

A Hybrid Model for Image Aesthetic Assessment Integrating Theme Prediction and ViT-based Composition Analysis

Biao Xiang *

Carey Business School, Johns Hopkins University, Baltimore, United State

* Corresponding Author Email: bxiang2@jh.edu

Abstract. This paper proposes a novel hybrid model for Image Aesthetic Assessment (IAA), which integrates theme prediction and Vision Transformer (ViT)-based composition analysis. Unlike traditional deep learning methods that rely solely on aesthetic scores for training, this approach combines semantic and photographic themes with the spatial relationships of visual elements. By leveraging theme labels and compositional relationships, the model captures aesthetic qualities from a more human-centric perspective. The model comprises three modules: a pre-trained Theme Understanding Network (TUN) to extract semantic and photographic theme features, a ViT module to analyze the spatial dependencies of visual elements, and a scoring module to concatenate features from the previous modules and then generate the prediction of images' aesthetic score. This integration enhances interpretability by providing attention heat maps and detailed theme and semantic labels, facilitating a deeper exploration of human aesthetic intelligence. The Theme and Composition-aware IAA (TC-IAA) model achieves a Mean Squared Error (MSE) of 0.47 on the Aesthetic Visual Analysis (AVA) dataset, demonstrating its effectiveness. Its practical implications extend to images aesthetic recommendation, photo-editing, and targeted image advertising, highlighting its potential impact on real-world applications.

Keywords: Image Aesthetic Assessment; Theme Prediction; Vision Transformer (ViT); Interpretability.

1. Introduction

Image Aesthetic Assessment (IAA) refers to the evaluation of the aesthetic appeal of images, mapping visual features to classifications (high, medium, low) or aesthetic scores [1]. Automatic image assessment has significant influence on a various application such as image editing, advertising recommendations, and user-generated content filtering, making it essential in social media, tourism, e-commerce, and other domains.

Historically, IAA research has focused on using manually designed visual features and conventional machine learning techniques such as Bayesian approaches and support vector machines [2]. The emergence of deep learning has led to substantial advancement in IAA, with Convolutional Neural Networks (CNN) markedly enhancing performance by training models from pixel-level, continuously producing state-of-the-art performance in the IAA field [3, 4]. However, previous studies directly predicted aesthetic scores from images annotated with human ratings. This pure data-driven manner tends to ignore the mechanism of human's aesthetic perception and the distribution of aesthetic preferences [5]. The determination of aesthetic value involves intricate reasoning processes, research has shown that image aesthetics are influenced by various factors such as theme, lighting, color, spatial relationships, and style [6]. Among these, theme is recognized as a crucial factor in both photography and aesthetics [7]. With different themes, people generally have distinct aesthetic standards for images. Photographers first pay attention to the theme of the image and then apply different evaluation criteria [5]. Incorporating the concept of theme into IAA, rather than solely generating a single aesthetic score mapping, can enhance the interpretability of the model and improve its application in real-world scenarios.

Despite the great success of CNN-based models, the reliance on convolution operations presents certain limitations. One of the major limitations is the restricted receptive field of convolutions. Although this can be expanded by stacking convolutional layers such as Visual Geometry Group Network (VGGNet), or by employing techniques like pyramid pooling layers to capture more global

features, convolutions still struggle to capture the spatial relationships and positional structures of visual elements in an image [8]. This is essential for IAA, as visual elements in images are often mutually dependent on each other and collectively contribute the aesthetic value of whole image [9]. The advent of Vision Transformers (ViT) push forwarded transformer into visual domain, marked a significant breakthrough relevant study [10]. By utilizing the attention mechanism, ViT effectively captures global features, compensating for the shortcomings of convolution. ViT divides images into small size patches (typically 16x16), after positional embedding, transformer encoder gets the attention weight of each patch to construct an attention heat map, analyzing the dependencies among different regions of the image to understand the relationships between visual elements. Many transformer-based models have been proposed in the IAA field, achieving remarkable performance [11].

Motivated by the need for more accurate and interpretable IAA, this study proposes a hybrid model that integrates theme prediction and ViT-based composition analysis in a parallel framework. Specifically, the approach begins with a pre-trained Theme Understanding Network (TUN) to extract semantic and photographic theme features from the image. Concurrently, a Vision Transformer (ViT) module is employed to analyze the compositional relationships of visual elements within the image, utilizing the attention mechanism to generate comprehensive global features. The integration of these parallel modules represents the key innovation of this study. By combining the outputs of both the theme prediction and composition prediction modules, the model can effectively capture both thematic and spatial elements of the image. These comprehensive features are subsequently input into the aesthetic scoring module to predict the final aesthetic score. This approach addresses the limitations of CNN in capturing spatial relationships, enhances the interpretability of aesthetic assessment models, and broadens the potential applications of IAA across various domains. The result of experiment on the Aesthetic Visual Analysis (AVA) dataset shows that the proposed Theme and Composition-aware IAA (TC-IAA) model achieves a Mean Squared Error (MSE) of 0.47 in aesthetic score prediction. Additionally, the model provides an attention heat map that elucidates the foundation of the aesthetic score by highlighting the key regions of the image that significantly influence the score.

2. Methodology

2.1. Dataset Description and Preprocessing

The AVA dataset stands as a significant resource for evaluating image aesthetics on a large scale [12]. This dataset comprises 255,508 images collected from an online community for amateur photographers. Each image is evaluated by an average of 200 users, with ratings ranging from 1 to 10, and is tagged with 66 semantic labels and 1,397 photographic theme labels. In this study, the dataset was partitioned into 230,000 images for training, 20,000 for testing, and the remaining 5,508 for validation. Each image includes two semantic labels, one photographic theme labels, and an average aesthetic score as ground truth. The AVA dataset provides 2 semantic labels from each image, in this study, prediction on semantic labels is considered correct if it matches either one of the labels.

2.2. Proposed Approach

This study aims to predict an aesthetic score along with semantic and theme labels for a given input image I . The model combines TUN and ViT in a parallel framework, extracting both thematic and compositional features to enhance the aesthetic assessment.

The aesthetic scoring function f_{aes} parametrized with θ_{aes} can be formulated as:

$$I \xrightarrow{f_{aes}(\theta_{aes})} (s, y_t, y_s) \quad (1)$$

Where s the prediction of images' aesthetic score is, y_s represents the semantic labels, and y_t denotes the theme labels.

The model consists of three key components: TUN, ViT, and the scoring module. TUN is based on a pre-trained ResNet-18, extracts high-level semantic and photographic theme features from the input image. ViT processes the image to capture global and compositional features. It also provides an attention heat map of the image based on the average attention weight. The scoring module combines the features from TUN and ViT to predict the final aesthetic score. As depicted in Fig. 1, the workflow of the proposed model is begun with the input image being concurrently processed by both TUN and ViT. TUN extracts semantic and theme labels, while ViT generates a compositional feature representation. These features are then concatenated and processed through the scoring module to predict the aesthetic score.

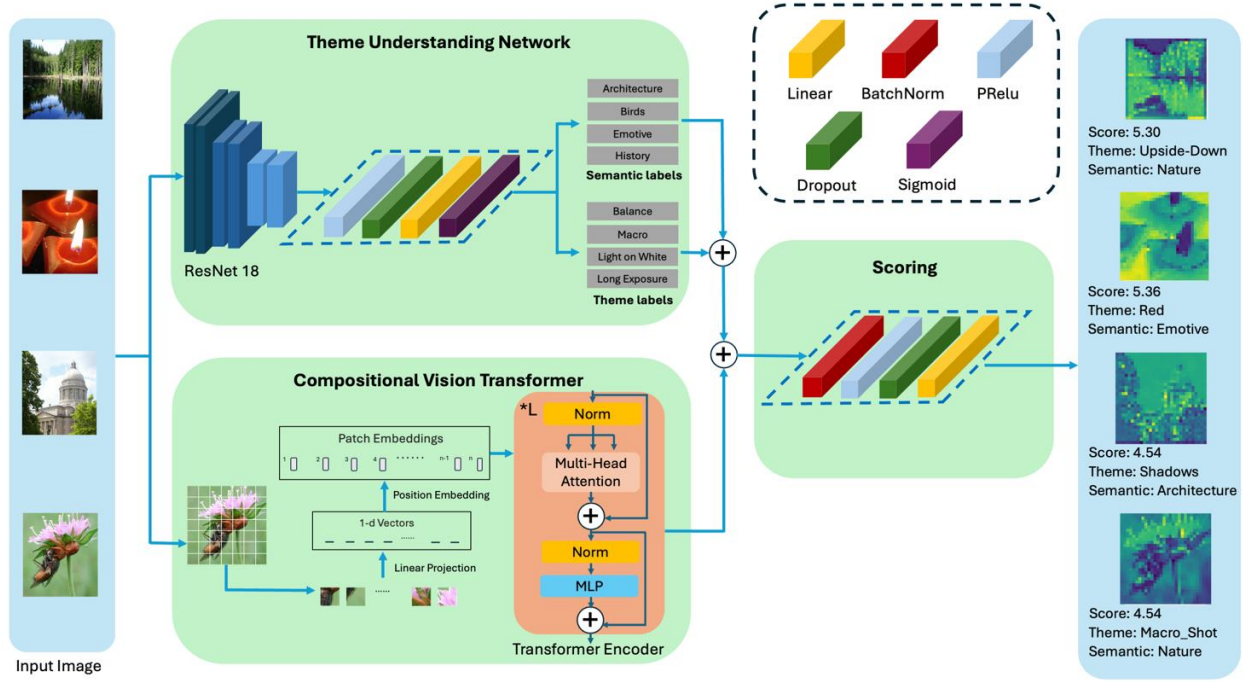


Fig 1. Model Pipeline.

2.2.1. Theme understanding network

TUN is designed to extract both semantic and photographic theme features of an image. It uses a ResNet18 backbone, pre-trained on ImageNet, to extract deep features from input images [3]. The network is trained separately for theme and semantic labels to ensure specialized learning for each category.

TUN is trained using dataset $D_t = \{(x_i, t_i, s_i)\}_{i=1}^N$, where x_i represents the input image, t_i and s_i represents corresponding theme and semantic labels. The features extracted by ResNet18 are passed through a PReLU activation and multi-layer perceptron (MLP) to produce the final theme label probabilities. The prediction process can be formulated as:

$$\hat{t}_i = MLP\theta_t(F_{\theta_{Res}}(x_i)), \quad \hat{s}_i = MLP\theta_s(F_{\theta_{Res}}(x_i)) \quad (2)$$

Where θ_t and θ_s denotes parameters of theme and semantic labels, $F_{\theta_{Res}}$ denotes the feature extracted by ResNet18, and $MLP\theta$ represent theme prediction MLP.

Both models are trained using the cross-entropy loss function:

$$L_{theme} = -\sum_{i=1}^{N_t} t_i \log(\hat{t}_i), \quad L_{semantic} = -\sum_{i=1}^{N_s} s_i \log(\hat{s}_i) \quad (3)$$

The parameters θ_t and θ_s are frozen during the final prediction to avoid ensure integrity of pre-trained features. During the prediction, the input image x first pass through TUN with parameter θ_t to obtain theme label. Then, the same input image is passed though TUN again with parameter θ_s to obtain semantic label.

2.2.2. Compositional vision transformer

The ViT module aims to extract global and commotional feature from input image, which is important for understanding the spatial relationship of image's visual elements. The ViT processes an image by first dividing it into non-overlapping patches, each of which is linearly embedded into a fixed-size vector. This sequence of embedded patches is then concatenated with a class token (CLS), a special token that accumulates the global information from the image.

Given an input image X in size of 244×244 , the ViT module first splits X into N parches, each of size 16×16 . The patches are flattened to vectors and then be transformed into patch embeddings E through linearly projection. A classification token E_{CLS} is appended to the sequence of patch embeddings. To preserve positional information, positional embeddings E_{pos} are incorporated, resulting in the final embedding Z_0 . This embedding Z_0 is inputed into the transformer encoder, which comprises L layers featuring multi-head self-attention mechanisms and feed-forward neural networks. Following the traversal through these transformer layers, the output coreesponding to the CLS token $Z_{L,0}$ severed as the final representation of the image. The attention weights A generated during the self-attention mechanism are used to create attention heat maps. For each attention head h , the attention weights A_h are computed as.

$$A_h = \text{soft max}(\frac{Q_h K_h^T}{\sqrt{d_k}}) V_h \quad (4)$$

The query, key, and value matrices for head h are denoted as Q_h , K_h^T , and V_h respectively, with d_k representing the dimensionality of the key vectors. By averaging the attention weights across all heads, attention heat maps are created to emphasize the regions of the image that significantly influence the final representation. The CLS token representation, which serves as the output of the ViT module, is then concatenated with outputs of the TUN module to generate the final prediction of image's aesthetic score. The ViT module is initialized using a pretrained ViT-B_16 model to leverage learned representations, enhancing the feature extraction process.

2.2.3. Scoring

The scoring module integrates outputs from TUN and ViT modules to generate the prediction of aesthetic score. Specifically, the TUN provides the semantic and theme labels y_s and y_t , while the ViT extracts global features CLS token $Z_{L,0}$. Concatenate outputs from these two modules to get the combined feature. The whole process can be described as:

$$s = f_{aes}(y_s \oplus y_t \oplus Z_{L,0}, \theta_{aes}) \quad (5)$$

The scoring module is trained using the Mean Squared Error (MSE) loss function. The MSE quantifies the discrepancy between the predicted score s and the actual score $\hat{s}$. The loss function is defined as follows:

$$\mathcal{LMSE} = \frac{1}{N} \sum_i (s_i - \hat{s}_i)^2 \quad (6)$$

2.3. Implementation Details

The implementation of proposed method is under the environment of Python 3.10 and leverages the PyTorch framework. The model training is conducted on a macOS device with an Apple M3 graphics processing unit (GPU). The input images are resized to 244x244 pixels. To avoid over-fitting and enhance the robustness, data augmentation techniques such as random rotation, horizontal flipping, and random scaling of the input images with probability of 0.5 are used during training. The Adam optimizer is used for training the model for 30 epochs in total, starting with an initial learning rate of 1×10^{-4} and a batch size of 64 images. A weight decay of 1e-5 is applied, and batch normalization is performed before each weight layer.

3. Result and Discussion

This chapter will provide an analysis and discussion of the training result of the TC-IAA model. The outputs of the model include the photographic theme label, the semantic label, an attention heat map, and predicted aesthetic score. This chapter will first analysis the accuracy and distribution of predicted aesthetic score, then will analysis the impact of photographic theme label and semantic label on the score prediction. Fig. 2 shows some sample outputs.

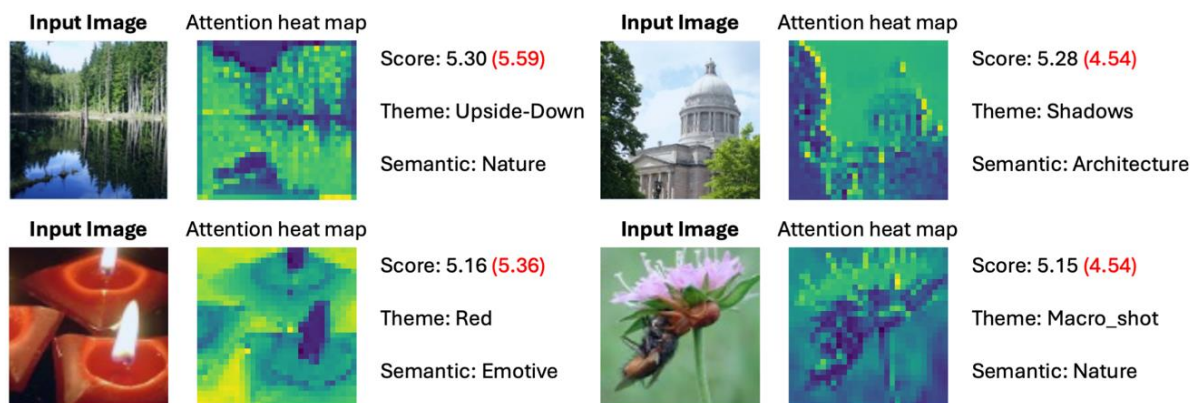


Fig 2. Sample outputs.

3.1. Predicted Aesthetic Score

The pretrained TUN achieved an accuracy of 52.43% in predicting semantic label and photographic theme label. After training the TC-IAA for 30 epochs, the result for predicting aesthetic score were: MSE = 0.3783, MAE = 0.5256, MAPE = 10.81%. As shown in Fig. 2, TC-IAA not only predict the aesthetic score of an image, but also provide attention heat map through ViT module. The heat map can identify the key aesthetic region of an image, which is helpful for analyzing why the image receives a certain score. The labels and the attention heat map provide more interpretability to the model's prediction, which broadens its applicability in various aspects, such as providing aesthetic recommendation and assisting in image modification.

Fig. 3 illustrates the distribution of aesthetic scores in the test set alongside the predicted scores generated by the TIC-IAA model. For the ground truth of test set, the sores are concentrated between 4 to 7 with an average of 5.39. Compared to the score distribution of test set, the predicted scores are more closely centered around 5. This indicates that the model performs poorly in predicting images with relatively high or low aesthetic score, its generalization still needs to be improved.

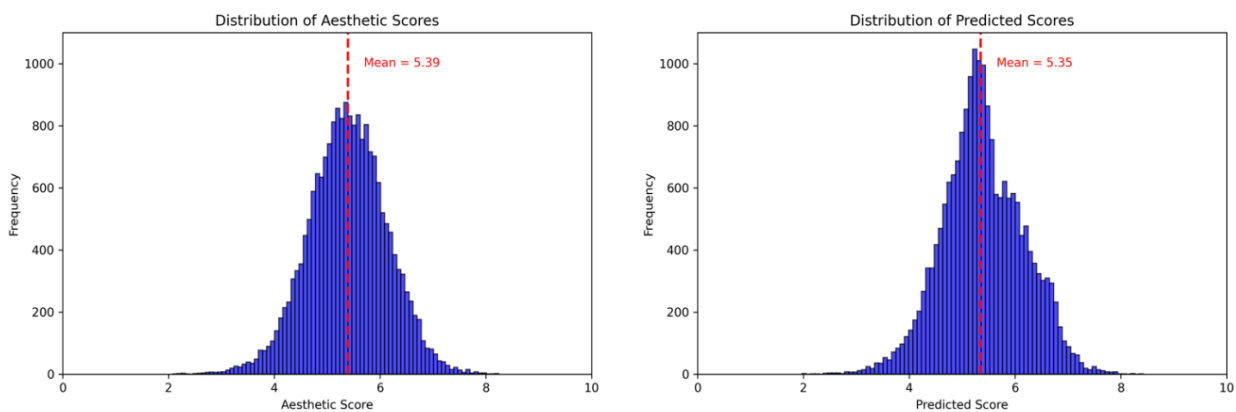


Fig 3. Score distribution.

3.2. Theme and Semantic Labels

Fig. 4 shows the MSE heatmap of the aesthetic score prediction for images belongs to top 20 different semantic labels and photographic theme labels. The heatmap shows that theme labels have

a significant effect on model predictions. For example, prediction on images with the theme label ‘Selective_Desaturation_II’ have an MSE of 0.28, whereas those with the theme ‘Free_Study_X’ have an MSE of 0.69. However, semantic labels show a minimal effect on score predictions, MSE for images with different semantic labels hovering around 0.5.

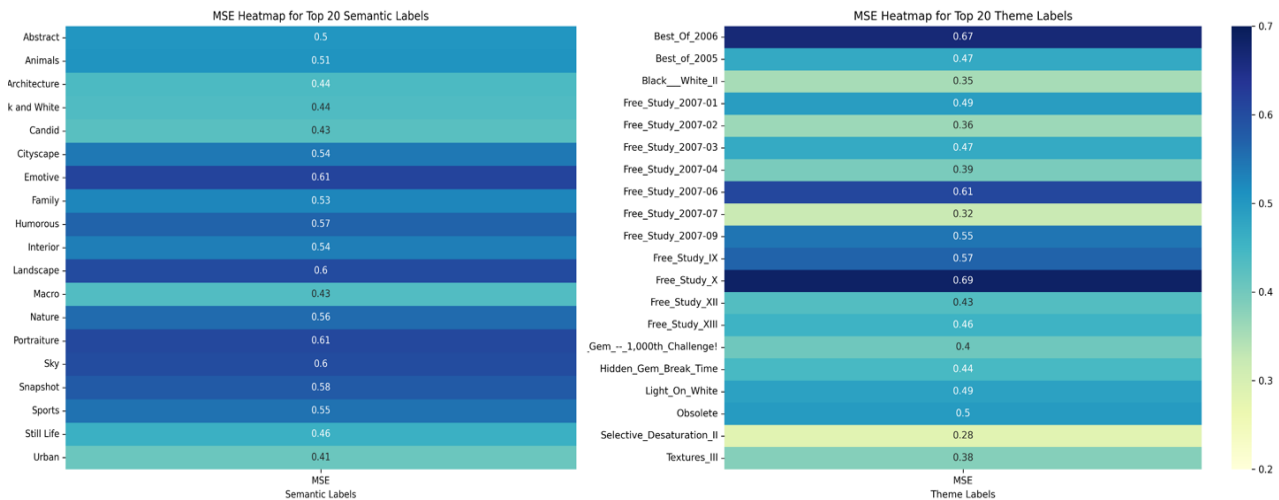


Fig 4. MSE heat map for labels.

This discrepancy can be attributed to the annotation of semantic labels in AVA dataset. AVA dataset has a consistent manner for annotate photographic theme labels but lacks clear definitions for semantic labels [12]. Fig. 5 shows the distributions of the two types of labels. As the graph shown, there's a highly uneven distribution of semantic labels, with most images lacking a semantic label (label with 0). For TUN module to have a better performance, it may require a dataset with more precise and accurate label annotations.

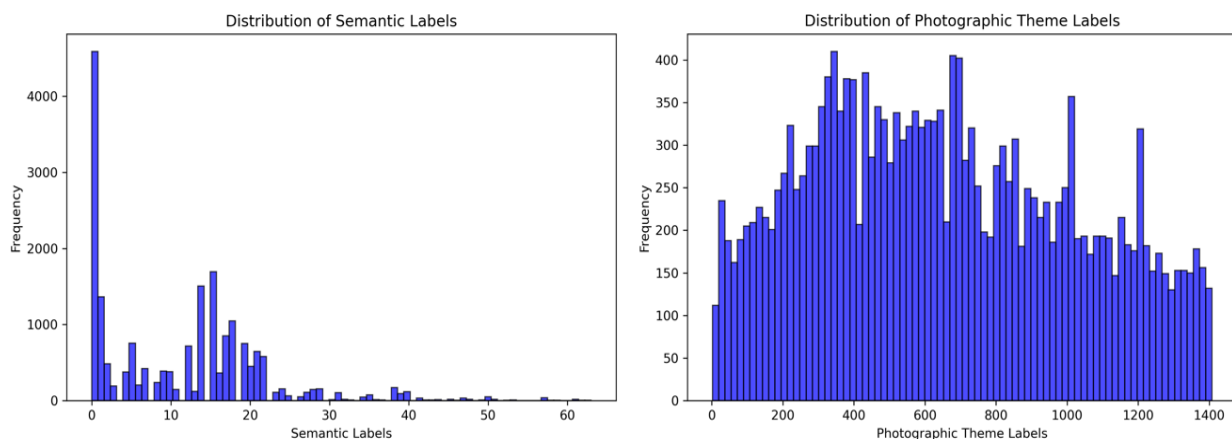


Fig 5. Distribution of labels.

4. Conclusion

This study introduced the TC-IAA model, a novel network architecture for IAA. The TC-IAA model comprises a pre-trained TUN, a Compositional ViT, and a Scoring module. Unlike traditional approaches that simply learn a mapping from image to score, TC-IAA analyzes and understands images from a human perspective by integrating semantic and photographic theme labels from TUN with the spatial relationships of visual elements from the ViT module. This integration makes the model can capture both global and local features effectively. The result of experiment on the AVA dataset shows that TUN achieves an accuracy of 52.43% in predicting semantic and photographic labels. The attention heat maps generated by the ViT module accurately identify the salient regions of images, providing greater interpretability of the results. The final scoring module predicts aesthetic

scores with a MSE of 0.47. Future research will focus on incorporating additional art-related features, such as lighting and color relationships, to augment the model's capacity to interpret images from an aesthetic viewpoint. This effort is directed at enhancing the model's performance and broadening its applicability across diverse artistic and photographic domains.

References

- [1] Deng Y. Loy C. C. & Tang X. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 2017, 34(4): 80-106.
- [2] Datta R. Joshi D. Li J. & Wang J.Z. Studying Aesthetics in Photographic Images Using a Computational Approach. In *European Conference on Computer Vision*. 2006.
- [3] Krizhevsky A. Sutskever I. & Hinton G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 2012, 25.
- [4] Talebi H. & Milanfar P. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 2018, 27(8): 3998-4011.
- [5] He S. Zhang Y. Xie R. Jiang D. & Ming A. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. In *IJCAI*, 2022: 942-948.
- [6] Celona L. Leonardi M., Napoletano P. & Rozza A. Composition and style attributes guided image aesthetic assessment. *IEEE Transactions on Image Processing*, 2022, 31: 5009-5024.
- [7] Li L. Huang Y. Wu J. Yang Y. Li Y. Guo Y. & Shi G. Theme-aware visual attribute reasoning for image aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(9): 4798-4811.
- [8] Liu D. Puri R. Kamath N. & Bhattacharya S. Composition-aware image aesthetics assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020: 3569-3578.
- [9] Mai L. Jin H. & Liu F. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016: 497-506.
- [10] Dosovitskiy A. Beyer L. Kolesnikov A. Weissenborn D. Zhai X. Unterthiner T. & Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021.
- [11] You J. & Korhonen J. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021: 1389-1393.
- [12] Murray N. Marchesotti L. & Perronnin F. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, 2408-2415.