Image Aesthetic Assessment Based on GNN-Guided Deformable Attention for Electronic Photography

Lin Li , Jichun Zhu , Mingxing Jiang * , Jingli Fang

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1595.v1

Keywords: image aesthetics assessment; graph neural network; deformable attention; deep learning; Transformer

Article

Image Aesthetic Assessment Based on GNN-Guided Deformable Attention for Electronic Photography

Lin Li [1], Jichun Zhu [1], Mingxing Jiang [2,*] and Jingli Fang [1]

[1] School of Computer and Information, Hefei University of Technology, Hefei 230602, China

[2] School of Computer Science and Artificial Intelligence, Chaohu University, Hefei 238024, China

* Correspondence: mx0551@163.com

Abstract

With the increasing demand for high-quality imaging in consumer electronics, image aesthetic assessment (IAA) has been widely applied to electronic cameras and display devices. Although the deformable attention mechanism has been introduced into IAA due to its perceptual capabilities, enabling models to refine attention regions by learning interest points and their corresponding offsets, existing methods often lack guidance from aesthetic composition features during the offset generation process, which limits their performance in aesthetic evaluation tasks. To address this issue, we propose a Graph Neural Network (GNN)-guided deformable attention module that incorporates composition information into the generation of interest points by modeling image features as graphs and applying GNN to guide interest point selection. In addition, we design an improved Transformer model that employs neighborhood attention to further enhance IAA performance. We evaluate the proposed model on two aesthetic datasets, AVA and TAD66K, and the experimental results demonstrate its effectiveness in improving overall model performance.

Keywords: image aesthetics assessment; graph neural network; deformable attention; deep learning; Transformer

1. Introduction

With the development of technology, consumer electronic devices have become increasingly prevalent. People have an increasing demand for high-quality imaging, especially in electronic photography. Consequently, it has become particularly important to evaluate these images from the perspective of human aesthetic perception. Image quality assessment (IQA) methods have been applied to consumer electronic devices. For example, some methods focus on display devices in consumer electronics, enhancing image presentation by evaluating color quality or leveraging brightness information to reduce distortion [1,2]. In addition, certain IQA methods are designed for specialized domains, such as vegetable quality detection [3] or image reconstruction in dental equipment [4]. However, these approaches primarily address objective image quality and often overlook subjective factors, such as human aesthetic perception. In general, IQA focuses on technical fidelity and distortion measurement, whereas IAA aims to evaluate subjective aesthetic appeal from a human perspective. Therefore, IQA methods are insufficient for meeting the practical demands of photography-related applications in consumer electronics. Image Aesthetic Assessment (IAA) has attracted growing attention in recent years. The IAA has been integrated into a variety of consumer electronics applications, such as assisting users in capturing aesthetically pleasing photographs and recommending content based on users' aesthetic preferences in software systems [5,6]. When users take photos using consumer electronic devices such as cameras or smartphones, the IAA-based methods can provide attribute scores to assist users, thereby enhancing the quality of their photography. A feasible solution is shown in Figure 1. A schematic diagram of an aesthetic evaluation solution that helps users improve their photography. After taking photos, the IAA tool provides scores from multiple aspects, helping users identify shortcomings and make targeted improvements

 doi:10.20944/preprints202602.1595.v1

to enhance image quality.. This solution significantly enhances the photo-taking experience and is not limited to consumer electronics, it can also be applied to general-purpose devices such as office terminals and human-computer interaction systems.

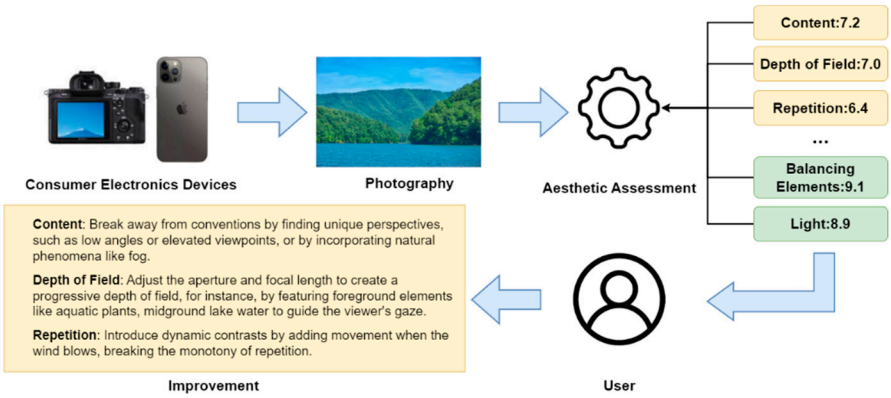


Figure 1. A schematic diagram of an aesthetic evaluation solution that helps users improve their photography. After taking photos, the IAA tool provides scores from multiple aspects, helping users identify shortcomings and make targeted improvements to enhance image quality.

Recently, most IAA methods are deep learning-based methods. Representative methods include those based on Graph Neural Network (GNN) [7–9], large language models (LLM) [10], and Transformer [11–13]. Among these, the Transformer methods [14] stand out for its superior performance in IAA tasks, largely due to its ability to model long-range dependencies and its remarkable scalability. However, current Transformer-based methods still exhibit certain limitations. In particular, they often fail to adequately capture the importance of image composition, which is a critical factor in aesthetic evaluation. For example, in photographic works that follow specific compositional principles [15,16], the model should be able to focus on features guided by these principles. However, existing Transformer-based models lack this compositional awareness.

The deformable attention can be used to solve this problem. This attention originates from deformable convolution, is a method that flexibly focuses on spatial positions based on the input. It has achieved satisfactory results in fields such as object detection [17,18]. This mechanism allows for attention to different locations in an image based on guided interest points. He *et al.* refine features using deformable attention to balance the attention between foreground and background, thereby optimizing aesthetic scoring outcomes [11]. However, the generation of interest point does not fully consider layout factors. As is well known, the GNN is a neural network specifically designed to handle graph-structured data, which typically consists of nodes and edges representing entities and the relationships between them, respectively. Therefore, we employ the GNN into this process to guide interest point selection. Compared to traditional neural networks, the GNN can more effectively capture the relationships between image regions, thereby enhancing the expressive power of image features and improving the performance of IAA tasks.

To fully utilize the layout information of an image for feature extraction, this paper proposes a Transformer model guided by graph convolution and deformable attention. This model consists of four stages. In each stage, we first use an neighborhood attention block [19] to preliminarily extract image features. This block is based on a flexible improved sliding window attention mechanism. Second, we design a deformable attention block to refine the features. Specifically, in each deformable attention block, we first model the input image features as a graph structure and obtain initial interest point coordinates through a graph neural network-based generation module. Subsequently, we use an offset generation module to obtain displacements, allowing the interest points to further shift to reasonable areas. Finally, these features are sampled based on the coordinates of these interest points.

Our contributions are concluded as follows:

 doi:10.20944/preprints202602.1595.v1

1. We propose a GNN-guided deformable attention module that introduces the GNN into interest point generation process. The GNN captures the relationships between different regions of an image, enabling the displacement of interest points to better align with aesthetic structures. This design addresses the lack of guidance in traditional deformable attention.
2. We propose an improved deformable attention model that utilizes neighborhood attention as a feature extractor, enhancing the model's ability to extract aesthetic features. Compared to the sliding window attention mechanism, it achieves superior performance in terms of accuracy.
3. Extensive experiments on public aesthetic datasets confirm the effectiveness of our model, comparing to recent Transformer-based and composition-aware aesthetic assessment models.

2. Related Work

2.1. Image Aesthetics Assessment

Image Aesthetic Assessment (IAA) aims to simulate human aesthetic judgment by analyzing visual factors such as composition, color, and style to generate an aesthetic score for an image. The development of IAA method can be broadly categorized into two stages:

2.1.1. Handcrafted Feature-Based Methods

When the concept of IAA was first proposed, mainstream research methods primarily relied on manually selecting features and designing handcrafted extraction techniques. For instance, Datta *et al.* [20] analyzed scoring trends and heuristics to identify features such as tone, texture, and image composition as scoring attributes. Obrador *et al.* [21] defined features based on photographic rules, such as sharpness and hue, specifically for photographic images, and constructed categorized datasets to study the impact of these features across different classifications. Tang *et al.* [22] extracted features from both the subject and background, combining them with the global features of the image to evaluate its aesthetic quality. These designed image features focus on specific aspects of the image; however, the limitations of feature design hinder the ability to accurately and comprehensively score images.

2.1.2. Deep Learning-Based Methods

With the continuous development of deep learning, researchers have increasingly employed various neural networks for IAA. The emergence of large-scale aesthetic datasets has further established these methods as the mainstream method for image aesthetic evaluation. Duan *et al.* [23] considered the semantic attributes of images, using several attention-generating networks to obtain attention maps for different semantic attributes, resulting in semantically enhanced feature representation for aesthetic scoring predictions. Hou *et al.* [24] improved the performance of image aesthetic evaluation by employing knowledge distillation, using a pre-trained object classification (POC) model as a feature extractor. Shi *et al.* [25] utilized multi-reference learning to consider the consistency between similar images, deriving aesthetic scores based on fine-grained aesthetic features. Zhou *et al.* [26] introduced multi-modal large language models (MLLMs) into the IAA task to generate high-quality textual descriptions, addressing the issue of noise present in current datasets. However, these methods do not take into account the impact of image composition on the IAA task.

2.2. Vision Transformer (ViT)

Since the introduction of ViT [27], the Transformer has been integrated into the field of image, such as image generation [28] and image segmentation [29]. Following this, numerous improvements have been proposed. Wu *et al.* proposed CvT [30], which incorporates convolution into the Transformer model, retaining the advantages of both CNNs and Transformer. Zhu *et al.* [31] referenced the mechanism of deformable convolutions to propose deformable detection transformer,

 doi:10.20944/preprints202602.1595.v1

combining the sparse spatial sampling of deformable convolutions with the relational modeling capabilities of Transformer.

However, traditional Transformer-based methods face challenges such as high computational complexity and inefficiency when applied to visual tasks. To handle this challenge, Liu *et al.* [32] proposed the Swin Transformer, which introduces a hierarchical model and a sliding window mechanism, achieving a balance between performance and efficiency. This method has been extensively applied in image classification, object detection, and related tasks. Subsequently, many variants of Swin have emerged. Hassani *et al.* [19] fixed self-attention to the nearest neighboring pixels, proposing a simple and flexible explicit sliding window attention mechanism called neighborhood attention. Xia *et al.* [33] incorporated a deformable attention mechanism into the Swin transformer model, proposing the deformable attention transformer (DAT), which enhances the flexibility and efficiency of the attention patterns while improving the ability to model long-range relationships. Zhang *et al.* [34] introduced a innovative quadrilateral attention module that extends window-based attention to quadrilateral shapes. This design allows the network to represent targets of various shapes and orientations while capturing rich contextual information.

Currently, Transformer-based methods are widely adopted in IAA tasks. Lan *et al.* [13] extract both emotional and aesthetic features from images and Transformer to explore the relationship between aesthetics and emotion. Ke *et al.* [12] propose a two-stream IAA model that combines Transformer and CNN features, effectively integrating the strengths of both models to achieve improved performance. He *et al.* [11] introduce a deformable mechanism to balance foreground and background features. However, these methods generally lack explicit guidance from image composition principles.

2.3. Graph Neural Network

Currently, an increasing number of studies focus on processing graph-structured data, such as relational information between locations, personal user information for recommendation algorithms, and molecular modeling in chemistry and pharmacology. Research on graph-structured neural networks is also on the rise, leading to the emergence of Graph Neural Networks (GNN) [35]. The GNN-based methods can be broadly classified into two categories: the first category is spatial-based methods, with early forms proposed by Micheli *et al.* [36]. The second category is spectral-based methods, first introduced by Bruna *et al.* [37]. Researchers have modeled images as graph structures by dividing them into patches [9] or extracting regional features [7], enabling the application of GNN to process image. In recent years, the GNN has gained popularity among computer vision researchers because it can effectively handle irregular information and non-Euclidean data. Hao *et al.* [38] utilized GNN to extract and integrate 3D information from CT slices to differentiate cases of parapneumonic effusion. Zheng *et al.* [39] proposed an efficient point cloud analysis framework called PointViG, which leveraged the powerful capabilities of GNN to handle complex datasets, thereby ensuring high performance while controlling computational costs.

In recent studies, the GNN has been applied to IAA tasks due to its excellent global modeling capabilities. Ghosal *et al.* [9] modeled input images as graph structures, preserving their original aspect ratio and resolution, thereby utilizing GNN to extract aesthetic information related to aspect ratio and layout for scoring. Liu *et al.* [7] employed fully convolutional networks (FCN) to map local regions of the input image to different 3D features, representing the image as a graph composed of regions, then they used graph convolution operations to capture the aesthetic features of the image. She *et al.* [8] proposed a hierarchical layout-aware graph convolutional network, utilizing two GCN modules. The first module constructs an aesthetically relevant graph structure in the coordinate space and performs inference on spatial nodes, while the second module aggregates salient features for graph inference. The final aesthetic features are obtained by combining the outputs of both modules.

 doi:10.20944/preprints202602.1595.v1

3. Method

In this section, a detailed introduction to our IAA model is provided. First, we introduce the GNN-based deformable attention module (GAT). Then, we describe the overall architecture of our model and the improvements designed to enhance its performance.

3.1. GAT Module

Figure 2 illustrates the internal architecture of the GAT module. After the image features pass through the interest points generation module and offsets generation module, the interest points and offsets are obtained, and their superposition yields the final coordinates. Bilinear interpolation is then applied to extract sampling features. Finally, multi-head attention processes sampling features and produce the output features. Next, we explain the principles of each module in detail.

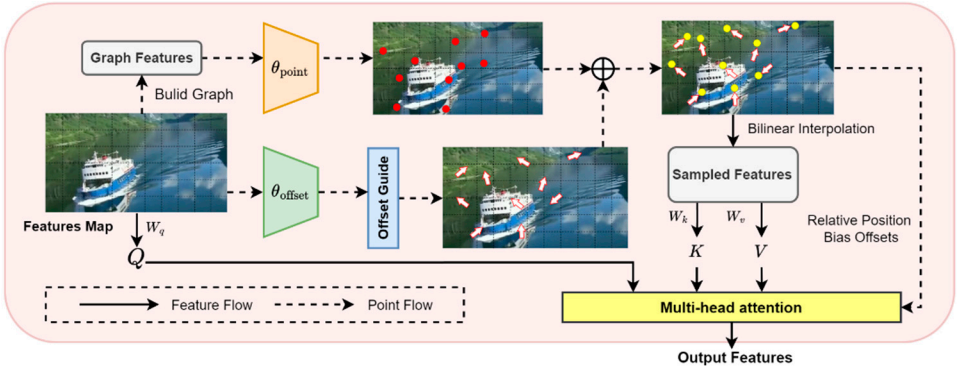


Figure 2. The internal structure of the GAT block presents the developed deformable attention mechanism. It consists of several key components, such as the interest point generation module, offset generation module, and offset guidance module.

3.1.1. Interest Points Generation

We employ a GNN-based module to generate the initial interest point coordinates. The input features X of each GAT module have size $X \in \mathbb{R}^{H\times W\times 3}$. To conform to the processing method of GNN, we first construct a feature matrix. After dividing the image into N patches, each image patch is transformed into a feature vector $\mathbf{x}_i \in \mathbb{R}^{H\times W\times 3}, i = 1,2,\dots,N$. We treat these feature vectors as nodes, resulting in a node set $V = \{v_1, v_2, \dots, v_N\}$. Next, we use a Dilated K-Nearest Neighbors algorithm [40] to find the K nearest neighbors $V'(v_i)$, for each node v_i, establishing a directed edge e_{ij} for each $v_j \in V'(v_i)$. This yields a graph-structured representation of the feature data $G(X) = (V, E)$, where V is the set of nodes and E is the set of edges. Finally, we input this graph structure into the GNN module θ_{point}. This module includes a custom dynamic graph convolutional layer and a fully connected layer that maps features to coordinates, outputting the initial interest point coordinates. To ensure numerical stability and facilitate uniform processing with the subsequently generated offsets, the coordinates are normalized to the range $[-1, +1]$ using the tanh function. This process can be expressed as follows:

$$(P_x, P_y) = \tanh\left(\theta_{\text{Point}}(G(X))\right) \tag{1}$$

3.1.2. Interest Offsets Generation

The input features Xundergo a linear transformation via the matrix W_q to obtain the query token q. Then, q is input into the interest point offset module θ_{offset} to obtain the offset value (O_x, O_y). Our constructed offset generation network consists of two convolutional layers. The first layer extracts basic features related to the offsets. After activation and normalization, these features

 doi:10.20944/preprints202602.1595.v1

are passed to the second layer, which refines them and adjusts dimensions to produce the offset coordinates.

In traditional tasks, such as super-resolution [41] and object detection [25], the interest points of the deformable attention typically focus only on salient regions. However, this method may overlook the overall layout features of the image. Therefore, we add a guiding module $Guide$ to fine-tune the offsets in accordance with aesthetic properties. Specifically, we calculate whether the interest points and offset values lie within the same quadrant. If they do not, we multiply the offset values by a coefficient $\eta \in (0,1)$ to reduce the magnitude of the offsets. This restriction helps prevent the interest points from clustering solely in the salient regions of the image, allowing for the extraction of the overall aesthetic feature information. The entire interest point generation process can be expressed as follows:

$$(O_x, O_y) = \text{Guide}\left(\theta_{\text{offset}}\left(W_q(X)\right)\right) \tag{2}$$

3.1.3. Feature Extraction

After obtaining the interest points (P_x, P_y) and the offsets (O_x, O_y), as shown in **Figure 2**, we add the offsets to the interest points to obtain the new coordinates (P'_x, P'_y), which represent the coordinates of the sampling points. We then use bilinear interpolation sampling to obtain the feature X'. The feature X' is processed through matrices W_k and W_v to get the deformed query token $k$and value token v. Next, we use a multi-head attention mechanism to aggregate the features. Here, we add relative position bias offsets to enhance the attention.

Existing studies have shown that introducing positional encoding into the computation of the attention mechanism can significantly improve model performance. For example, the Conditional Position encoding Vision Transformer (CPVT) proposed in reference [42] dynamically generates positional encodings to enhance the Transformer's effectiveness. Since the deformable attention method computes coordinate information through queries before the multi-head attention operation, we can directly use this positional information to enhance the attention without additional computation. The operation on the n-th attention head in multi-head attention can be expressed as follows:

$$\text{Output}_n = \text{Softmax}\left(\frac{q_n k_n^{\top}}{\sqrt{d}} + \phi(\hat{B}; R)\right) v_n \tag{3}$$

Specifically, we construct a relative position bias table $\hat{B}$of size $(2H-1) \times (2W-1)$. We then obtain the relative displacement $R = P' - P$ based on the previously calculated interest points P and the adjusted interest points P'. Based on this displacement coordinate, we sample using bilinear interpolation in $\hat{B}$, denoted as $\phi(\hat{B}; R)$. After the computation, this is added to the multi-head attention operation to enhance the features. After these processes, we obtain the output features of the GAT module.

3.2. Method Architecture

3.2.1. Architecture of Each Transformer Block

The preprocessing and postprocessing architecture of each attention block is consistent, as shown in Figure 3. The input features first pass through a Local Perception Unit (LPU) [43], which is a module based on Depth-wise Convolution (DWConv). A residual connection is used to process the features before and after, capturing both global and local information. This process can be expressed as follows:

This is example 1 of an equation:

$$\text{LPU}(X) = \text{DWConv}(X) + X \tag{4}$$

 doi:10.20944/preprints202602.1595.v1

Next, the features are processed by the attention module, which similarly aggregates features using a residual structure. Before being input into the next Transformer block, the features pass through a multi-layer perceptron (MLP). Unlike the typical MLP, we also add a DWConv residual structure before the activation layer to further enhance local feature modeling. This output is then fused with the output of the attention module, improving the capability to represent overall compositional features.

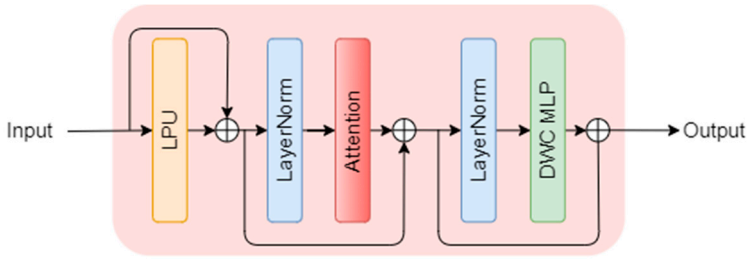


Figure 3. This figure shows the detailed architecture of each Transformer block in the GAT. Each attention block is processed in the same way.

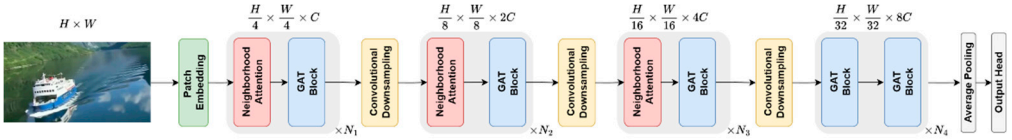
3.2.2. Neighborhood Attention

Before inputting the features into the deformable attention block, we first perform basic feature extraction using a standard attention block. This method preliminarily extracts local features, providing effective feature representations for the deformable attention operation, which significantly improves computational efficiency without sacrificing performance. Additionally, from an aesthetic standpoint, early features may struggle to distinguish the structural layout between the foreground and background, making it challenging for deformable attention to shift features to the desired locations.

Current research includes some methods that use local attention for feature extraction [11] and others that employ sliding window attention [44]. However, for aesthetic evaluation tasks, these methods fail to effectively capture features related to aesthetic attributes, as they do not account for the interconnections between different parts of the image and inadequately focus on corner features, which results in the loss of structural information. Given these limitations, we utilize neighborhood attention [19] for feature extraction, which is a simple and flexible improvement over sliding window attention. The attention range for each pixel is localized to its nearest neighborhood. As computational costs increase, it gradually approaches self-attention while still allowing the receptive field to expand. Different from blocked and window self-attention, neighborhood attention can maintain translational equivariance and attend to corner features. In the experiments in the next section, we demonstrate the effectiveness of this module.

3.2.3. Overall Architecture

The overall architecture of the model is illustrated in Figure 4. The input image is first divided into patches and embedded into a sequence of patch tokens. These tokens are then processed through four hierarchical stages of feature extraction. Each stage consists of stacked attention modules tailored for aesthetic feature modeling. Between stages, convolutional down-sampling layers reduce spatial resolution while increasing semantic richness. The final feature representation is aggregated via global average pooling and passed through a regression head to predict the aesthetic score of the image.



 doi:10.20944/preprints202602.1595.v1

Figure 4. Overview of the GAT model. N_1 to N_4 indicate the number of times the module combination is repeated in each stage.

4. Experiments

4.1. Dataset and Evaluation Metrics

To validate our method, we conduct experiments on two IAA datasets: AVA [15], and TAD66K [45].

AVA: Currently the largest publicly available aesthetic dataset, it contains approximately 255,000 images sourced from the DPChallenge website, covering over 60 categories of semantic labels and photography style-related tags. Each image is rated by an average of 210 users, with scores ranging from 1 to 10.

TAD66K: This is a theme-oriented aesthetic dataset containing 66,000 carefully selected images, covering 47 popular themes. It features extensive annotations, with each image having at least 1,200 valid labels. Each image is given an aesthetic score ranging from 1 to 10.

These datasets mentioned above are divided into training set, testing set, and validation set with a ratio of 90:5:5. The initial learning rate during training is set to 1e-5, optimized using the Adam optimizer. Since AVA provides score distributions, we use the Earth Mover's Distance (EMD) loss for training on this dataset to predict score distributions. The TAD66K dataset provides real-valued scores, so we use Mean Squared Error (MSE) as the loss for training. To ensure the generalization ability of our method, we pre-trained the model on ImageNet-1K dataset [46].

The effectiveness of the proposed method is evaluated using three widely adopted metrics in image aesthetics assessment: Accuracy (ACC), Pearson Correlation Coefficient (PLCC), and Spearman's Rank Correlation Coefficient (SRCC) [47]

For the accuracy, a binary classification strategy is employed by setting the midpoint of the score range as the threshold: samples with scores above the midpoint are categorized as high quality, while those below are considered low quality. The accuracy is calculated as $\mathrm{ACC} = \frac{N_c}{N_a}$, where N_c is the number of correctly predicted samples and N_a is the total number of samples. The SRCC measures the monotonicity of the predictions, while the PLCC assesses the precision of the predicted scores. Additionally, we compare the mean squared error (MSE) of the models on the AVA dataset, as this metric is reported in several existing works.

Table 1. Comparison of Experimental Results and Evaluation Metrics on the AVA Dataset.

Method	Year	ACC↑	PLCC↑	SRCC↑	EMD1↓	EMD2↓	MSE↓
A-Lamp[52]	2017	82.5%	0.671	0.666	-	-	-
NIMA[58]	2018	81.5%	0.636	0.612	0.050	-	-
AFDC[53]	2020	83.2%	0.671	0.649	0.00447	-	0.270
HLAGCN[8]	2021	84.1%	0.678	0.656	0.045	0.065	0.264
MUSIQ[48]	2021	81.5%	0.738	0.726	-	-	**0.242**
SA-IAA[23]	2022	-	-	0.740	-	-	-
GATP[9]	2022	76.4%	0.762	-	-	-	-
[54]	2022	-	0.707	0.685	-	-	-
Maxvit[49]	2022	80.8%	0.733	0.732	0.0439	-	0.263
EAT[11]	2023	81.7%	**0.770**	**0.759**	-	-	-
Tavar[55]	2023	**85.1%**	0.736	0.725	-	-	-
SPTF-CNN[12]	2023	84.5%	0.709	0.687	0.043	0.064	0.264
DAT++[50]	2023	81.8%	0.742	0.733	-	0.074	-
CADAS[51]	2024	85.0%	0.702	0.687	**0.042**	**0.061**	-
UNIAA[10]	2024	-	0.704	0.713	-	-	-
Ours	-	82.3%	0.750	0.741	**0.042**	0.072	0.243

 doi:10.20944/preprints202602.1595.v1

4.2. Experimental Results

Table 1 presents our experimental results on AVA dataset, comparing our method with several classic methods, including those using Transformers for IAA [11,12,48–51] and those considering layout factors with graph convolution for IAA [8,9,12,23,52–55]. As the experimental results show, our method achieves overall performance. Specifically, our approach achieves the best EMD1. Furthermore, it attains a high PLCC and SRCC, which are among the top results. In terms of MSE, our method reaches 0.243, which is extremely close to the best value. Our method achieves a more balanced overall performance. For example, although we did not achieve higher results than [11] in PLCC and SRCC, our accuracy exceeds that model. This improvement can be attributed to the advantages of the attention mechanism, as well as the integration of image composition information into the model via the GNN. Specifically, the GNN guides the generation of initial reference points, shifting the attention regions toward areas that better align with aesthetic composition, thereby improving the accuracy of the IAA task and overall performance. In Figure 5, we illustrate the scoring outcomes of our model, which effectively simulates human aesthetic preferences by outputting scores close to the ground truth (GT) based on image features and composition. The Transformer architecture shares structural similarities with GNN, as both enhance feature extraction by modeling relationships between elements. Our proposed Transformer, integrated with GNN, further strengthens this capability, leading to improved performance.

Table 2. Comparison of Experimental Results and Evaluation Metrics on the TAD66K Dataset.

Method	ACC↑	PLCC↑	SRCC↑
A-Lamp[48]	-	0.422	0.411
MUSIQ[51]	-	<u>0.517</u>	<u>0.489</u>
Maxvit[53]	-	0.513	0.484
DAT++[55]	67.5%	0.514	0.486
AesMamba[57]	**72.0%**	0.511	0.483
[25]	65.0%	0.475	0.452
Ours	<u>68.7%</u>	**0.521**	**0.490**

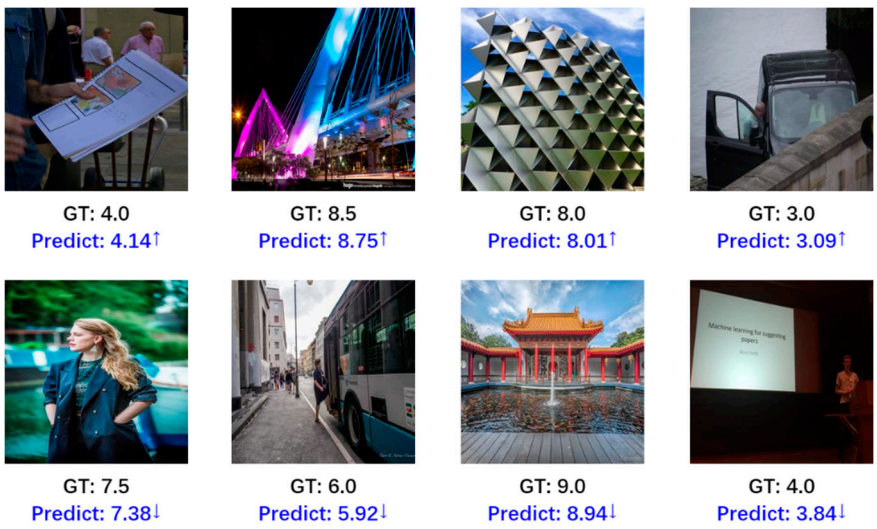


Figure 5. Comparison of model prediction scores and GT.

In This improvement can be attributed to the advantages of the attention mechanism, as well as the integration of image composition information into the model via the GNN. Specifically, the GNN guides the generation of initial reference points, shifting the attention regions toward areas that better align with aesthetic composition, thereby improving the accuracy of the IAA task and overall performance. In Figure 5, we illustrate the scoring outcomes of our model, which effectively simulates

 doi:10.20944/preprints202602.1595.v1

human aesthetic preferences by outputting scores close to the ground truth (GT) based on image features and composition. The Transformer architecture shares structural similarities with GNN, as both enhance feature extraction by modeling relationships between elements. Our proposed Transformer, integrated with GNN, further strengthens this capability, leading to improved performance.

Table 2. Comparison of Experimental Results and Evaluation Metrics on the TAD66K Dataset.

Method	ACC↑	PLCC↑	SRCC↑
A-Lamp [48]	-	0.422	0.411
MUSIQ [51]	-	<u>0.517</u>	<u>0.489</u>
Maxvit [53]	-	0.513	0.484
DAT++[55]	67.5%	0.514	0.486
AesMamba [57]	**72.0%**	0.511	0.483
[25]	65.0%	0.475	0.452
Ours	<u>68.7%</u>	**0.521**	**0.490**

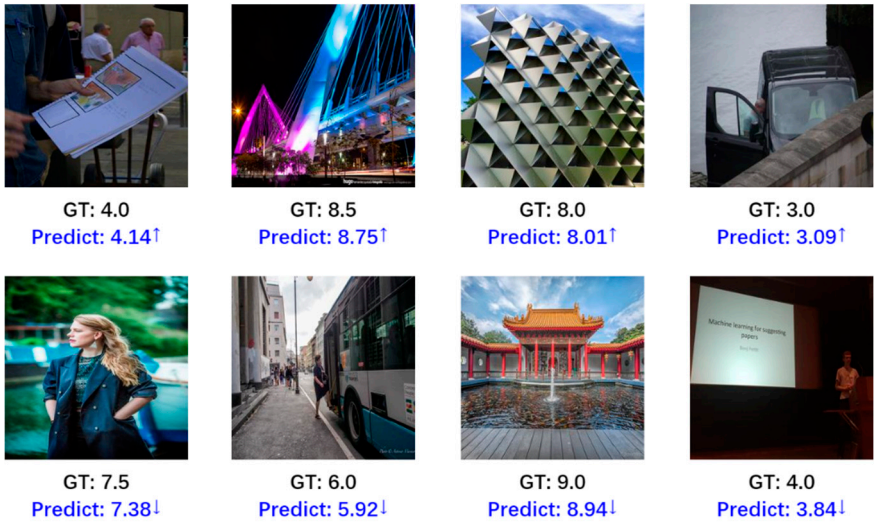


Figure 5. Comparison of model prediction scores and GT.

we further compare our method's performance on the TAD66K dataset. Unlike traditional aesthetic datasets, TAD66K provides more granular scores by performing classification and scoring based on different themes, which poses greater challenges to existing IAA methods [45]. The results demonstrate that our method performs well on this dataset. Specifically, our method achieves second-best accuracy alongside high PLCC and SRCC and outperforms both the traditional layout-based method [52] and Transformer-based models [48,49]. Moreover, compared with recent methods such as Mamba [56], our model still maintains a better performance. This can be attributed to the fact that the TAD66K dataset provides aesthetic scores across multiple categories without a unified standard. The incorporation of composition information enhances the feature expressiveness, allowing the model to more effectively perceive the aesthetic characteristics of the images.

 doi:10.20944/preprints202602.1595.v1

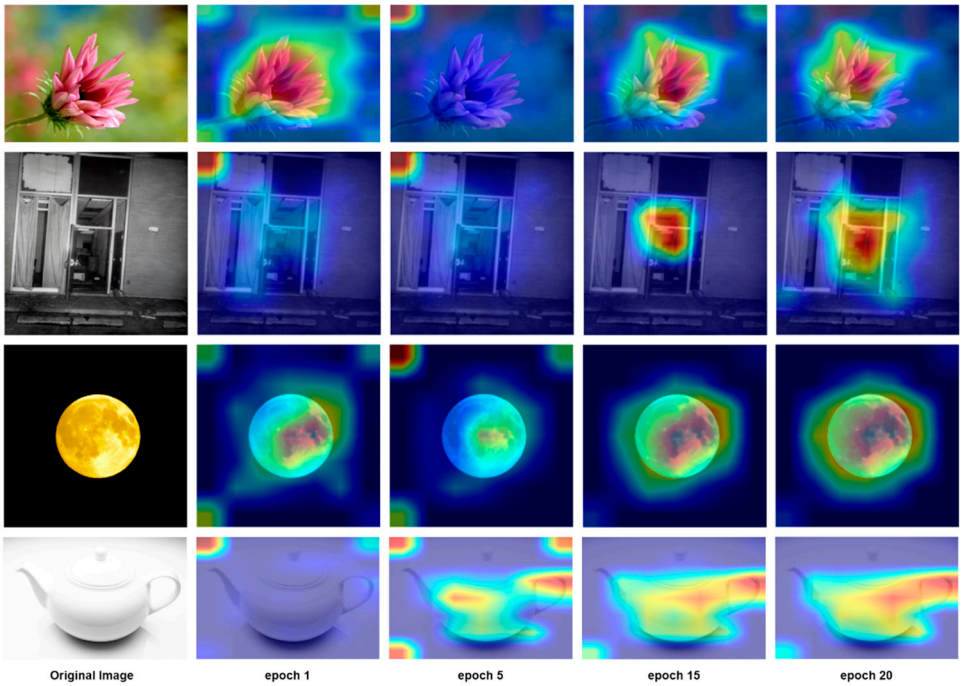


Figure 6. Original input images and corresponding GradCAM maps at different training epochs.

We further employ GradCAM [57] to visualize the model's learning process during training. As shown in Figure 6, the regions of interest gradually become more aligned with aesthetic properties as the number of training epochs increases. We observe that, in the early stages of training, the model primarily attends to image boundaries and corners. As training progresses, attention shifts to more aesthetic meaningful regions, which may possibly due to suboptimal initialization from the pre-trained weights. Through observation, we found that in early training stages, interest points were mainly concentrated on image boundaries and corners. As training progressed, they gradually shift to more compositionally meaningful regions, likely due to the influence of pre-trained weights.

We also utilize the style labels from the AVA dataset to classify and analyze the scoring results of the GAT method. A stricter evaluation criterion is applied: if the difference between the model-predicted score and the GT is less than a threshold of 0.5, the score is considered reliable.

The analysis results are shown in Figure 7. It can be observed that our method performs well in composition and composition-related styles. For instance, the reliability rate for the Rule of Thirds label achieves 85.34%, while the Shallow DOF label achieves 83.28%. Other labels similarly attain reliability rates around 80%. However, the Soft Focus label exhibit a lower reliability rate of only 70.95%. We speculate that this is due to the lack of clear edges in Soft Focus images, which causes inaccuracies in locating offset points and consequently reduces scoring accuracy.

 doi:10.20944/preprints202602.1595.v1

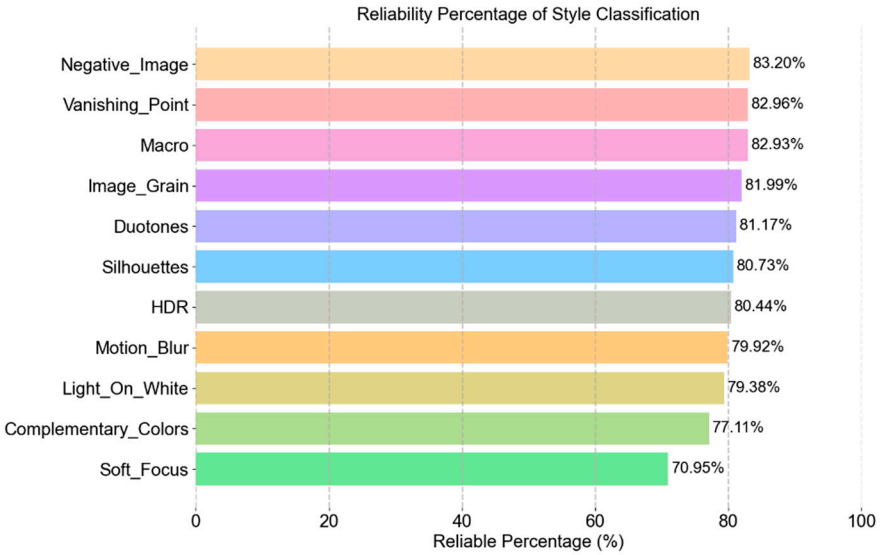


Figure 7. Histogram results of the style analysis, a threshold of 0.5 was applied.

For consumer electronic devices, efficiency is crucial. Inference efficiency affects user experience, while training efficiency impacts iteration speed and development difficulty. Our method is significantly more efficient than existing transformer-based approaches. To validate this, we train 10,000 images using three transformer variants under identical conditions. The dataset is split into training, test, and validation sets (90:5:5), and each model is trained for 10 epochs. The results are presented in Figure 8. It is evident that our method requires significantly less training time compared to Swin attention, and also improves upon local attention in terms of efficiency. While Swin attention achieves higher performance than local attention, it incurs a substantially greater time cost. Our method not only reduces training time relative to local attention but also surpasses Swin attention in overall performance. This will be further demonstrated in the subsequent experiments. We also evaluate inference speed by averaging the inference time over 500 images. The average inference time per image is 64.51 ms for Swin attention, 50.40 ms for local attention and 44.35 ms for our method. As the result shows, our method achieves faster inference. This improvement benefits from the neighborhood attention mechanism, which simplifies computation while maintaining strong performance.

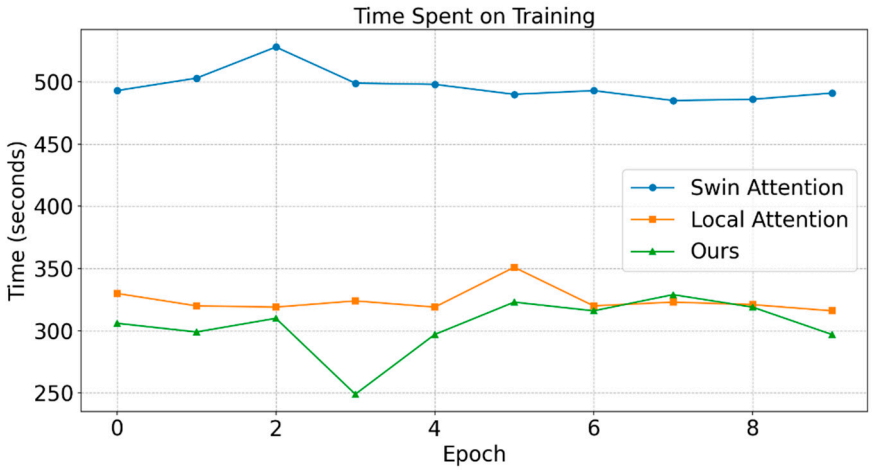


Figure 8. Training time per epoch comparison among three methods. Our proposed method consistently demonstrates better computational efficiency.

4.3. Ablation Study

 doi:10.20944/preprints202602.1595.v1

To investigate the effect of the proposed modules and demonstrate the effectiveness of our method, we conducted ablation studies on the TAD66K dataset to analyze the contributions of different components.

4.3.1. Neighborhood Attention

We replace the neighborhood attention module in our model with local attention and Swin attention, and the comparison results are present in Table 3. As shown, both Swin attention and neighborhood attention improve performance, with the latter yielding a more significant improvement. This demonstrates the effectiveness of neighborhood attention in feature extraction.

Table 3. Ablation Study on Different Transformer Blocks.

Block Type	ACC↑	PLCC↑	SRCC↑
local attention	67.0%	0.503	0.474
Swin attention	67.3%	0.510	0.482
neighborhood attention	**68.7%**	**0.521**	**0.490**

4.3.2. LPU and DWConv MLP

We conduct four groups of comparative experiments: using only the LPU, using only the DWConv MLP, using neither of the two modules, and using both. When the DWConv MLP is not used, it is replaced by a standard MLP, which is widely adopted in current Transformer models [11]. And LPU is directly removed without substitution.

As shown in Table 4, removing either the LPU or the DWConv MLP result in varying degrees of performance degradation, demonstrating the effectiveness of both modules.

Table 4. Ablation Study on the Effectiveness of Two Modules.

LPU	DWConv MLP	ACC↑	PLCC↑	SRCC↑
✓	-	64.9%	0.417	0.393
-	✓	61.2%	0.467	0.441
-	-	60.5%	0.376	0.354
✓	✓	**68.7%**	**0.521**	**0.490**

4.3.3. GAT Block

For GAT blocks, we compare our proposed GNN-based interest points generation method with two alternative methods: convolution-based generation and uniform sampling. The results are presented in Table 5.

As the results show, the proposed GNN module outperform both the convolution-based and uniform generation methods. This demonstrates that our graph convolutional module effectively incorporates layout features into the generation process, leading to improved performance. Furthermore, from the results we observe that the convolution-based method performs better than uniform sampling, indicating that leveraging feature information to guide the localization of interest points is beneficial for the IAA task.

Table 5. Ablation Study on Different Interest Points Generation Methods.

Generator	ACC↑	PLCC↑	SRCC↑
Conv	67.7%	0.517	0.488
Uniform	67.5%	0.512	0.482
GNN	**68.7%**	**0.521**	**0.490**

Furthermore, the visualization study in Figure 6 reveals that, the deformable attention mechanism tends to produce a consistent distribution of initial interest points across different inputs.

 doi:10.20944/preprints202602.1595.v1

In different scenarios, these initial points form a general layout pattern that best fits the scenarios. Regardless of the input image, the spatial structure of the generated interest points remains largely similar. The final sampling locations are then obtained by applying learned offsets to these initial points.

The GNN exhibits strong layout-awareness and has capability of modeling the relationships among different regions of an image during training. These properties provide an advantage for tasks involving composition. As a result, the proposed interest points generation module achieves competitive performance.

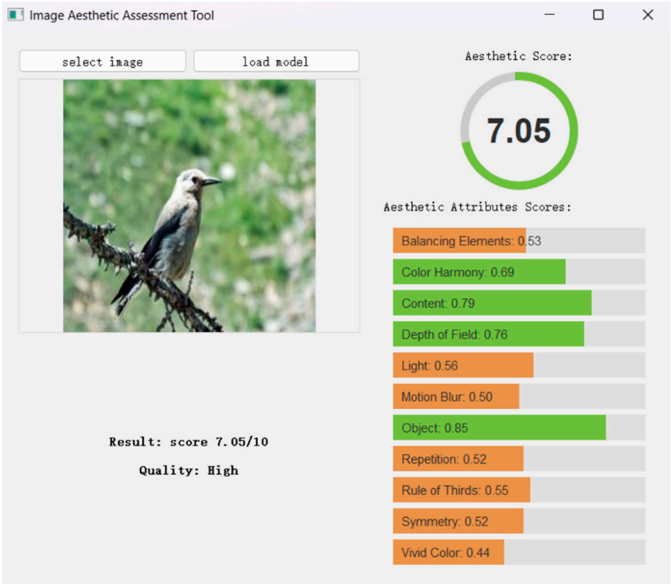


Figure 9. Schematic of multi-attribute aesthetic scoring results.

4.3.4. GAT Block

We further utilize the proposed GAT to design a multi-attribute aesthetic scoring tool, which assists users in electronic photography by providing multi-attribute aesthetic scores. By modifying the output layer in GAT, we enable it to predict both overall aesthetic scores and individual attribute scores, based on the annotations on the datasets. The model is trained using Mean Squared Error (MSE) loss for the multi-attribute output and aesthetic score output. After training, the model was integrated into the application. As illustrated in the Figure 9, the users select an image for evaluation. The tool then performs IAA scoring and provides both the overall aesthetic score and multi-attribute aesthetic scores. These attributes encompass various aspects of photography, including color, composition, and photographic techniques. This tool can be applied in various scenarios. As illustrated in the Figure 10. Application scenarios of IAA in different scenarios. it can be used for horizontal comparisons to evaluate the strengths and weaknesses of different photos within a task, helping to optimize the output. It can also support vertical comparisons to identify suitable photographic subjects and develop appropriate shooting plans. By referring to the corresponding attribute scores, users can refine their photographic techniques, enhance image quality, and ultimately produce works that align with their creative intentions.

 doi:10.20944/preprints202602.1595.v1

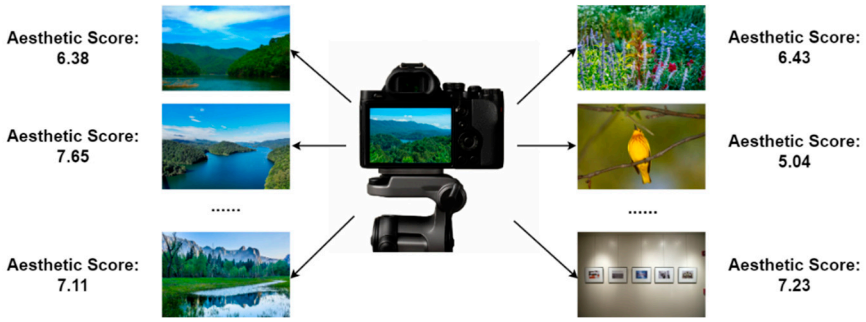


Figure 10. Application scenarios of IAA in different scenarios.

5. Conclusions

In this paper, we propose a GNN-guided deformable attention model for image aesthetic assessment. Specifically, we propose an improved deformable attention module. This module leverages a GNN to integrate compositional information and guides the interest points in deformable attention toward regions that better align with compositional intent, thereby enabling the extracted features to conform more closely to aesthetic principles. Based on this module, we further design an aesthetic scoring model, which achieves promising performance on several aesthetic benchmarks. In addition, we develop a multi-attribute aesthetic scoring application using this method, which can be leveraged to assist users in enhancing their photography. Our model can be integrated into consumer electronics to assist users in enhancing their photography or in filtering images more effectively.

Author Contributions: Conceptualization, Z.J. and J.M.; methodology, Z.J.; software, J.M.; validation, Z.J.; formal analysis, L.L.; investigation, Z.J.; resources, L.L.; data curation, F.J.; writing—original draft preparation, Z.J.; writing—review and editing, L.L.; visualization, Z.J.; supervision, L.L. and J.M.; project administration, L.L.; funding acquisition, L.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
IAA	Image Aesthetic Assessment
GNN	Graph Neural Network
CNN	Convolutional Neural Network
ViT	Vision Transformer
DAT	Deformable Attention Transformer
GAT	GNN-based Deformable Attention Module
ACC	Accuracy
PLCC	Pearson Linear Correlation Coefficient
SRCC	Spearman's Rank Correlation Coefficient
GT	Ground Truth
MSE	Mean Squared Error
DWConv	Depth-wise Convolution
GradCAM	Gradient-weighted Class Activation Mapping

 doi:10.20944/preprints202602.1595.v1

References

1. Jiang, M.; Shen, L.; Zheng, L.; Zhao, M.; Jiang, X. Tone-Mapped Image Quality Assessment for Electronics Displays by Combining Luminance Partition and Colorfulness Index. *IEEE Trans. Consum. Electron.* 2020, *66*, 153–162.
2. Jiang, M.; Shen, L.; Hu, M.; An, P.; Ren, F. Blind Quality Evaluator of Tone-Mapped HDR and Multi-Exposure Fused Images for Electronic Display. *IEEE Trans. Consum. Electron.* 2021, *67*, 350–362.
3. Biswas, S.; Barma, S. A Low-Cost Vegetable Quality Assessment System Based on Microscopy Images in Deep Learning Edge Computing: A Pilot Study on Potato Tuber. *IEEE Trans. Consum. Electron.* 2024, *70*, 6343–6353.
4. Mirzaei, S.; Tohidypour, H.R.; Nasiopoulos, P.; Vora, S.R.; Mirabbasi, S. An Advanced Denoising Technique for Low-Dose CBCT Imaging: Enhancing Image Quality and Consumer Safety in Dental Diagnostics. *IEEE Trans. Consum. Electron.* 2025.
5. Hong, R.; Zhang, L.; Tao, D. Unified Photo Enhancement by Discovering Aesthetic Communities from Flickr. *IEEE Trans. Image Process.* 2016, *25*, 1124–1135.
6. Wang, W.; Shen, J.; Ling, H. A Deep Network Solution for Attention and Aesthetics Aware Photo Cropping. *IEEE Trans. Pattern Anal. Mach. Intell.* 2018, *41*, 1531–1544.
7. Liu, D.; Puri, R.; Kamath, N.; Bhattacharya, S. Composition-Aware Image Aesthetics Assessment. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision; 2020; pp. 3569–3578.
8. She, D.; Lai, Y.-K.; Yi, G.; Xu, K. Hierarchical Layout-Aware Graph Convolutional Network for Unified Aesthetics Assessment. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021; pp. 8475–8484.
9. Ghosal, K.; Smolic, A. Image Aesthetics Assessment Using Graph Attention Network. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR); 2022; pp. 3160–3167.
10. Zhou, Z.; Wang, Q.; Lin, B.; Su, Y.; Chen, R.; Tao, X.; Zheng, A.; Yuan, L.; Wan, P.; Zhang, D. Uniaa: A Unified Multi-Modal Image Aesthetic Assessment Baseline and Benchmark. *ArXiv Prepr. ArXiv240409619* 2024.
11. He, S.; Ming, A.; Zheng, S.; Zhong, H.; Ma, H. Eat: An Enhancer for Aesthetics-Oriented Transformers. In Proceedings of the Proceedings of the 31st ACM international conference on multimedia; 2023; pp. 1023–1032.
12. Ke, Y.; Wang, Y.; Wang, K.; Qin, F.; Guo, J.; Yang, S. Image Aesthetics Assessment Using Composite Features from Transformer and CNN. *Multimed. Syst.* 2023, *29*, 2483–2494.
13. Lan, G.; Xiao, S.; Yang, J.; Zhou, Y.; Wen, J.; Lu, W.; Gao, X. Image Aesthetics Assessment Based on Hypernetwork of Emotion Fusion. *IEEE Trans. Multimed.* 2023, *26*, 3640–3650.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* 2017, *30*.
15. Murray, N.; Marchesotti, L.; Perronnin, F. AVA: A Large-Scale Database for Aesthetic Visual Analysis. In Proceedings of the 2012 IEEE conference on computer vision and pattern recognition; IEEE, 2012; pp. 2408–2415.
16. Kong, S.; Shen, X.; Lin, Z.; Mech, R.; Fowlkes, C. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. In Proceedings of the Computer Vision–ECCV 2016; 2016; pp. 662–679.
17. Chen, Z.; Zhu, Y.; Zhao, C.; Hu, G.; Zeng, W.; Wang, J.; Tang, M. Dpt: Deformable Patch-Based Transformer for Visual Recognition. In Proceedings of the Proceedings of the 29th ACM international conference on multimedia; 2021; pp. 2899–2907.
18. Wei, X.; Yin, L.; Zhang, L.; Wu, F. DV-DETR: Improved UAV Aerial Small Target Detection Algorithm Based on RT-DETR. *Sensors* 2024, *24*, 73–76.
19. Hassani, A.; Walton, S.; Li, J.; Li, S.; Shi, H. Neighborhood Attention Transformer. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2023; pp. 6185–6194.
20. Datta, R.; Joshi, D.; Li, J.; Wang, J.Z. Studying Aesthetics in Photographic Images Using a Computational Approach. In Proceedings of the Computer Vision–ECCV 2006; Springer, 2006; pp. 288–301.

 doi:10.20944/preprints202602.1595.v1

21. Obrador, P.; Saad, M.A.; Suryanarayan, P.; Oliver, N. Towards Category-Based Aesthetic Models of Photographs. In Proceedings of the Advances in Multimedia Modeling: 18th International Conference, MMM 2012, Klagenfurt, Austria, January 4-6, 2012. Proceedings 18; Springer, 2012; pp. 63–76.
22. Tang, X.; Luo, W.; Wang, X. Content-Based Photo Quality Assessment. *IEEE Trans. Multimed.* 2013, *15*, 1930–1943.
23. Duan, J.; Chen, P.; Li, L.; Wu, J.; Shi, G. Semantic Attribute Guided Image Aesthetics Assessment. In Proceedings of the 2022 IEEE International Conference on Visual Communications and Image Processing (VCIP); IEEE, 2022; pp. 1–5.
24. Hou, J.; Ding, H.; Lin, W.; Liu, W.; Fang, Y. Distilling Knowledge from Object Classification to Aesthetics Assessment. *IEEE Trans. Circuits Syst. Video Technol.* 2022, *32*, 7386–7402.
25. Shi, T.; Chen, C.; Li, X.; Hao, A. Semantic and Style Based Multiple Reference Learning for Artistic and General Image Aesthetic Assessment. *Neurocomputing* 2024, *582*, 127434.
26. Zhou, H.; Tang, L.; Yang, R.; Qin, G.; Zhang, Y.; Hu, R.; Li, X. Uniqa: Unified Vision-Language Pre-Training for Image Quality and Aesthetic Assessment. *ArXiv Prepr. ArXiv240601069* 2024.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *ArXiv Prepr. ArXiv201011929* 2020.
28. Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; Freeman, W.T. Maskgit: Masked Generative Image Transformer. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022; pp. 11315–11325.
29. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-Attention Mask Transformer for Universal Image Segmentation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022; pp. 1290–1299.
30. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing Convolutions to Vision Transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2021; pp. 22–31.
31. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable Detr: Deformable Transformers for End-to-End Object Detection. *ArXiv Prepr. ArXiv201004159* 2020.
32. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2021; pp. 10012–10022.
33. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision Transformer with Deformable Attention. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022; pp. 4794–4803.
34. Zhang, Q.; Zhang, J.; Xu, Y.; Tao, D. Vision Transformer with Quadrangle Attention. *IEEE Trans. Pattern Anal. Mach. Intell.* 2024, *46*, 3608–3624.
35. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Philip, S.Y. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* 2020, *32*, 4–24.
36. Micheli, A. Neural Network for Graphs: A Contextual Constructive Approach. *IEEE Trans. Neural Netw.* 2009, *20*, 498–511.
37. Bruna, J.; Zaremba, W.; Szlam, A.; LeCun, Y. Spectral Networks and Locally Connected Networks on Graphs. *ArXiv Prepr. ArXiv13126203* 2013.
38. Hao, J.; Liu, J.; Pereira, E.; Liu, R.; Zhang, J.; Zhang, Y.; Yan, K.; Gong, Y.; Zheng, J.; Zhang, J.; et al. Uncertainty-Guided Graph Attention Network for Parapneumonic Effusion Diagnosis. *Med. Image Anal.* 2022, *75*, 102217.
39. Zheng, Q.; Qi, Y.; Wang, C.; Zhang, C.; Sun, J. PointViG: A Lightweight GNN-Based Model for Efficient Point Cloud Analysis. *ArXiv Prepr. ArXiv240700921* 2024.
40. Li, G.; Muller, M.; Thabet, A.; Ghanem, B. DeepGCNs: Can GCNs Go as Deep as CNNs? In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2019; pp. 9267–9276.

41. Cao, J.; Liang, J.; Zhang, K.; Li, Y.; Zhang, Y.; Wang, W.; Gool, L.V. Reference-Based Image Super-Resolution with Deformable Attention Transformer. In Proceedings of the European conference on computer vision; Springer, 2022; pp. 325–342.
42. Chu, X.; Tian, Z.; Zhang, B.; Wang, X.; Shen, C. Conditional Positional Encodings for Vision Transformers. *ArXiv Prepr. ArXiv210210882* 2021.
43. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. Cmt: Convolutional Neural Networks Meet Vision Transformers. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022; pp. 12175–12185.
44. Shi, P.; Chen, X.; Qi, H.; Zhang, C.; Liu, Z. Object Detection Based on Swin Deformable Transformer-BiPAFPN-YOLOX. *Comput. Intell. Neurosci.* 2023, *2023*, 4228610, doi:https://doi.org/10.1155/2023/4228610.
45. He, S.; Zhang, Y.; Xie, R.; Jiang, D.; Ming, A. Rethinking Image Aesthetics Assessment: Models, Datasets and Benchmarks. In Proceedings of the International Joint Conferences on Artificial Intelligence; 2022; pp. 942–948.
46. 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; Ieee, 2009; pp. 248–255.
47. Daryanavard Chounchenani, M.; Shahbahrami, A.; Hassanpour, R.; Gaydadjiev, G. Deep Learning Based Image Aesthetic Quality Assessment-A Review. *ACM Comput. Surv.* 2025, *57*, 1–36.
48. Ma, S.; Liu, J.; Wen Chen, C. A-Lamp: Adaptive Layout-Aware Multi-Patch Deep Convolutional Neural Network for Photo Aesthetic Assessment. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition; 2017; pp. 4535–4544.
49. Talebi, H.; Milanfar, P. NIMA: Neural Image Assessment. *IEEE Trans. Image Process.* 2018, *27*, 3998–4011.
50. Chen, Q.; Zhang, W.; Zhou, N.; Lei, P.; Xu, Y.; Zheng, Y.; Fan, J. Adaptive Fractional Dilated Convolution Network for Image Aesthetics Assessment. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020; pp. 14114–14123.
51. Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; Yang, F. Musiq: Multi-Scale Image Quality Transformer. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision; 2021; pp. 5148–5157.
52. Celona, L.; Leonardi, M.; Napoletano, P.; Rozza, A. Composition and Style Attributes Guided Image Aesthetic Assessment. *IEEE Trans. Image Process.* 2022, *31*, 5009–5024.
53. u, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. Maxvit: Multi-Axis Vision Transformer. In Proceedings of the Computer Vision – ECCV 2022; 2022; pp. 459–479.
54. Li, L.; Huang, Y.; Wu, J.; Yang, Y.; Li, Y.; Guo, Y.; Shi, G. Theme-Aware Visual Attribute Reasoning for Image Aesthetics Assessment. *IEEE Trans. Circuits Syst. Video Technol.* 2023, *33*, 4798–4811.
55. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Dat++: Spatially Dynamic Vision Transformer with Deformable Attention. *ArXiv Prepr. ArXiv230901430* 2023.
56. Huang, Y.; Li, L.; Chen, P.; Wu, J.; Yang, Y.; Li, Y.; Shi, G. Coarse-to-Fine Image Aesthetics Assessment with Dynamic Attribute Selection. *IEEE Trans. Multimed.* 2024.
57. Gao, F.; Lin, Y.; Shi, J.; Qiao, M.; Wang, N. AesMamba: Universal Image Aesthetic Assessment with State Space Models. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia; 2024; pp. 7444–7453.
58. 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 Proceedings of the IEEE international conference on computer vision; 2017; pp. 618–626.