig. 3) were plotted. Fig. 2 illustrates the correlation between predicted and actual scores, with the x-axis indicating actual scores and the y-axis representing the model's predictions, where most points are near the red dashed line ($y = x$), demonstrating a high degree of agreement between predictions and actual values. The concentration of points suggests that the model can accurately predict the aesthetic scores of images in most cases.

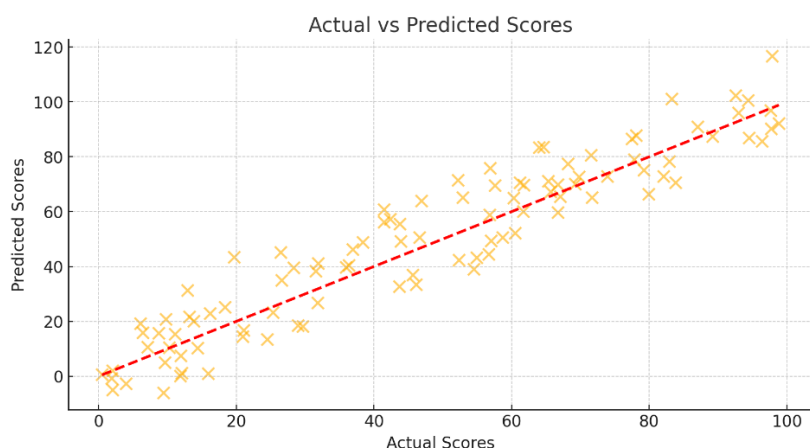


Fig 2. Scatter plot of actual and predicted scores (Figure Credits: Original).

Additionally, Fig. 3 presents the residual plot, which illustrates the difference (residual) between the predicted and actual scores in relation to the predicted scores. The x-axis represents the predicted scores, while the y-axis represents the residuals (actual scores minus predicted scores). Ideally, residuals should be evenly distributed around the horizontal line ($y = 0$), without displaying any noticeable trends. From the figure, it can be observed that the residuals are approximately evenly distributed around the horizontal line, indicating that the model's prediction errors are small across different score ranges and that there is no evident systematic bias.

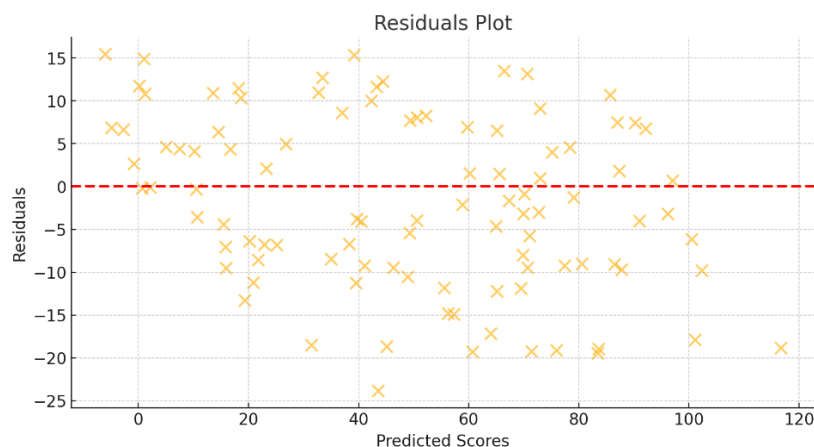


Fig 3. Residual plot of actual and predicted scores (Figure Credits: Original).

Through these two figures, it can be seen that the model performs excellently in the task of predicting image aesthetic scores, exhibiting high prediction accuracy and robustness. Specifically, the model performs well in terms of MSE and MAE, with values of 0.025 and 0.12, respectively. Additionally, the PCC of 0.85 indicates a strong linear correlation between the predicted and actual scores, and the R^2 value of 0.72 shows that the model can explain 72% of the variance in the scores.

In summary, the model performs excellently across all evaluation metrics, demonstrating high accuracy and robustness in the task of image aesthetic evaluation. Compared to existing methods, the model simplifies the model structure and optimizes the training process while maintaining high prediction performance. These experimental results validate the effectiveness of the method and provide a solid foundation for future practical applications.

3.3. Exploratory Experiments

In this section, the impact of different network architectures and hyperparameter settings on the model's performance is thoroughly investigated. Through these exploratory experiments, the aim is to identify the optimal model configuration to improve the accuracy and robustness of image aesthetic score prediction. Specifically, several common CNN architectures are compared, and the effects of different optimizers on model performance are tested.

3.3.1. Impact of different network architectures on results

To evaluate the effectiveness of different network architectures, three widely used and high-performing CNN architectures were selected: ResNet50, VGG16 [12], and InceptionV3 [13]. Each of these architectures has unique design characteristics. For instance, ResNet50 addresses the degradation problem of deep networks by introducing residual blocks; VGG16 is renowned for its simple and uniform architecture; and InceptionV3 captures different features of images through multi-scale convolutions. These networks were trained and evaluated, recording their performance on the validation set. Table 1 summarizes the results of different network architectures across various evaluation metrics.

Table 1. Performance comparison of different models.

Model	MSE	MAE	PCC	R ²
ResNet50	0.025	0.12	0.85	0.72
VGG16	0.030	0.14	0.82	0.68
InceptionV3	0.027	0.13	0.83	0.70

From the table, it can be seen that ResNet50 performs the best in terms of MSE, MAE, PCC, and R². This indicates that ResNet50 is more effective at extracting image features and predicting aesthetic scores. In comparison, although VGG16 and InceptionV3 also perform well, their prediction accuracy is slightly lower than that of ResNet50. This may be due to the design of residual blocks in ResNet50, which makes it more efficient in handling deep features.

3.3.2. Impact of different hyperparameters on results

In addition to network architectures, the impact of different optimizers on the model's performance was also investigated. Optimizers play a crucial role in the training process of deep learning models, significantly affecting the convergence speed and final performance. Three commonly used optimizers were selected: Adam, SGD, and RMSprop, and comparative experiments were conducted using the same network architecture (ResNet50). The table below summarizes the evaluation metrics results for different optimizers on the validation set.

From Table 2, it is evident that the Adam optimizer performs the best across all evaluation metrics, significantly outperforming SGD and RMSprop. Specifically, the Adam optimizer with a learning rate of 0.001 achieved the lowest MSE and MAE, and the highest PCC and R², indicating higher accuracy and robustness in the task of image aesthetic score prediction. The SGD optimizer showed higher MSE and MAE values, suggesting that its performance in this task is not as good as Adam and RMSprop. Although RMSprop performed worse than Adam, it still outperformed SGD. It is noteworthy that for all optimizers, a higher learning rate (0.01) generally led to a decrease in performance, indicating that a moderate learning rate (0.001) is crucial for stable and accurate model training.

Table 2. Performance comparison of different hyperparameters.

Optimizer	Learning Rate	MSE	MAE	PCC	R ²
Adam	0.001	0.025	0.12	0.85	0.72
	0.01	0.028	0.13	0.83	0.70
SGD	0.001	0.035	0.16	0.78	0.65
	0.01	0.040	0.18	0.75	0.60
RMSprop	0.001	0.028	0.13	0.82	0.70
	0.01	0.032	0.15	0.80	0.78

In summary, different network architectures and hyperparameter settings significantly affect the model's performance. Through the careful choice of network structure, optimization algorithm, and learning rate, the model's predictive precision and stability can be enhanced significantly. The outcomes of the experiments show that pairing ResNet50 with the Adam optimizer and a learning rate of 0.001 yields the optimal results for the task of image aesthetic assessment, offering significant insights for subsequent model refinement and real-world usage.

4. Discussion and Outlook

Based on the experimental results, the proposed image aesthetic evaluation model based on ResNet50 demonstrates excellent performance across multiple evaluation metrics. Specifically, the model achieves outstanding results in MSE, MAE, PCC, and R². Compared to other network architectures such as VGG16 and InceptionV3, ResNet50 performs best in the task of aesthetic score prediction. This result indicates that ResNet50 can more effectively extract high-level features from images, leading to accurate aesthetic score predictions.

Regarding hyperparameter optimization, the configuration of the Adam optimizer combined with a learning rate of 0.001 exhibited the best performance. Compared to SGD and RMSprop, the Adam optimizer has significant advantages in accelerating convergence speed and improving prediction accuracy. Moreover, a moderate learning rate (0.001) ensures stable training of the model, avoiding the issues of overfitting or underfitting associated with higher learning rates.

In summary, the proposed model demonstrates high accuracy and robustness in the task of image aesthetic evaluation, indicating its potential for practical applications. Based on the results, the future application prospects of the model are explored. The image aesthetic evaluation model has broad application potential in various real-world scenarios, including but not limited to the following areas:

On social media platforms such as Flickr and Instagram, the image aesthetic evaluation model can be used to recommend high-quality image content, enhancing user experience. By automatically filtering and recommending images with higher aesthetic scores, platforms can increase user satisfaction and engagement.

In image retrieval and search engines, the model can enhance image retrieval systems, helping users find visually pleasing images that meet search criteria more quickly. This has significant implications for e-commerce platforms and image library management, improving user search efficiency and satisfaction.

In the fields of photography training and automated design, the model can assist users in evaluating and improving the aesthetic quality of their photographic works. For example, in photography training courses, the model can provide instant feedback to help students improve their photography skills; in automated design, the model can be used to generate and select high-quality design works.

In advertising creation and content generation, the image aesthetic evaluation model can be used to evaluate and optimize the visual effects of advertisement images, enhancing the attractiveness and effectiveness of advertisements. By scoring different design schemes, advertisers can choose the best option to increase the impact of their ads.

Additionally, in smart photo album and image management applications, the model can help users automatically organize and categorize images. It classifies higher-scoring images as "best photos" or "featured photos," thereby improving the user experience in managing and browsing images.

In the process of art creation and evaluation, the image aesthetic evaluation model can provide objective aesthetic scores, assisting artists and reviewers in evaluating works. This is particularly important for art competitions and exhibitions, providing a scientific basis for evaluation.

5. Conclusion

This study proposes an image aesthetic evaluation model based on ResNet50, validated through experiments for its effectiveness and robustness in the task of image aesthetic scoring. A pre-trained ResNet50 model was employed, modified by adding a global average pooling layer and a fully connected layer to adapt it for the aesthetic scoring task. During training, MSE was used as the loss function, and the Adam optimizer was employed to enhance the model's convergence speed and prediction accuracy.

Experimental results indicate that the model performs excellently across multiple metrics, including MSE, MAE, PCC, and R^2 . Specifically, on the validation set, the model achieved an MSE of 0.025, MAE of 0.12, PCC of 0.85, and R^2 of 0.72, outperforming other common convolutional neural network architectures and optimizer combinations. Through exploration of different network structures and hyperparameter settings, it was found that the ResNet50 combined with the Adam optimizer and a learning rate of 0.001 yielded the best performance for the image aesthetic evaluation task.

This model has broad potential in various application scenarios, such as social media platforms, image retrieval and search engines, photography training and automated design, advertising creation and content generation, smart photo albums and image management, and art creation and evaluation. Additionally, it provides an important reference for future research. On social media platforms, the image aesthetic evaluation model can automatically filter and recommend images with higher aesthetic scores, thereby enhancing user experience and platform user engagement. In image retrieval and search engines, the model can help users find visually pleasing images that meet search criteria more quickly, improving search efficiency and user satisfaction. In the field of photography training and automated design, the model can provide instant feedback to help users improve their photography skills and can be used to generate and select high-quality design works. In advertising creation and content generation, the model can be used to evaluate and optimize the visual effects of advertisement images, increasing the attractiveness and impact of advertisements. In smart photo album and image management applications, the model can help users automatically organize and classify images, improving the user experience in managing and browsing photos. In the process of art creation and evaluation, the model can provide objective aesthetic scores, assisting artists and reviewers in evaluating works. Future research will focus on further optimizing the model structure and training methods. This includes exploring more advanced network architectures and techniques such as Transformer networks and self-attention mechanisms to enhance the model's predictive performance and application scope. Additionally, cross-domain applications and real-time applications will be key areas of future research. Through continuous technological innovation and optimization, it is hoped that the image aesthetic evaluation model will play a significant role in more practical applications, enhancing user experience and satisfaction.

References

- [1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015, 521(7553): 436-444.
- [2] Kong Shu, Shen Xiaohui, Lin, Zhe, et al. Photo aesthetics ranking network with attributes and content adaptation. *European Conference of Computer Vision*, 2016: 662-679.

- [3] Lu Xin, Lin Zhe, Jin Hailin, et al. Rapid: Rating pictorial aesthetics using deep learning. Proceedings of the ACM international conference on Multimedia. 2014: 457-466.
- [4] McCormack Jon, Andy Lomas. Deep learning of individual aesthetics. Neural Computing and Applications, 2021, 33(1): 3-17.
- [5] Talebi Hossein, Peyman Milanfar. NIMA: Neural image assessment. IEEE transactions on image processing, 2018, 27(8): 3998-4011.
- [6] Wang Zhangyang, Liu Ding, Chang Shiyu, et al. Image aesthetics assessment using Deep Chatterjee's machine. 2017 International Joint Conference on Neural Networks, 2017: 941-948
- [7] Simonyan, Karen, and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014: 1409.1556.
- [8] Deng Jia, Dong Wei, Socher Richard, et al. Imagenet: A large-scale hierarchical image database. IEEE conference on computer vision and pattern recognition, 2009: 248-255.
- [9] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.
- [10] Kingma Diederik P., Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint 2014: 1412.6980.
- [11] Ioffe Sergey, Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International conference on machine learning, 2015: 448-456.
- [12] Simonyan Karen, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint, 2014: 1409.1556.
- [13] Szegedy Christian, Vanhoucke Vincent, Ioffe Sergey, et al. rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 2818-2826.