Theme-aware Visual Attribute Reasoning for Image Aesthetics Assessment

Leida Li, Yipo Huang, Jinjian Wu, Yuzhe Yang, Yaqian Li, Yandong Guo, and Guangming Shi, *Fellow, IEEE*

***Abstract*—People usually assess image aesthetics according to visual attributes, e.g., interesting content, good lighting and vivid color, etc. Further, the perception of visual attributes depends on the image theme. Therefore, the inherent relationship between visual attributes and image theme is crucial for image aesthetics assessment (IAA), which has not been comprehensively investigated. With this motivation, this paper presents a new IAA model based on Theme-Aware Visual Attribute Reasoning (TAVAR). The underlying idea is to simulate the process of human perception in image aesthetics by performing bilevel reasoning. Specifically, a visual attribute analysis network and a theme understanding network are first pre-trained to extract aesthetic attribute features and theme features, respectively. Then, the first level Attribute-Theme Graph (ATG) is built to investigate the coupling relationship between visual attributes and image theme. Further, a flexible aesthetics network is introduced to extract general aesthetic features, based on which we built the second level Attribute-Aesthetics Graph (AAG) to mine the relationship between theme-aware visual attributes and aesthetic features, producing the final aesthetic prediction. Extensive experiments on four public IAA databases demonstrate the superiority of the proposed TAVAR model over the state-of-the-arts. Furthermore, TAVAR features better explainability due to the use of visual attributes.**

***Index Terms*—image aesthetics assessment; visual attribute; image theme; bilevel reasoning**

I. Introduction

The perception of visual aesthetics is an innate ability of human. With the continuous advancement of human-centric visual perception technology, we hope machines can simulate the human aesthetic processes and possess the same perception ability of aesthetics [1]–[4]. Image aesthetics assessment (IAA), which can automatically assess the aesthetic quality of images [5], has been attracting considerable interest due to its extensive applications in such as image retrieval [6], album curation [7], smart photography [8], [9] and image editing [10]. Although people can effortlessly judge the aesthetic quality of images, it remains a great challenge for computational aesthetics models.

This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2021B0101400002, the National Natural Science Foundation of China under Grants 62171340, 61991451 and 61771473, the OPPO Research Fund, the Key Project of Shaanxi Provincial Department of Education (Collaborative Innovation Center) under Grant 20JY024. *(Corresponding authors: Leida Li, Yipo Huang.)*

L. Li, Y. Huang, J. Wu and G. Shi are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mails: ldli@xidian.edu.cn, huangyipo@stu.xidian.edu.cn; jinjian.wu@mail.xidian.edu.cn; gmshi@xidian.edu.cn).

Y. Yang, Y. Li and Y. Guo are with the Intelligent Perception and Interaction Research Department, OPPO Research Institute, Shanghai, China (emails: ippllewis@gmail.com; liyaqian@oppo.com; yandong.guo@live.com).

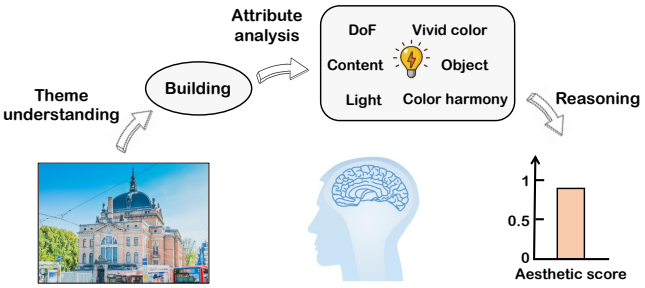


Fig. 1. An illustration of our motivation. When judging image aesthetics, people first understand the image theme, and then analyze the visual attributes according to the perceived theme. Finally, aesthetic judgment is made through reasoning.

In the literature, a number of IAA models have been reported [3]. Early efforts mainly focus on mapping the hand-crafted visual features to a high or low aesthetic category [11]. Although hand-crafted features have explicit physical meanings, they cannot comprehensively model people's aesthetic perception, which are highly complex and abstract. Recently, deep convolutional neural networks (CNN) have demonstrated their advantage in IAA [12], [13]. Despite the notable advances achieved, the existing deep IAA models largely work in a pure data-driven manner, which is different from the mechanism of human aesthetic perception.

Generally, people's judgment on image aesthetics comes from the perception of various visual attributes [14]. As illustrated in Fig. 1, during image aesthetics rating, people first understand the image theme, based on which they proceed to analyze the visual attributes, e.g., interesting content, good lighting, vivid color and depth of field, etc. Finally, the aesthetic judgment is made through a complex reasoning. This characteristic has been confirmed in related research on photography [15] and aesthetics [12], [16]. However, most of the existing deep IAA models usually map the image into a latent aesthetic feature space directly, which is not consistent with the human perception of image aesthetics [3]. In addition, most of the models can only generate a single scalar aesthetic score or aesthetic distribution, and the lack of explainability might hamper their applications in real-world scenarios.

Motivated by the above facts, this paper presents a Theme-Aware Visual Attribute Reasoning (TAVAR) model for image aesthetics assessment, with the objective to model the aesthetic perception process of human based on a bilevel reasoning framework. Specifically, considering that people's judgment on image aesthetics depends on the perception of visual attributes, which are further coupled with image theme, a

Visual Attribute Analysis Network (VAAN) and a Theme Understanding Network (TUN) are first pre-trained to extract aesthetic attribute features and theme features, respectively. Then, the first level Attribute-Theme Graph (ATG) is built to investigate the coupling relationship between visual attributes and image theme, producing the theme-aware visual attribute features. Further, a flexible aesthetics network is introduced to extract the general aesthetic features, based on which we built the second level Attribute-Aesthetics Graph (AAG) to mine the relationship between theme-aware visual attributes and aesthetic features, producing the final aesthetic prediction. Extensive experimental results demonstrate that the proposed TAVAR model not only accurately assesses the image aesthetic quality, but also predicts the visual attributes simultaneously to facilitate model explainability.

The contributions of this work can be summarized with the following points:

- We propose a new IAA model based on theme-aware visual attribute reasoning, dubbed TAVAR, which simulates the aesthetic perception process of human and delivers the state-of-the-art performance. In contrast to the existing IAA models that only predict a single quality score or aesthetic distribution, the proposed model can also output the aesthetic attributes, which in turn facilitate better model explainability.
- We propose a bilevel aesthetic reasoning framework based on the attribute-theme graph (ATG) and the attribute-aesthetics graph (AAG). ATG is designed to investigate the interaction between image theme and visual attributes, based on which the theme-aware visual attribute features are further combined with the general aesthetic features to perform the second-level reasoning, producing the final aesthetic score.
- We conduct extensive experiments and comparisons on four public IAA databases, and the experimental results demonstrate the superiority of the proposed TAVAR model over the state-of-the-arts. Visual analysis is also provided to demonstrate the explainability of the proposed model.

The remainder of this paper is structured as follows. In Section II, we review the related works on IAA and graph reasoning networks. Section III describes the details of the proposed model. Experimental results and analysis are presented in Section IV. Finally, we conclude this paper in Section V.

II. Related Works

In this section, we briefly review the literature related to image aesthetics assessment and graph reasoning networks that are closely related to our work.

A. Image Aesthetics Assessment

Early works on IAA are mainly based on hand-crafted features. For example, Ke *et al.* [11] proposed an IAA model to distinguish high-quality and low-quality images by designing a set of perceptual features, including spatial distribution of edges, color distribution, and hue count, etc. A Bayes classifier was used to integrate the features and achieve the aesthetic decision. In [17], the authors focused on the foreground subject and developed a set of high-level features for describing photo aesthetic quality. Tang *et al.* [18] proposed a content-based IAA model by first separating the subject from the background. Then, aesthetic features were extracted from both the subject region and the background region to compute the image aesthetic score. In addition to aesthetic-related features, generic features were also used to measure image aesthetic quality, e.g. the global image descriptor (GIST) [19] and scale-invariant feature transform (SIFT) [20]. While these hand-crafted features have explicit physical meanings, they are typically built based on the still limited understanding of aesthetics and cannot describe image aesthetics comprehensively.

With the considerable progress in deep learning, a variety of CNN models have been developed and become the de facto configuration in building modern IAA models. Lu *et al.* [21] designed a double-column CNN model to learn aesthetic features from both the global and local views. The style features of images were also utilized to improve the prediction accuracy. In [22], Jin *et al.* proposed a deep IAA model based on GoogLeNet named ILGNet, which integrated both the inception modules and a connected layer of local and global features. In [16], the authors proposed a new CNN architecture by joint learning of meaningful photographic features and image scene information. Chen *et al.* [23] proposed an IAA model based on the adaptive fractional dilated convolution (AFDC), which can explicitly relate the perception of image aesthetics to the aspect ratios while preserving the composition. Li *et al.* [24] proposed a personality-assisted multi-task deep IAA model, which can improve the aesthetic representation ability by jointly learning personality features. In [25], Shu *et al.* proposed a deep IAA model called PI-DCNN, which utilized the prior knowledge of photo and photographic elements as privileged information and transferred the privileged information to formulate image aesthetics assessment. More recently, She *et al.* [26] proposed an end-to-end graph-based representation learning framework for image aesthetics assessment, called HLA-GCN, which utilized two LA-GCN modules to capture layout information. While notable advances have been achieved, these methods do not explicitly model the process of human perception of aesthetics, which is typically characterized by an interaction between image theme and visual attributes in determining the overall aesthetic quality.

B. Graph Reasoning Networks

With the capability of relationship reasoning, graph-based networks have been applied to a variety of high-level vision tasks [27]. Early works mainly focus on modeling graph data using simple discriminative models, e.g. Conditional Random Fields (CRFs) [28] and random walk networks [29]. Recently, Graph Convolution Network (GCN) [30] has been proposed by generalizing CNN to graph data based on two core operations, *i.e.*, aggregating and transforming node embeddings, which has demonstrated notable advantage in modeling complex relationships [31]. For example, Kipf *et al.* [30] proposed a GCN model by introducing a localized and well-behaved propagation rule for semi-supervised learning on graph-structured

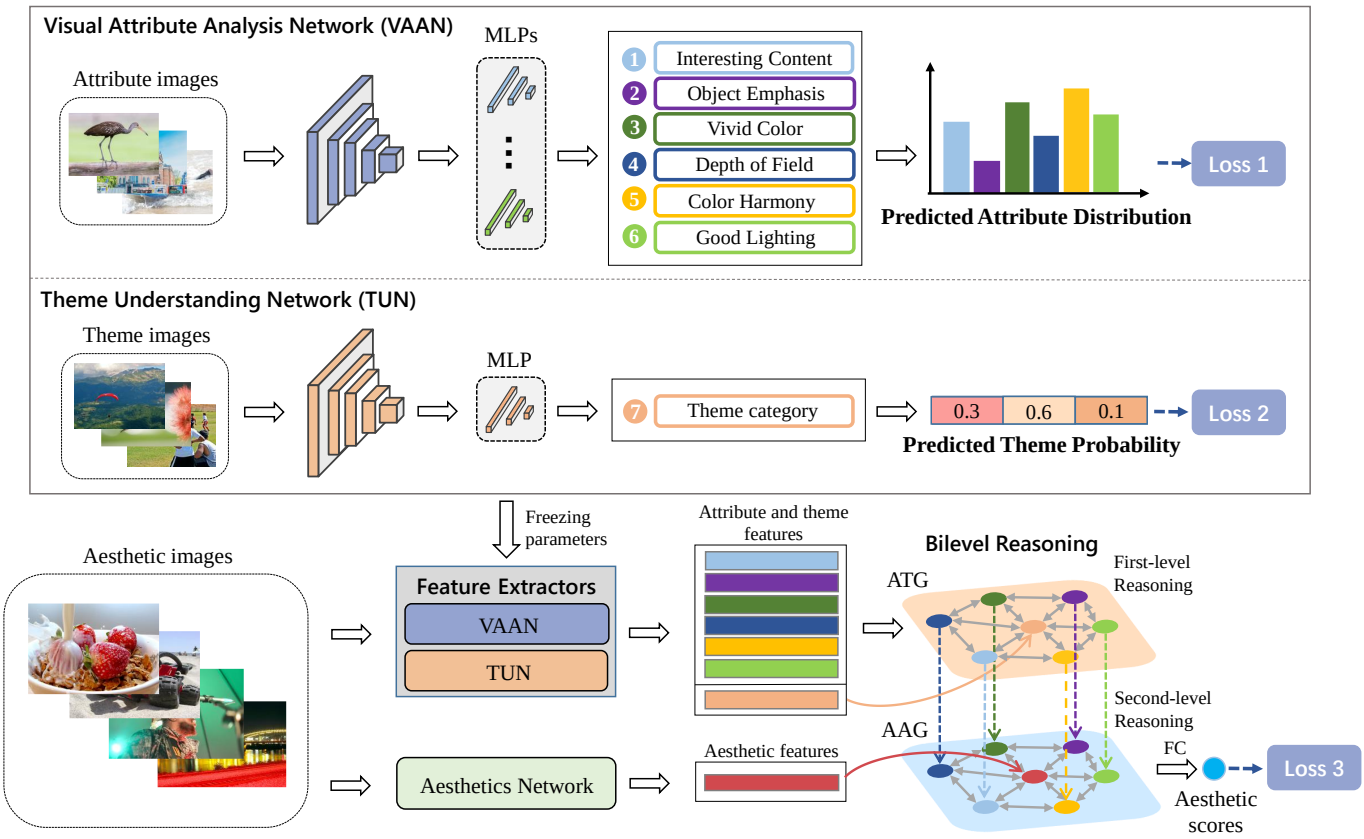


Fig. 2. The overall structure of the proposed Theme-Aware Visual Attribute Reasoning (TAVAR) model for image aesthetics assessment. 1) Visual Attribute Analysis Network (VAAN) is pre-trained to extract visual attribute features. 2) Theme Understanding Network (TUN) is pre-trained to extract image theme features. 3) Aesthetic network is used to extract the general aesthetic features. 4) Attribute-Theme Graph (ATG) is used to mine the relationship between image theme and visual attributes. 5) Attribute-aesthetic Graph (AAG) is used to further mine the relationship between the theme-aware visual attributes and general aesthetic features.

data. Yang *et al.* [32] used GCN to investigate the contextual information between objects and relations for scene graph generation. Nie *et al.* [33] utilized GCN to learn the utterance features for emotion detection in conversation. In this paper, we leverage GCN to build a bilevel reasoning framework, aiming to model the aesthetic perception process of human.

III. Proposed Model

The overall structure of the proposed TAVAR model is illustrated in Fig. 2. Specifically, inspired by human evaluation process of image aesthetics, we first train two feature extractors to obtain visual attribute and theme features based on the Visual Attribute Analysis Network (VAAN) and the Theme Understanding Network (TUN). Then, the first level Attribute-Theme Graph (ATG) is designed to mine the coupling relationship between visual attributes and image theme in determining aesthetic perception. Further, a flexible aesthetic network was introduced to extract the general aesthetic features, based on which we develop the second level Attribute-Aesthetics Graph (AAG) to further explore the relationship between theme-aware attribute features and aesthetic features, producing the final aesthetic score.

A. Feature Extractors

Visual Attribute Analysis Network (VAAN). Study in [16] has proven that people judge image aesthetic quality mainly based on visual attributes. Therefore, visual attributes strongly correlate with image aesthetics, which can be regard as the aesthetic elements to describe an image intuitively, *e.g.* interesting content, good lighting or vivid color. In this part, we design a multi-branch CNN model to learn the aesthetics-aware visual attributes, which is illustrated in the upper part of Fig. 2. Specifically, we build the network using ResNet-50 [34] by removing the fully connected layers, which shares the feature extraction among branches. Then, we employ six multi-layer perceptrons (MLPs) to further map the shared features to six visual attributes, including interesting content, object emphasis, vivid color, depth of field, color harmony, and good lighting [35]. Specifically, for an input image x, the hidden features $\boldsymbol{h}_a$ are first obtained from the shared feature extraction network $\mathbb{F}_{\theta_a}$ as:

$$\boldsymbol{h}_a = \mathbb{F}_{\theta_a}(x), \tag{1}$$

where θ_a denotes the parameter set of the shared feature extraction network $\mathbb{F}_{\theta_a}$.

Then, six MLPs are used to build the six attribute branches. Since the Parametric Rectified Linear Unit (PReLU) [36] can

improve the performance of deep model and reduce the risk of overfitting compared with ReLU, we introduce PReLU as activation function in MLPs. Next, we leverage the six attribute branches to further map the hidden features $\boldsymbol{h}_a$ to the visual attributes $\hat{\boldsymbol{a}}$, which is defined as:

$$\hat{\boldsymbol{a}} = MLP_{\theta_m}(\boldsymbol{h}_a), \tag{2}$$

where $\theta_m, (m = 1, 2, ..., 6)$ denotes the parameters of each attribute branch MLP_{θ_m}, and $\hat{\boldsymbol{a}} = \{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_6\}$ denotes six predicted visual attributes.

During the pre-training of VAAN, we assume that a visual attribute dataset $\mathcal{D}_a = \{x_i, \boldsymbol{a}_i\}_{i=1}^{N_a}$ can provide images and the corresponding visual attribute labels, where $\boldsymbol{a}_i$ denotes the labeled visual attributes of image x_i $(i = 1, 2, \ldots, N_a)$, and N_a represents the number of samples in $\mathcal{D}_a$. In this work, the training goal of VAAN is to predict the values of visual attributes as a regression task. Considering that L1 loss has a stronger robustness and can avoid the problem of gradient explosion in the regression task [37], L1 loss is leveraged to optimize the parameters θ_a and θ_m based on the $\mathcal{D}_a$ dataset, which is defined as:

$$\mathcal{L}_1 = \frac{1}{N_a} \sum_{i=1}^{N_a} \mid \boldsymbol{a}_i - \hat{\boldsymbol{a}}_i \mid, \tag{3}$$

where $\hat{\boldsymbol{a}}_i$ denote the predicted visual attributes of image x_i, which is computed by:

$$\hat{\boldsymbol{a}}_i = MLP_{\theta_m}(\mathbb{F}_{\theta_a}(x_i)). \tag{4}$$

The visual attribute analysis network can be built by training on $\mathcal{D}_a$ and can simultaneously extract the features of all visual attributes, which will be used as inputs of the subsequent graph reasoning module.

Theme Understanding Network (TUN). To understand the aesthetics of an image, people first pay attention to the theme because image aesthetics is strongly related to the theme of the image [15]. In other words, it has been acknowledged that people would first figure out what they see in an image before they make aesthetic judgment [12]. Further, image theme and aesthetic attributes are always tightly coupled, so people's understanding of image theme will affect the judgment of visual attributes [16]. As a result, when using visual attributes for building the IAA model, the theme category of an image should be simultaneously considered to achieve comprehensive prediction. With this consideration, a theme understanding network is also integrated in this work. Specifically, another ResNet-50 backbone [34] is first employed to build the theme understanding network. Then, a multi-layer perceptron (MLP) with PReLU activation function is utilized to map the input image x_i to the predicted theme categories, where the last fully connected layer produces six outputs, which represent six theme categories. Finally, a softmax nonlinearity operation is performed to generate the predicted theme probablities, which is formulated as:

	The.	IC	OE	VC	DoF	CH	GL
The.	1	1	1	1	1	1	1
IC	1	1	0	0	1	1	0
OE	1	0	1	1	0	0	1
VC	1	0	1	1	1	0	0
DoF	1	1	0	1	1	0	0
CH	1	1	0	0	0	1	1
GL	1	0	1	0	0	1	1

(a)

	Aes.	IC	OE	VC	DoF	CH	GL
Aes.	1	1	1	1	1	1	1
IC	1	1	0	0	1	1	0
OE	1	0	1	1	0	0	1
VC	1	0	1	1	1	0	0
DoF	1	1	0	1	1	0	0
CH	1	1	0	0	0	1	1
GL	1	0	1	0	0	1	1

(b)

Fig. 3. Illustration of the adjacency matrices $\boldsymbol{A}^{ac}$ and $\boldsymbol{A}^{aa}$. (a): the adjacency matric $\boldsymbol{A}^{ac}$; (b): the adjacency matric $\boldsymbol{A}^{aa}$; IC: Interesting Content; OE: Object Emphasis; VC: Vivid Color; DoF: Depth of Field; CH: Color Harmony; GL: Good Lighting; The.: Theme; Aes.: Aesthetic.

$$\hat{\boldsymbol{c}} = MLP_{\theta_k}(\mathbb{F}_{\theta_c}(x_i)), \tag{5}$$

where θ_c and θ_k denote the parameters of the theme understanding network $\mathbb{F}_{\theta_c}$ and the multi-layer perceptron MLP_{θ_k} respectively, and $\hat{\boldsymbol{c}}$ denotes the predicted theme probablities.

To train the theme understanding network, we adopt a dataset $\mathcal{D}_c = \{x_i, \boldsymbol{c}_i\}_{i=1}^{N_c}$ with images and the corresponding theme category labels, where $\boldsymbol{c}_i$ denotes the labeled theme category of image x_i $(i = 1, 2, 3, \ldots, N_c)$, and N_c represents the number of images in $\mathcal{D}_c$. In this work, the proposed Them Understanding Network (TUN) is used to predict the theme category, which is trained as a classification task. Therefore, we introduce the popular cross-entropy loss to optimize the parameters θ_c and θ_k, which is defined as:

$$\mathcal{L}_2 = -\sum_{i=1}^{N_c} \mathbf{c}_i \log\ (\hat{\mathbf{c}}_i), \tag{6}$$

where $\hat{\boldsymbol{c}}_i$ denotes the predicted theme category of image x_i, which is computed by:

$$\hat{\boldsymbol{c}}_i = MLP_{\theta_k}(\mathbb{F}_{\theta_c}(x_i). \tag{7}$$

The theme understanding network is trained on $\mathcal{D}_c$, which can not only predict the theme category of the image but also generate the theme features, which will be also input into the subsequent graph reasoning module. In our design, although both the VAAN and TUN take advantage of ResNet-50 as the backbone network to extract feature, there are two major differences between them. First, VAAN contains a multi-branch regressor for predicting six visual attributes, while TUN only uses an MLP to achieve theme category prediction. Therefore, VAAN and TUN have different prediction heads. Second, the training objectives of them are different. VAAN is trained to predict the value of aesthetic attributes, which belongs to the regression task. TUN is built to predict the category probabilities of theme, which belongs to the classification task. As a result, VAAN and TUN use different loss functions to optimize the parameters of model.

B. Bilevel Aesthetic Reasoning

Attribute-Theme Graph (ATG). As mentioned above, people usually assess image aesthetics according to visual attributes. Further, the perception of visual attributes depends on the image theme. Therefore, we first build an attribute-theme graph to investigate the relationship between image theme and visual attribute, which is formulated as:

$$\mathcal{G}^{ac} = (\boldsymbol{H}^{ac}, \boldsymbol{A}^{ac}), \tag{8}$$

where $\boldsymbol{H}^{ac} = \{\boldsymbol{H}^{a}, \boldsymbol{H}^{c}\}$, $\boldsymbol{H}^{a}$ and $\boldsymbol{H}^{c}$ denote the node features of visual attributes and image theme extracted form the aforementioned feature extractors, and $\boldsymbol{A}^{ac}$ is the theme-centric adjacency matrix, which is built based on the pre-defined manner and can be calculated as illustrated in Fig. 3(a). The motivation to build $\boldsymbol{A}^{ac}$ is two-fold. First, considering the relationship between image theme and visual attributes, we take the theme feature as the central node, and all the attribute nodes are connected to the central theme node. The underlying reason is that the perception of visual attributes depends on the image theme. Second, considering the relationship between different visual attributes, we connect interesting content and DoF, interesting content and color harmony, object emphasis and vivid color, object emphasis and good lighting, vivid color and DoF, color harmony and good lighting, respectively. Finally, the attribute-theme GCN is built as:

$$\boldsymbol{H}^{ac*} = \text{GCN}_{\theta_{ac}}(\mathcal{G}^{ac}), \tag{9}$$

where $\text{GCN}_{\theta_{ac}}$ denotes a graph convolution network, θ_{ac} denotes the learnable parameters in $\text{GCN}_{\theta_{ac}}$. In addition, $\boldsymbol{H}^{ac*} = \{\boldsymbol{H}^{a*}, \boldsymbol{H}^{c*}\}$, where $\boldsymbol{H}^{a*}$ and $\boldsymbol{H}^{c*}$ denote the updated node feature of visual attributes and image theme, respectively.

In implementation, $\text{GCN}_{\theta_{ac}}$ can be described as:

$$\boldsymbol{H}^{ac*} = \text{ReLU}(\widehat{\boldsymbol{A}^{ac}}\boldsymbol{H}^{ac}\boldsymbol{W}^{ac}), \tag{10}$$

where $\boldsymbol{W}^{ac}$ denotes the transformation matrix, and $\widehat{\boldsymbol{A}^{ac}}$ is the normalized version of the adjacency matrix $\boldsymbol{A}^{ac}$. Following the above operations, we can build a GCN to obtain the attribute-theme graph.

Attribute-Aesthetics Graph (AAG). From the above theme-attribute GCN, we can obtain the theme-aware visual attribute features $\boldsymbol{H}^{a*}$. Then, taking into account the complementary role of visual attribute features and general aesthetic features in determining the overall image aesthetic quality [16], we propose the second-level attribute-aesthetics graph reasoning. To this end, we first utilize an aesthetics network to extract the general aesthetic features. Then, we construct the AAG to mine the relationship between theme-aware visual attribute features and aesthetic features, aiming to achieve more comprehensive feature representation ability of aesthetics. In this work, the recently proposed Swin Transformer [38] is utilized as the aesthetics network $\mathbb{F}_{\theta_p}$, which can be formulated as:

$$\boldsymbol{H}_p = \mathbb{F}_{\theta_p}(x), \tag{11}$$

Algorithm 1 The proposed TAVAR model.

Input: IAA training set $\mathcal{D}$, which consists of three subsets including $\mathcal{D}_a = \{x_i, \boldsymbol{a}_i\}_{i=1}^{N_a}, \mathcal{D}_c = \{x_i, \boldsymbol{c}_i\}_{i=1}^{N_c}, \mathcal{D}_{IAA} = \{x_i, \boldsymbol{y}_i\}_{i=1}^{N_z}$.
Output: Predicted aesthetic score $\hat{\boldsymbol{y}}$;

```
Initialize all the parameters of the proposed model;
// Feature Extractor Pre-training ;
For iteration = 1, 2, ..., do;
    Sample a batch of k images from D_a;
    For j = 1, 2, ..., N do;
        Output {â}_{i=1}^k by using F_{θ_a} and MLP_{θ_m};
        Update θ_a and θ_m by computing L_1;
    end For
end For
For iteration = 1, 2, ..., do;
    Sample a batch of k images from D_c;
    For j = 1, 2, ..., N do;
        Output {ĉ}_{i=1}^k by using F_{θ_c} and MLP_{θ_k};
        Update θ_c and θ_k by computing L_2;
    end For
end For
Freezing the parameters of feature extractors;
// Bilevel Aesthetic Reasoning ;
Building the adjacency matrixes A^{ac} and A^{aa};
For iteration = 1, 2, ..., do;
    Sample a batch of k images from D_{IAA};
    For j = 1, 2, ..., N do;
        Output aesthetic score {ŷ}_{i=1}^k by using GCN_{θ_ac},
        F_{θ_p}, GCN_{θ_aa} and FC_{θ_l};
        Update the parameters of θ_ac, θ_p, θ_aa and θ_l by
        computing L_3;
    end For
end For
```

where θ_p denotes the parameter set of the network $\mathbb{F}_{\theta_p}$, and $\boldsymbol{H}_p$ denotes the extracted aesthetic features.

Then, the attribute-aesthetics graph is formulated as:

$$\mathcal{G}^{aa} = (\boldsymbol{H}^{aa}, \boldsymbol{A}^{aa}), \tag{12}$$

where $\boldsymbol{A}^{aa}$ is the aesthetics-centric adjacency matrix of $\mathcal{G}^{aa}$, which is illustrated in Fig. 3(b). Like $\boldsymbol{A}^{ac}$, the motivation to build the adjacency matrix $\boldsymbol{A}^{aa}$ includes two aspects. First, considering the relationship between different visual attributes, we connect different theme-aware visual attributes in the same way as $\boldsymbol{A}^{ac}$. Moreover, based on the relationship between theme-aware visual attributes and general aesthetics, we regard the features of general aesthetics as the central node, and all the theme-aware visual attribute nodes are connected to it. Based on the attribute-aesthetics graph, we can obtain the updated node features that integrate theme-aware visual attribute and aesthetic features,

$$\boldsymbol{H}^{aa*} = \text{GCN}_{\theta_{aa}}(\mathcal{G}^{aa}), \tag{13}$$

where θ_{aa} denotes the learnable parameters in $\text{GCN}_{\theta_{aa}}$, and $\text{GCN}_{\theta_{aa}}$ can be described as:

$$\boldsymbol{H}^{aa*} = \text{ReLU}(\widehat{\boldsymbol{A}^{aa}}\boldsymbol{H}^{aa}\boldsymbol{W}^{aa}), \tag{14}$$

where $\boldsymbol{W}^{aa}$ denotes the transformation matrix, and $\widehat{\boldsymbol{A}^{aa}}$ is the normalized version of the adjacency matrix $\boldsymbol{A}^{aa}$. Based on the above operations, the attribute-aesthetics graph can be built.

Finally, we append a FC layer to map the updated node features H^{aa*} to the overall aesthetic quality score $\hat{\boldsymbol{y}}$, which is defined as:

$$\hat{\boldsymbol{y}} = \text{FC}_{\theta_l}(H^{aa*}), \tag{15}$$

where θ_l represents the parameters of the FC layer.

During the training of the whole TAVAR model, we use $\mathcal{D}_{IAA} = \{x_i, \boldsymbol{y}_i\}_{i=1}^{N_z}$ to denote the image aesthetics assessment dataset, where $\boldsymbol{y}_i$ denotes the ground truth aesthetic score of image x_i $(i = 1, 2, \ldots, N_z)$, and N_z represents the number of samples in $\mathcal{D}_{IAA}$. Based on $\mathcal{D}_{IAA}$, we employ L1 loss to optimize the parameters of the whole model, which is defined as:

$$\mathcal{L}_3 = \frac{1}{N_z}\sum_{i=1}^{N_z} \mid \boldsymbol{y}_i - \hat{\boldsymbol{y}}_i \mid . \tag{16}$$

The training process of the proposed TAVAR model is summarized in Algorithm 1.

IV. Experimental Results

A. Databases

To verify the performance of the proposed TAVAR model, we conduct experiments on four popular IAA databases, including AADB [16], EVA [39], AVA [40] and PARA [41].

AADB database [16]. This database contains 10,000 images rated by a total of 190 users. Each image is annotated with the overall aesthetic score and 11 visual attributes by at least 5 users. These attributes include light, content, object, color harmony, vivid color, depth of field (DoF), motion blur, rule of thirds, balancing element, repetition, and symmetry. The visual attributes and overall aesthetic scores range from [-1, 1] and [1, 5] respectively. In our experiment, six of the above eleven visual attributes are used to pre-train the visual attribute analysis branch. During aesthetic quality prediction, following the widely adopted setting [13], [35], [42], 8,500 images are used for model training, 500 images are used for validation, and the rest 1,000 images are used for testing.

EVA database [39]. This database includes 4,070 images, where each image is annotated by 30 to 40 users. In addition to the overall aesthetic score, it also provides labels for 6 common theme types, including human, animals, natural and rural scenes, architectures and city scenes, still life, and others. In this work, we use the EVA database to pre-train the theme understanding branch. During performance evaluation, 3,500 images are used for training, and the rest 570 images are used for testing.

AVA database [40]. This database contains more than 250,000 images collected from the website of *DPChallenge*, where each image is annotated by 210 users on average. The overall aesthetic scores range from [1, 10]. Following the common setting in IAA [26], we utilize a total of 235, 503 images for training, and 19, 997 images for testing. For the aesthetic binary classification, images with overall aesthetic scores above 5 are categorized into high aesthetic quality, and the other images are categorized into low aesthetic quality.

PARA database [41]. This database contains a total of 31,220 images and each image is annotated by 25 subjects in average and 438 subjects in total. Each image is annotated with 4 human-oriented subjective attributes and 9 image-oriented objective attributes. In addition, PARA also provides labels for 10 theme types, including portrait, animal, plant, scene, building, still life, night scene, food, indoor, and others. The overall aesthetic scores range from [1, 5]. Following the standard setting [41], 28,220 images are used for training, and the rest 3000 images are used to test the performance of models.

B. Implementation Details

For all the experiments, we first resize images into $244 \times 244 \times 3$, and then randomly crop into $224 \times 224 \times 3$ for input. In the test stage, we directly resize original images into $224 \times 224 \times 3$. In implementation, we first employ the AADB database [16] and EVA database [39] to pre-train the attribute and theme feature extractors. Then, we freeze the parameters of feature extractors, and target databases are used to train the bilevel reasoning model and fine-tune the aesthetic network simultaneously. Specifically, the stochastic gradient descent (SGD) is used as the optimizer, and the initial learning rate is 0.03 with a warm-up strategy. We utilize Pytorch to implement the proposed model and train it on a computer with Intel Core i7-9700K CPU @ 3.60GHz, and NVIDIA GeForce RTX 3090 24G GPU.

For comparison with the existing IAA models, Pearson linear correlation coefficient (PLCC) is used to evaluate the accuracy of the prediction results, and Spearman rank order correlation coefficient (SRCC) is used to measure the prediction monotonicity [43], [44]. Before computing PLCC, the predicted scores need to passe through a five-parameter nonlinear mapping,

$$\hat{z} = \xi_1\left(0.5 - \frac{1}{1 + e^{\xi_2(z-\xi_3)}}\right) + \xi_4 z + \xi_5, \tag{17}$$

where z donates the prediction score, $\hat{z}$ denotes the mapped score, and ξ_i, i=1, 2, …, 5, are the fitting parameters. Then, the PLCC is computed by,

$$\text{PLCC} = \frac{\sum_{i=1}^{N}(s_i - \mu_{s_i})(\hat{z}_i - \mu_{\hat{z}_i})}{\sqrt{\sum_{i=1}^{N}(s_i - \mu_{s_i}) * \sum_{i=1}^{N}(\hat{z}_i - \mu_{\hat{z}_i})}}, \tag{18}$$

where N is the number of the testing samples, s_i and $\hat{z}_i$ represent the ground truth score and the mapped score of i-th test image respectively.

The SRCC is computed by,

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{19}$$

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED TAVAR MODEL WITH THE STATE-OF-THE-ART IAA MODELS ON THE AADB [16] DATABASE.

Method	SRCC	PLCC	ACC (%)
RegNet(AlexNet) [16]	0.678	-	-
Hou *et al.*(VGG16) [47]	0.689	-	-
Malu *et al.*(ResNet-50) [35]	0.689	-	-
PI-DCNN(ResNet-152) [25]	0.705	-	-
RGNet(ResNet-101) [48]	0.710	-	-
Unified_IAA(ResNet-101) [42]	0.726	-	-
NIMA(ResNet-50) [13]	0.708*	0.711*	80.1*
MLSP(Inception-v2) [49]	0.725*	0.726*	78.2*
PA_IAA(DenseNet-121) [24]	0.720*	0.728*	70.9*
TANet(MobileNet-v2) [12]	0.738*	0.737*	79.8*
MUSIQ(ViT) [50]	0.706*	0.712*	76.3*
Celona *et al.*(EfficientNet) [51]	0.757	0.762	81.6
TAVAR (Proposed)	**0.761**	**0.763**	**81.9**

where d_i represents the difference between the ranks of the ground truth and predicted scores. A good IAA model is expected to deliver higher PLCC and SRCC values [45], [46].

Moreover, the overall accuracy (ACC) is employed to evaluate the performance of aesthetic binary classification on AVA database, which is computed by,

$$\text{ACC} = \frac{TP + TN}{P + N}, \tag{20}$$

where P and N denote the number of high and low aesthetic images, respectively. TP and TN are the number of correctly classified images. The ACC is in the range [0, 1], and higher ACC value represents better classification performance.

C. Performance Evaluation

Performance on AADB Database. We first compare the performance of the proposed TAVAR model with the relevant state-of-the-arts on the AADB database. In experiments, we evaluate the performance of these models on two aesthetic tasks, *i.e.*, aesthetic binary classification and aesthetic score regression. The experimental results are summarized in Table I, where the best results are marked in boldface and the results with * are obtained from our experiments. Since several previous IAA models only test the SRCC values and the source codes of those works have not been released, their PLCC and ACC results are marked by "-". It is known from the experimental results that the proposed TAVAR model outperforms other comparison models in terms of SRCC, PLCC and ACC values. This demonstrates that our TAVAR using theme-aware visual attribute reasoning is very effective for the IAA task.

Performance on EVA and PARA Databases. EVA [39] and PARA [41] databases are two recently released new IAA databases, and most of the existing models did not report experimental results on them. For comparison, we retrain five popular IAA models with public codes including NIMA [13], MLSP [49], PA_IAA [24], MUSIQ [50] and TANet [12]. All experiments are conducted under the same setting as the proposed TAVAR. The experimental results are summarized in Table II, where the best results are marked in boldface. It is known from the experimental results that the proposed TAVAR model achieves the best performances on both EVA [39] and PARA [41] databases. On the EVA database, TAVAR achieves the best PLCC, SRCC and ACC of 0.799, 0.810 and 89.6%,

TABLE II
PERFORMANCE COMPARISON OF THE PROPOSED TAVAR MODEL WITH THE STATE-OF-THE-ART IAA MODELS ON THE EVA [39] AND PARA [41] DATABASES.

Method	SRCC	PLCC	ACC (%)
EVA database			
NIMA(ResNet-50) [13]	0.725*	0.738*	70.4*
MLSP(Inception-v2) [49]	0.677*	0.684*	85.5*
PA_IAA(DenseNet-121) [24]	0.742*	0.761*	72.4*
MUSIQ(ViT) [50]	0.715*	0.747*	88.3*
TANet(MobileNet-v2) [12]	0.794*	0.769*	88.5*
TAVAR	**0.799**	**0.810**	**89.6**
PARA database			
NIMA(ResNet-50) [13]	0.886*	0.923*	89.0*
MLSP(Inception-v2) [49]	0.842*	0.892*	84.2*
PA_IAA(DenseNet-121) [24]	0.877*	0.919*	87.5*
MUSIQ(ViT) [50]	0.882*	0.918*	88.1*
TANet(MobileNet-v2) [12]	0.883*	0.917*	89.2*
TAVAR	**0.911**	**0.940**	**89.7**

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED TAVAR MODEL WITH THE STATE-OF-THE-ART IAA MODELS ON THE AVA [40] DATABASE.

Method	SRCC	PLCC	ACC (%)
A-Lamp(VGG16) [52]	-	-	82.5
ILGNet(GoogLeNet) [22]	-	-	82.7
RegNet(AlexNet) [16]	0.558	-	77.3
USAR(AlexNet) [53]	0.578	-	78.1
PA_IAA(DenseNet-121) [24]	0.666	-	82.9
PA_IAA(Inception-v3) [24]	0.677	-	83.7
GPF-CNN(InceptionNet) [54]	0.690	0.704	81.8
NIMA(VGG16) [13]	0.592	0.610	80.6
NIMA(Inception-v2) [13]	0.612	0.636	81.5
NIMA(ResNet-50) [13]	0.690	0.694	79.3
AFDC(ResNet-50) [23]	0.649	0.671	83.0
Unified_IAA(ResNet-101) [42]	0.719	0.720	80.8
HLA-GCN(ResNet-50) [26]	0.665	0.687	84.6
TAVAR	**0.725**	**0.736**	**85.1**

TABLE IV
PLCC/SRCC VALUES OF BACKBONE NETWORKS AND TAVAR IN LEAVE-ONE-THEME-OUT CROSS VALIDATION ON THE EVA [39] AND PARA [41] DATABASES. ANI.: ANIMALS, A.&C.: ARCHITECTURES AND CITY SCENES, NAT.: NATURAL AND RURAL SCENES, STI: STILL LIFE, OTH.: OTHERS, BUI.:BUILDING, IND.: INDOOR, NIG.: NIGHT SCENE, POR.: PORTRAITURE, AVE.: AVERAGE, MOB.-V3: MOBILENET-V3, S-TRANS.: SWIN TRANSFORMER.

Database	Theme	Mob.-v3	TAVAR	Gain (%)	ResNet-50	TAVAR	Gain (%)	S-Trans.	TAVAR	Gain (%)
EVA	Ani.	0.395/0.383	0.465/0.455	7.0/7.2	0.444/0.439	0.508/0.490	6.4/5.1	0.478/0.453	0.572/0.566	9.4/11.3
	A.&C.	0.622/0.614	0.667/0.676	4.5/6.2	0.639/0.631	0.703/0.705	6.4/7.4	0.656/0.655	0.725/0.731	6.9/7.6
	Human	0.445/0.425	0.495/0.490	5.0/6.5	0.461/0.450	0.521/0.501	6.0/5.1	0.532/0.495	0.561/0.534	2.9/3.9
	Nat.	0.669/0.670	0.712/0.717	4.3/4.7	0.676/0.687	0.740/0.743	6.4/5.6	0.667/0.662	0.739/0.746	7.2/8.4
	Sti.	0.483/0.479	0.576/0.577	9.3/9.8	0.538/0.512	0.580/0.561	4.2/4.9	0.538/0.516	0.603/0.576	6.5/6.0
	Oth.	0.704/0.683	0.720/0.702	1.6/1.9	0.687/0.682	0.752/0.711	6.5/2.9	0.722/0.704	0.779/0.724	5.7/2.0
	Ave.	0.553/0.542	0.606/0.603	5.3/6.1	0.574/0.567	0.634/0.619	6.0/5.2	0.599/0.581	0.663/0.646	6.4/6.5
PARA	Ani.	0.797/0.765	0.803/0.776	0.6/1.1	0.804/0.786	0.812/0.792	0.8/0.6	0.824/0.802	0.856/0.835	3.2/3.3
	Bui.	0.864/0.769	0.889/0.814	2.5/4.5	0.874/0.807	0.893/0.825	1.9/1.8	0.887/0.811	0.915/0.856	2.8/4.5
	Food	0.790/0.748	0.810/0.774	2.0/2.6	0.787/0.752	0.798/0.764	1.1/1.2	0.836/0.805	0.870/0.839	3.4/3.4
	Ind.	0.843/0.763	0.897/0.834	5.4/7.1	0.882/0.810	0.904/0.838	2.2/2.8	0.895/0.840	0.932/0.890	3.7/5.0
	Nig.	0.866/0.848	0.879/0.860	1.3/1.2	0.884/0.870	0.941/0.936	5.7/6.6	0.892/0.877	0.923/0.913	3.1/3.6
	Oth.	0.907/0.731	0.941/0.802	3.4/7.1	0.937/0.743	0.954/0.821	1.7/7.8	0.943/0.801	0.966/0.849	2.3/4.8
	Plant	0.901/0.786	0.912/0.815	1.1/2.9	0.904/0.810	0.910/0.822	0.6/1.2	0.926/0.837	0.939/0.858	1.3/2.1
	Por.	0.852/0.796	0.861/0.811	0.9/1.5	0.858/0.813	0.868/0.822	1.0/0.9	0.882/0.841	0.910/0.873	2.8/3.2
	Scene	0.913/0.866	0.919/0.880	0.6/1.4	0.916/0.879	0.923/0.882	0.7/0.3	0.932/0.899	0.946/0.920	1.4/2.1
	Sti.	0.918/0.869	0.923/0.877	0.5/0.8	0.923/0.875	0.940/0.899	1.7/2.4	0.931/0.894	0.946/0.915	1.5/2.1
	Ave.	0.865/0.794	0.883/0.824	1.8/3.0	0.877/0.815	0.894/0.840	1.7/2.5	0.895/0.841	0.920/0.875	2.5/3.4

surpassing the second-best model by 0.5% (SRCC), 4.1% (PLCC) and 1.1% (ACC), respectively. On the PARA database, the proposed TAVAR is advantageous over the existing IAA models by a sizable margin in terms of both the aesthetic score regression task (SRCC and PLCC) and the binary classification task (ACC). This further confirms the advantage of the proposed TAVAR model.

Performance on AVA Database. We further evaluate the performance of the proposed TAVAR model on the large-scale AVA database [40] and compare it with the state-of-the-art IAA models. The experimental results are summarized in Table III, where the best results are marked in boldface and the "-" means that the results is not available. From Table III, we can observe that the proposed TAVAR surpasses all the competing IAA models on the AVA database. For the aesthetic score regression task, TAVAR also achieves the best results both on the prediction monotonicity (SRCC) and accuracy (PLCC). For the binary classification task, the proposed TAVAR achieves 85.1% classification accuracy and outperforms the second-best model HLA-GCN [26] by 0.5%. This demonstrates that the proposed TAVAR achieves the best performance on AVA database in terms of both the binary classification task and the aesthetic score regression task.

D. Generalization Ability

Cross Theme Evaluation. The proposed theme-aware TAVAR model is inspired by the mechanism of human perception of aesthetics. To validate the generalization performance of the proposed model for unknown themes, we compare our model with three backbone networks by using the Leave-One-Theme-Out cross validation on EVA and PARA databases, which have theme annotations. In implementation, suppose there are N kinds of themes in a database, we use $(N-1)$ kinds of themes for training and the remaining one theme is used for performance test. The tested backbone networks include Mobilenet-v3 [55], ResNet-50 [34] and Swin Transformer [38], ranging from the lightweight, most commonly used, and most advanced backbones. Specifically, we first directly employ these backbone networks to train and test on the AADB, EVA, AVA and PARA databases, respectively. Then, we embed them into the proposed theme-aware IAA model as the aesthetics network to repeat the experiments respectively. The performances in terms of PLCC/SRCC values are listed in Table IV. To the best of our knowledge, the Leave-One-Theme-Out experiment has not been conducted in existing IAA works, but it is very important for real-world applications of IAA models.

From Table IV, we have the following observations. (1) The proposed TAVAR achieves very promising performance the Leave-One-Theme-Out cross validation on both databases. Especially on the PARA database, TAVAR using any backbone network can achieve PLCC and SRCC values greater than 0.87. This indicates that our TAVAR model can effectively learn the influence of themes on aesthetic perception and

TABLE V
PERFORMANCE COMPARISON OF THE PROPOSED TAVAR MODEL WITH THE STATE-OF-THE-ART IAA MODELS WHEN TRAINING AND TEST ON DIFFERENT DATABASES.

Model	Training	Test			
		AADB	EVA	AVA	PARA
NIMA [13]	AADB	-	0.542	0.272	0.715
	EVA	0.457	-	0.278	0.642
	AVA	0.471	0.531	-	0.626
	PARA	0.685	0.655	0.335	-
	Average	0.538	0.576	0.295	0.661
PA-IAA [24]	AADB	-	0.541	0.285	0.729
	EVA	0.429	-	0.295	0.640
	AVA	0.527	0.552	-	0.655
	PARA	0.670	0.692	0.352	-
	Average	0.542	0.595	0.311	0.675
TANet [12]	AADB	-	0.602	0.468	0.771
	EVA	0.486	-	0.546	0.711
	AVA	0.328	0.512	-	0.471
	PARA	0.674	0.669	0.587	-
	Average	0.496	0.594	0.534	0.651
TAVAR	AADB	-	0.587	0.519	0.763
	EVA	0.489	-	0.570	0.728
	AVA	0.480	0.546	-	0.652
	PARA	0.688	0.723	0.644	-
	Average	**0.552**	**0.619**	**0.578**	**0.714**
	Joint training	0.711	0.746	0.697	0.770

fast adapt to IAA task with unknown theme types. (2) The proposed TAVAR model performs significantly better than these backbone networks on all themes of both databases. Moreover, on the EVA database, the performance gain of TAVAR using three backbone networks is higher than PARA database, which proves that our proposed theme-aware IAA framework has a more significant improvement on small-scale database. (3) On both databases, TAVAR using Swin Transformer can achieve the best performance and the largest performance gain. Overall, the proposed TAVAR model has good generalization performance and can consistently improve the generalization performance of backbone networks in the proposed theme-aware IAA framework.

Cross Database Evaluation. Further, we conduct cross-database evaluation to verify the generalization of the proposed TAVAR model on the AADB [16], EVA [39], AVA [40] and PARA [41] databases. In this experiment, we train TAVAR on one database and test it on other databases without doing any fine-tuning. For comparison, we select several top-performing deep IAA models with open source codes, and perform the experiments under the same setting. In addition, as proven in [58], joint training is a promising solution to enhance the model generalization performance. We introduce joint training into our experiments to explore its effectiveness for the proposed TAVAR model. The comparison results in terms of SRCC are summarized in Table V. From the experimental results, we can observe that the average cross-database results of the proposed TAVAR outperform other IAA models by a large margin on all databases. Especially when the model is trained on PARA database, the cross-database results are already better than the intra-database performance of many existing IAA models. In addition, experimental results demonstrate that joint training can further improve the generalization performance of the proposed TAVAR. These experimental results further demonstrate that the proposed TAVAR can be quickly generalized to other target databases without doing any fine-tuning, which is highly desired in real-world applications.

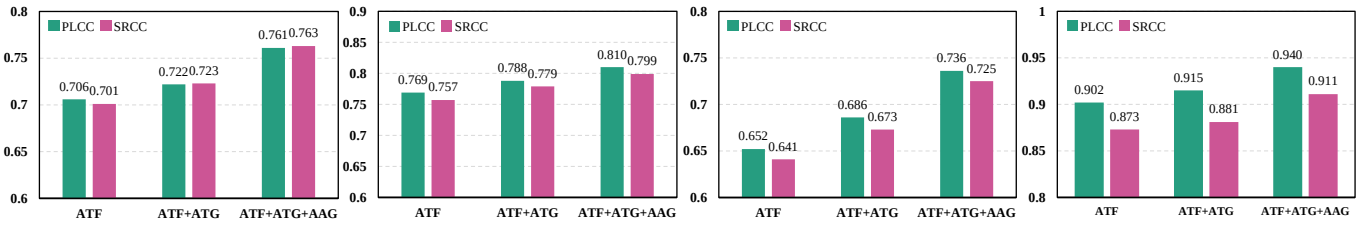
E. Ablation Study

Impact of Aesthetics Network: In this work, the aesthetic network is used to extract general aesthetic features, and the proposed bilevel reasoning framework is able to leverage the theme-aware visual attribute for enhancing the performance of the aesthetic network. To verify the effectiveness of bilevel reasoning framework for different aesthetics networks, we conduct an ablation experiment. Specifically, we first directly adopt six popular backbone networks to train and test on the AADB [16], EVA [39], AVA [40] and PARA [41] databases, respectively. These backbone networks include EfficientNet-b0 [56], Mobilenet-v3 [55], ResNet-50 [34], DenseNet-121 [57], DeiT [59] and Swin Transformer [38], which are all pre-trained on ImageNet. Then, we embed them into the proposed TAVAR to conduct comparison experiments respectively. The comparison results are given in Table VI.

From Table VI, it is observed that TAVAR is superior to these baseline models by a large margin on all databases. Specifically, for EfficientNet-b0 [56], our framework has the highest performance gains of 2.8% (PLCC) and 2.1% (SRCC), where average gains of 1.9% (PLCC) and 1.6% (SRCC) are obtained on all databases. For Mobilenet-v3 [55], TAVAR achieves the biggest performance gains of 2.1% (PLCC) and 2.2% (SRCC), while the average gains are 1.9% (PLCC) and 1.7% (SRCC). For ResNet-50 [34], TAVAR has the highest performance gains of 3.6% (PLCC) and 3.1% (SRCC), where average performance gains of 2.8% (PLCC) and 2.4% (SRCC) are obtained. For DenseNet-121 [57], TAVAR achieves the biggest performance gains of 3.4% (PLCC) and 3.3% (SRCC), while the average gains are 2.4% (PLCC) and 2.3% (SRCC). For DeiT [59], our TAVAR has the highest performance gains of 4.7% (PLCC) and 4.1% (SRCC), where average gains of 3.5% (PLCC) and 2.8% (SRCC) are obtained on all databases. Using Swin Transformer [38], TAVAR achieves the best experimental results while obtaining the highest performance gains of 4.5% (PLCC) and 3.9% (SRCC) and the average performance gains of 3.3% (PLCC) and 3.5% (SRCC). These results demonstrate that the proposed bilevel reasoning framework can effectively adapt to different backbone networks for the IAA task and consistently improve model performance. Even with the lightweight EfficientNet-b0 [56] model, our

TABLE VI
PERFORMANCES OF TAVAR USING VARIOUS BACKBONE NETWORKS ON THE AADB [16], EVA [39], AVA [40] AND PARA [41] DATABASES.

Model	AADB		EVA		AVA		PARA		Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
EfficientNet-b0 [56]	0.706	0.707	0.725	0.719	0.661	0.654	0.906	0.864	0.750	0.736
TAVAR	0.734	0.728	0.743	0.733	0.687	0.673	0.913	0.873	0.769	0.752
Gain	2.8% ↑	2.1% ↑	1.8% ↑	1.4% ↑	2.6% ↑	1.9% ↑	0.7% ↑	0.9% ↑	1.9% ↑	1.6% ↑
Mobilenet-v3 [55]	0.718	0.716	0.750	0.742	0.659	0.652	0.902	0.861	0.757	0.743
TAVAR	0.738	0.733	0.771	0.764	0.679	0.665	0.915	0.876	0.776	0.760
Gain	2.0% ↑	1.7% ↑	2.1% ↑	2.2% ↑	2.0% ↑	1.3% ↑	1.3% ↑	1.5% ↑	1.9% ↑	1.7% ↑
ResNet-50 [34]	0.704	0.703	0.756	0.743	0.667	0.656	0.906	0.870	0.758	0.743
TAVAR	0.736	0.730	0.792	0.771	0.694	0.687	0.920	0.881	0.786	0.767
Gain	3.2% ↑	2.7% ↑	3.6% ↑	2.8% ↑	2.7% ↑	3.1% ↑	1.4% ↑	1.1% ↑	2.8% ↑	2.4% ↑
DenseNet-121 [57]	0.708	0.703	0.746	0.722	0.670	0.655	0.909	0.872	0.758	0.738
TAVAR	0.742	0.736	0.776	0.751	0.689	0.676	0.919	0.879	0.782	0.761
Gain	3.4% ↑	3.3% ↑	3.0% ↑	2.9% ↑	1.9% ↑	2.1% ↑	1.0% ↑	0.7% ↑	2.4% ↑	2.3% ↑
DeiT [56]	0.721	0.723	0.751	0.739	0.661	0.652	0.911	0.876	0.761	0.748
TAVAR	0.754	0.748	0.798	0.780	0.705	0.689	0.926	0.886	0.796	0.776
Gain	3.3% ↑	2.5% ↑	4.7% ↑	4.1% ↑	4.4% ↑	3.7% ↑	1.5% ↑	1.0% ↑	3.5% ↑	2.8% ↑
Swin Transformer [38]	0.731	0.730	0.772	0.763	0.691	0.686	0.920	0.880	0.779	0.765
TAVAR	0.761	0.763	0.810	0.799	0.736	0.725	0.940	0.911	0.812	0.800
Gain	3.0% ↑	3.3% ↑	3.8% ↑	3.6% ↑	4.5% ↑	3.9% ↑	2.0% ↑	3.1% ↑	3.3% ↑	3.5% ↑



(a) AADB (b) EVA (c) AVA (d) PARA

Fig. 4. Ablation analysis on different model components on the AADB [16], EVA [39], AVA [40] and PARA [41] databases.

proposed model still achieves very encouraging performance. This further demonstrates the effectiveness of the proposed theme-aware visual attribute reasoning framework.

Impact of Model Components: To explore the relative contributions of different components of the proposed TAVAR model, ablation studies are further conducted. In the experiment, we first evaluate the effectiveness of visual attribute and theme features extracted from the pre-trained feature extractors (denoted as ATF), where we use a FC layer directly to predict the aesthetic quality score. Then, we add the proposed ATG module to infer the aesthetic quality of the image by mining the relationship between visual attributes and theme category (denoted as ATF+ATG). Finally, we further introduce the proposed AAG module to mine the relationship between theme-aware visual attribute features and general aesthetic features, producing the full TAVAR model (denoted as ATF+ATG+AAG). The experimental results are shown in Fig. 4. It is known from Fig. 4 that the attribute and theme features (ATF) achieve very encouraging results on all the databases, which indicates that the extracted attribute and theme features can effectively describe the image aesthetic quality. In addition, when the ATG is combined to mine the relationship between image theme and visual attribute, the performance further improves. This demonstrates that ATG is advantageous over the commonly used FC pooling. Finally, the proposed TAVAR consisting of all components achieves the best performance. These results demonstrate the necessity and reasonability of integrating all components for reasoning the image aesthetic score.

F. Effectiveness of GCN-based Bilevel Reasoning

In this work, we propose to build the GCN-based theme-aware visual attribute reasoning for handling the IAA task. To investigate the effectiveness of the proposed GCN-based

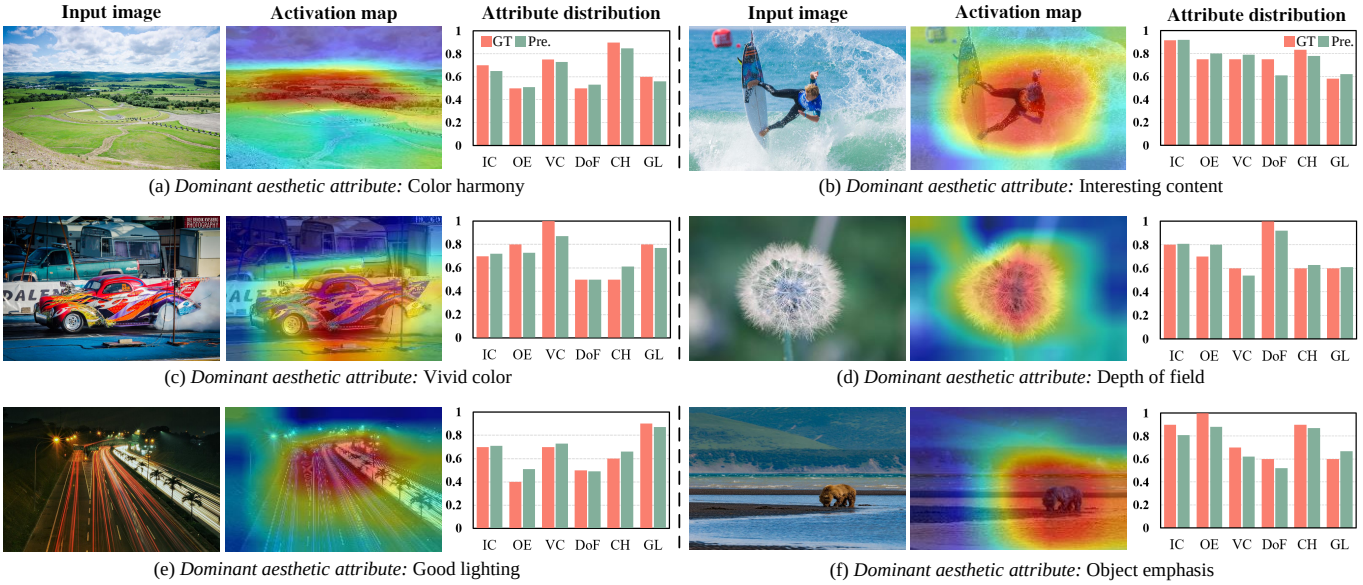


Fig. 5. Visual examples of the proposed TAVAR model on six testing images with different dominant aesthetic attributes. IC: Interesting Content; OE: Object Emphasis; VC: Vivid Color; DoF: Depth of Field; CH: Color Harmony; GL: Good Lighting; GT: Ground Truth; Pre.: Prediction. (Best viewed in color and zoomed in.)

TABLE VII
PERFORMANCE OF THE PROPOSED MODEL USING DIFFERENT FEATURE FUSING BLOCKS ON THE AADB [16], EVA [39], AVA [40] AND PARA [41] DATABASES.

Model	AADB		EVA		AVA		PARA	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Concatenation [60]	0.759	0.751	0.803	0.778	0.720	0.704	0.937	0.905
Point-wise Addition [34]	0.761	0.757	0.802	0.783	0.719	0.702	0.934	0.901
Self-attention [37]	0.761	0.753	0.798	0.779	0.725	0.709	0.936	0.908
Bilevel GCN [30]	**0.763**	**0.761**	**0.810**	**0.799**	**0.736**	**0.725**	**0.940**	**0.911**

TABLE VIII
PERFORMANCES OF THE PROPOSED VISUAL ATTRIBUTE ANALYSIS NETWORK (VAAN) AND THEME UNDERSTANDING NETWORK (TUN) USING DIFFERENT ACTIVATION FUNCTIONS. IC: INTERESTING CONTENT; OE: OBJECT EMPHASIS; VC: VIVID COLOR; DOF: DEPTH OF FIELD; CH: COLOR HARMONY; GL: GOOD LIGHTING; TR: THEME RECOGNITION.

Model	OE		VC		DoF		CH		GL		IC		TR
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	ACC (%)
using ReLU	0.677	0.681	0.698	0.703	0.707	0.512	0.541	0.529	0.534	0.497	**0.616**	**0.611**	84.8
using PReLU	**0.687**	**0.691**	**0.704**	**0.712**	**0.718**	**0.513**	**0.551**	**0.543**	**0.545**	**0.515**	0.614	0.608	**85.1**

bilevel reasoning, we further conduct comparison experiments. Specifically, we compare three popular feature fusion strategies including Concatenation [60], Point-wise Addition [34] and Self-attention Fusion [37] by replacing the proposed GCN-based bilevel reasoning. For fair comparisons, all comparison experiments are conducted using the same experimental setting and test on all the four databases including AADB [16], EVA [39], AVA [40] and PARA [41]. The experimental results in terms of PLCC/SRCC values are listed in Table VII. From the experimental results, it can be seen that the proposed model with GCN-based bilevel reasoning is superior to other feature fusing strategies on all databases, which further demonstrates the reasonability of using GCN to mine the interaction between image theme and visual attributes.

G. Effectiveness of Activation Function

Considering that the Parametric Rectified Linear Unit (PReLU) [36] can improve the performance of deep model and reduce the risk of overfitting compared with ReLU, we introduce PReLU as activation function in MLPs of the proposed visual attribute analysis network and the theme understanding network. To verify the effectiveness of PReLU, we further conduct comparison experiment by replacing PReLU with ReLU in VAAN and TUN. For fair comparison, both PReLU and ReLU are trained using the same experimental setting. The results are summarized in Table VIII, where the best results are marked in boldface. From the results, among the six attribute prediction tasks in VAAN, five tasks using PReLU

achieve better performance than using ReLU. For the theme classification task, PReLU improves the accuracy by 0.3% compared with ReLU. The experimental results demonstrate that the performance of our model using PReLU as activation function is significantly better than that using ReLU.

H. Model Interpretation

To intuitively demonstrate the explainability of the proposed TAVAR model, we perform a visual experiment on six high-aesthetics images, which are characterized by different dominant visual attributes. Fig. 5 shows the testing images and activation maps via the commonly used CAM [61] method, as well as the corresponding predicted visual attributes and ground truth values.

From Fig. 5, we have the following observations. (1) The appearances of all activation maps are consistent with the position of attention when people judge image aesthetics. For example, the dominant aesthetic attribute of Fig. 5 (a) is color harmony, where the green grass in the near vicinity and the cyan hills in the distance make the image look beautiful. The activation map corresponds well to these regions. In Fig. 5 (b), the local region of the surfer looks interesting, where people will pay more attention when judging the aesthetic quality, and the activation map also covers this region. The dominant aesthetic attribute of Fig. 5 (c) is vivid color, the activation map also shows higher attention values at the car region, which is of prime importance for aesthetic judgment. Similar results can be found for the other images. (2) The proposed TAVAR model can accurately predict the aesthetic attributes of the image, which are very close to the ground truth values. From these visual results, we know that TAVAR can capture the regions with dominant aesthetic attribute, which is consistent with the human perception. The possible reason is that the pipeline of the proposed TAVAR model simulates the process of human perception in image aesthetics, in which the pre-trained visual attribute analysis network and theme understanding network allow the model to adaptively focus on the regions with the dominant aesthetic attribute according to different image themes. By bilevel aesthetic reasoning, the proposed model can produce the final aesthetic prediction, which is more consistent with the human judgment on image aesthetics. In addition, the predicted visual attributes can provide reasonable explanations on why a particular aesthetic quality score is predicted, which is very useful in real-world applications.

I. Computational Efficiency

For the IAA task, computational complexity is also of great importance in real-world applications. To evaluate the computational efficiency of the proposed TAVAR model, we compare the average processing speed of TAVAR with several popular deep IAA models on AVA database. For fair comparison, all tests are implemented on the same computer with Intel Core i7-9700K CPU @ 3.60GHz, and NVIDIA GeForce RTX 3090 24G GPU and the same deep learning environment with Pytorch and Python 3.7. Moreover, we use the default setting of the source codes without any modification. The average images-per-second is calculated for evaluating the computational efficiency of each model. Fig. 6 shows the experimental results.

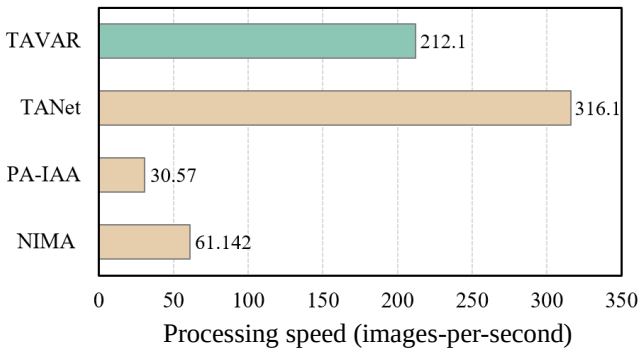


Fig. 6. Average processing speed of different IAA models.

From Fig. 6, it is observed that the proposed TAVAR model can process more than 212 images per second, which demonstrates the potential of TAVAR in practical applications. Since the proposed TAVAR contains a visual attribute analysis network and a theme understanding network, they would take more time. Therefore, TAVAR is slower than TANet. However, the visual attribute analysis network and the theme understanding network not only allow TAVAR to accurately predict aesthetic scores, but also make it interpretable.

V. Conclusion

In this paper, we have presented a new IAA model based on theme-aware visual attribute reasoning, dubbed TAVAR, which simulates the aesthetic perception process of human using a bilevel reasoning framework. Specifically, to investigate the coupled relationship between image theme and visual attribute, we propose the first level attribute-theme graph (ATG), which enables the model to more effectively capture dominant visual attributes according to different image themes, producing the theme-aware visual attribute features. Further, the second level attribute-aesthetics graph (AAG) is built to model the complementary relationship between visual attribute features and aesthetic features in determining the overall image aesthetic quality. We have demonstrated the superior performance of TAVAR on four public IAA databases. Furthermore, visual analysis has shown that the proposed TAVAR model can provide the reasons for aesthetic decision based on the predicted visual attributes, which makes TAVAR explainable.

References

[1] J. Hou, H. Ding, W. Lin, W. Liu, and Y. Fang, "Distilling knowledge from object classification to aesthetics assessment," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 32, no. 11, pp. 7386–7402, 2022.

[2] Y. Niu, S. Chen, B. Song, Z. Chen, and W. Liu, "Comment-guided semantics-aware image aesthetics assessment," *IEEE Trans. Circuits and Syst. Video Technol.*, p. Early Access, Jan. 2022.

[3] Y. Deng, C. C. Loy, and X. Tang, "Image aesthetic assessment: an experimental survey," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 80–106, Oct. 2017.

[4] Y. Niu and F. Liu, "What makes a professional video? a computational aesthetics approach," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 22, no. 7, pp. 1037–1049, Jul. 2012.

[5] C. Chen, B. Zhou, and W. H. Mow, "RA Code: A robust and aesthetic code for resolution-constrained applications," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 28, no. 11, pp. 3300–3312, Aug. 2018.

[6] F. Liu, C. Gao, Y. Sun, Y. Zhao, F. Yang, A. Qin, and D. Meng, "Infrared and visible cross-modal image retrieval through shared features," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 31, no. 11, pp. 4485–4496, Nov. 2021.

[7] C. Guo, X. Tian, and T. Mei, "Multigranular event recognition of personal photo albums," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1837–1847, Jul. 2018.

[8] Y. S. Rawat and M. S. Kankanhalli, "Clicksmart: A context-aware viewpoint recommendation system for mobile photography," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 27, no. 1, pp. 149–158, Jan. 2017.

[9] Y. S. Rawat, M. Shah, and M. S. Kankanhalli, "Photography and exploration of tourist locations based on optimal foraging theory," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 30, no. 7, pp. 2276–2287, Jul. 2020.

[10] X. Chai, F. Shao, Q. Jiang, and Y.-S. Ho, "Roundness-preserving warping for aesthetic enhancement-based stereoscopic image editing," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 31, no. 4, pp. 1463–1477, Jul. 2021.

[11] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, vol. 1, Jul. 2006, pp. 419–426.

[12] S. He, Y. Zhang, R. Xie, D. Jiang, and A. Ming, "Rethinking image aesthetics assessment: Models, datasets and benchmarks," *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2022.

[13] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Sep. 2018.

[14] K. Ligaya, S. Yi, I. A. Wahle, K. Tanwisuth, and J. P. Doherty, "Aesthetic preference for art can be predicted from a mixture of low-and high-level visual features," *Nat. Hum. Behav.*, vol. 5, no. 6, pp. 743–755, Jun. 2021.

[15] B. Barnbaum, *The art of photography : A personal approach to artistic expression*. Rocky Nook, 2017.

[16] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 662–679.

[17] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2008, pp. 386–399.

[18] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, Dec. 2013.

[19] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[20] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[21] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "RAPID: Rating pictorial aesthetics using deep learning," in *Proc. ACM Int. Conf. Multimedia*, Nov. 2014, pp. 457–466.

[22] X. Jin, L. Wu, X. Li, X. Zhang, J. Chi, S. Peng, S. Ge, G. Zhao, and S. Li, "ILGNet: inception modules with connected local and global features for efficient image aesthetic quality classification using domain adaptation," *IET Comput. Vis.*, vol. 13, no. 2, pp. 206–212, Jun. 2019.

[23] Q. Chen, W. Zhang, N. Zhou, P. Lei, Y. Xu, Y. Zheng, and J. Fan, "Adaptive fractional dilated convolution network for image aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2020, pp. 14 102–14 111.

[24] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multi-task learning for generic and personalized image aesthetics assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 3898–3910, Jan. 2020.

[25] Y. Shu, Q. Li, S. Liu, and G. Xu, "Learning with privileged information for photo aesthetic assessment," *Neurocomputing*, vol. 404, pp. 304–316, May 2020.

[26] D. She, Y. Lai, G. Yi, and K. Xu, "Hierarchical layout-aware graph convolutional network for unified aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2021, pp. 8471–8480.

[27] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Trans. Knowl. Data En.*, vol. 34, no. 1, pp. 249–270, Mar. 2022.

[28] S. Chandra, N. Usunier, and I. Kokkinos, "Dense and low-rank gaussian crfs using deep embeddings," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5103–5112.

[29] G. Bertasius, L. Torresani, S. X. Yu, and J. Shi, "Convolutional random walk networks for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jul. 2017, pp. 858–866.

[30] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, Sep. 2016.

[31] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, Jan. 2020.

[32] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph r-cnn for scene graph generation," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2018, pp. 670–685.

[33] W. Nie, R. Chang, M. Ren, Y. Su, and A. Liu, "I-GCN: Incremental graph convolution network for conversation emotion detection," *IEEE Trans. Multimedia*, p. Early Access, Jan. 2021.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 770–778.

[35] G. Malu, R. Bapi, and B. Indurkhya, "Learning photography aesthetics with deep cnns," *arXiv preprint arXiv:1707.03981*, Apr. 2017.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[37] L. Li, T. Song, J. Wu, W. Dong, J. Qian, and G. Shi, "Blind image quality index for authentic distortions with local and global deep feature aggregation," *IEEE Trans. Circuits and Syst. Video Technol.*, pp. 1–1, Early Access 2021.

[38] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Oct. 2021, pp. 10 012–10 022.

[39] C. Kang, G. Valenzise, and F. Dufaux, "EVA: An explainable visual aesthetics dataset," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2020, pp. 5–13.

[40] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2408–2415.

[41] Y. Yang, L. Xu, L. Li, N. Qie, Y. Li, P. Zhang, and Y. Guo, "Personalized image aesthetics assessment with rich attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2022, pp. 19 861–19 869.

[42] H. Zeng, Z. Cao, L. Zhang, and A. C. Bovik, "A unified probabilistic formulation of image aesthetic assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 1548–1561, Sep. 2020.

[43] L. Li, Y. Huang, J. Wu, K. Gu, and Y. Fang, "Predicting the quality of view synthesis with color-depth image fusion," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 31, no. 7, pp. 2509–2521, Jul. 2021.

[44] T. Song, L. Li, P. Chen, H. Liu, and J. Qian, "Blind image quality assessment for authentic distortions by intermediary enhancement and iterative training," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 32, no. 11, pp. 7592–7604, Jun. 2022.

[45] Z. Pan, H. Zhang, J. Lei, Y. Fang, X. Shao, N. Ling, and S. Kwong, "DACNN: Blind image quality assessment via a distortion-aware convolutional neural network," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 32, no. 11, pp. 7518–7531, Jul. 2022.

[46] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Trans. Circuits and Syst. Video Technol.*, vol. 32, no. 9, pp. 5944–5958, Sep. 2022.

[47] L. Hou, C. P. Yu, and D. Samaras, "Squared earth mover's distance-based loss for training deep neural networks," *arXiv preprint arXiv:1611.05916*, Nov. 2016.

[48] D. Liu, R. Puri, N. Kamath, and S. Bhattacharya, "Composition-aware image aesthetics assessment," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2020, pp. 3569–3578.

[49] V. Hosu, B. Goldlcke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2019, pp. 9367–9375.

[50] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 5128–5137.

[51] L. Celona, M. Leonardi, P. Napoletano, and A. Rozza, "Composition and style attributes guided image aesthetic assessment," *IEEE Trans. Image Process.*, vol. 31, pp. 5009–5024, Jul. 2022.

[52] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Nov. 2017, pp. 722–731.

[53] P. Lv, M. Wang, Y. Xu, Z. Peng, J. Sun, S. Su, B. Zhou, and M. Xu, "USAR: An interactive user-specific aesthetic ranking framework for images," in *Proc. ACM Int. Conf. Multimedia*. ACM, Oct. 2018, pp. 1328–1336.

[54] X. Zhang, X. Gao, W. Lu, and L. He, "A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2815–2826, Apr. 2019.

[55] A. Howard, M. Sandler, G. Chu, L. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, and V. Vasudevan, "Searching for mobilenetv3," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1314–1324.

[56] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, May 2019, pp. 6105–6114.

[57] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jul. 2017, pp. 2261–2269.

[58] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Trans. Image Process.*, vol. 30, pp. 3474–3486, Mar. 2021.

[59] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, Dec. 2021, pp. 10 347–10 357.

[60] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Oct. 2015, pp. 234–241.

[61] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Dec. 2016, pp. 2921–2929.

Leida Li (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. In 2008, he was a Research Assistant with the Department of Electronic Engineering, Kaohsiung University of Science and Technology, Kaohsiung, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics.

Yipo Huang (Student Member, IEEE) received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2017, and the M.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Xidian University, Xi'an, China. His research interests include multimedia quality assessment, computational aesthetics and perceptual image processing.

Jinjian Wu (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. From 2011 to 2013, he was a Research Assistant with Nanyang Technological University, Singapore, where he was a Postdoctoral Research Fellow from 2013 to 2014. From 2015 to 2019, he was an Associate Professor with Xidian University, where he has been a Professor since 2019. His research interests include visual perceptual modeling, biomimetic imaging, quality evaluation, and object detection.

Yuzhe Yang (Member, IEEE) received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2019. He received the M.S. degree from University of Southampton, Southampton, U.K., in 2020. Currently, he is a computer vision algorithm engineer at OPPO Research Institute. His research interests include multimedia affective computing, multimedia quality assessment and representation learning.

Yaqian Li received the B.S. degree from Lanzhou University, and the M.S. degree from Harbin Institute of Technology, China, in 2011 and 2013, receptively. He is currently the Technical Lead of visual search and understanding with OPPO Research Institute. His professional interests lie in the broad area of visual recognition, image retrieval, object detection, image quality assessment, and multimodality learning.

Yandong Guo received the B.S. and M.S. degrees in ECE from Beijing University of Posts and Telecommunications, China, in 2005 and 2008, receptively, and the Ph.D. degree in ECE from Purdue University at West Lafayette in 2013. He is currently the Chief Scientist of Intelligent Perception with OPPO and chair the AI strategic planning for OPPO. His professional interests lie in the broad area of computer vision, imaging systems, human behavior understanding and biometric, and autonomous driving.

Guangming Shi (Fellow, IEEE) received the B.S. degree in automatic control, the M.S. degree in computer control, and the Ph.D. degree in electronic information technology from Xidian University, Xi'an, China, in 1985, 1988, and 2002, respectively. He had studied at the University of Illinois and The University of Hong Kong. Since 2003, he has been a Professor with the School of Electronic Engineering, Xidian University. He is currently the Academic Leader in circuits and systems with Xidian University. He has authored and coauthored more than 200 papers in journals and conferences. His research interests include compressed sensing, brain cognition theory, multirate filter banks, image denoising, low-bitrate image and video coding, and implementation of algorithms for intelligent signal processing. He was awarded the Cheung Kong Scholar Chair Professor by the Ministry of Education in 2012. He served as the Chair for the 90th MPEG and 50th JPEG of the international standards organization, and the Technical Program Chair for FSKD06, VSPC in 2009, the IEEE Pulse Code Modulation in 2009, the SPIE Visual Communications and Image Processing in 2010, and the IEEE International Symposium on Circuits and Systems in 2013.