# Hierarchical Layout-Aware Graph Convolutional Network for Unified Aesthetics Assessment

Dongyu She[1], Yu-Kun Lai[2], Gaoxiong Yi[3], Kun Xu[1*]
[1]Tsinghua University [2]Cardiff University [3]Tencent

## Abstract

*Learning computational models of image aesthetics can have a substantial impact on visual art and graphic design. Although automatic image aesthetics assessment is a challenging topic by its subjective nature, psychological studies have confirmed a strong correlation between image layouts and perceived image quality. While previous state-of-the-art methods attempt to learn holistic information using deep Convolutional Neural Networks (CNNs), our approach is motivated by the fact that Graph Convolutional Network (GCN) architecture is conceivably more suited for modeling complex relations among image regions than vanilla convolutional layers. Specifically, we present a Hierarchical Layout-Aware Graph Convolutional Network (HLA-GCN) to capture layout information. It is a dedicated double-subnet neural network consisting of two LA-GCN modules. The first LA-GCN module constructs an aesthetics-related graph in the coordinate space and performs reasoning over spatial nodes. The second LA-GCN module performs graph reasoning after aggregating significant regions in a latent space. The model output is a hierarchical representation with layout-aware features from both spatial and aggregated nodes for unified aesthetics assessment. Extensive evaluations show that our proposed model outperforms the state-of-the-art on the AVA and AADB datasets across three different tasks. The code is available at* http://github.com/days1011/HLAGCN.

## 1. Introduction

Automatic image aesthetics assessment (IAA) has attracted increasing attention in recent years due to its potential applications, *e.g.*, image retrieval, album photo recommendation, image enhancement [7, 14, 44], *etc*. Early efforts focus on extracting elaborately designed hand-crafted features according to the known photographic principles, *e.g.*, the rule-of-thirds [8], color harmony [35], and global image layout [16, 30]. With the advance of Convolution Neural Networks (CNNs) [20], recent methods aim to map image aesthetics to different types of formulations using CNNs, *i.e.*, binary classification labels [25, 40], aesthetic scores [37] and their distributions [3, 10, 11]. Although significant progress has been achieved, the performance of employing CNNs for IAA is often compromised due to the following two main inherent constraints.

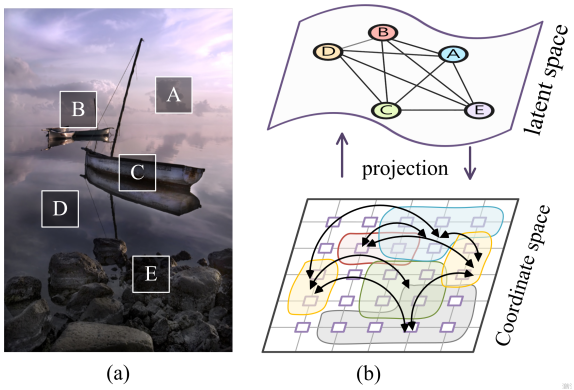


Figure 1. To capture layout information over the whole input space, we first partition an image into spatial nodes. Features from the disjoint regions (denoted in colors) in the coordinate space are then projected into the latent space for efficient graph reasoning.

First, general deep aesthetic models require additional operations (*e.g.*, cropping, warping, or padding) to generate the fixed-size input needed for mini-batch compatibility. However, the altered object aspect ratios or image layouts often impair the image aesthetics and introduce label noise for representation learning. Several methods try to address such limitations by either feeding the original-sized images [10, 29] or incorporating multiple patches [28, 41] into the network, which slows down the training and inference process significantly. Second, the layout information is crucial for assessing visual aesthetics since the appropriate arrangement of visual elements in the photograph can add balance and harmony [31, 46]. For example, Fig. 1 (a)

*Corresponding author, xukun@tsinghua.edu.cn

shows an example from *DPChallenge*[1], an on-line community for photography amateurs. In this landscape scene, five denoted elements including sky, sailboats, lake and reef, are presented with a comfortable and balanced layout collectively manifesting the high-level aesthetics, which won first place with an average vote of 7.83. However, due to the inherent limitation of the regular receptive field, the convolution operations of CNNs are typically inefficient at capturing relations among distant regions in the coordinate space.

To address these problems, we propose a double-subnetwork framework based on the Graph Convolution Networks (GCNs) [17], leveraging layout information for assessing visual aesthetics. Specifically, we first use a Fully Convolutional Network (FCN) to preserve the spatial information of the convolutional feature maps, which are viewed as representations of nodes throughout the entire spatial grid. Based on the spatial nodes, we construct an aesthetics graph by connecting every pair of nodes to form edges, and embedding the information regarding content similarity and aspect ratio-embedded spatial relations between nodes as edge weights. Then instead of relying solely on standard convolutions to model aesthetic information, the proposed 1$^{\text{st}}$ Layout-Aware Graph Convolutional Network (LA-GCN) module performs graph convolutions on the graph. To enable efficient global reasoning over disjoint regions, we further propose to aggregate nodes with similar semantics in a latent space and perform graph reasoning via the 2$^{\text{nd}}$ LA-GCN module, as shown in Fig. 1 (b). By fusing features from both spatial and aggregated nodes, our Hierarchical LA-GCN empowers the GCN model with the capacity of learning hierarchical representation for IAA.

Our contributions are summarized as follows: First, we present a layout-aware graph convolution module to explicitly relate the aesthetic perception to the image layout attributes in an end-to-end fashion; second, we propose the HLA-GCN (Hierarchical LA-GCN) by extending the LA-GCN module to a hierarchical architecture for learning visual representations from both coordinate and latent spaces. Our proposed framework performs favorably against the state-of-the-art methods on the AVA and AADB datasets for unified aesthetics assessment, *i.e*., quality classification, score regression, and distribution prediction.

## 2. Related Work

In this section, we provide a brief review of image aesthetics assessment (IAA) methods [3, 7, 42, 47] especially those on preserving image layouts and compositions as well as graph-based representation learning methods [4, 13, 17, 39] that are closely related to our work.

**Image Aesthetics Assessment** Conventional methods design image layouts by approximating simple photography composition guidelines, *e.g*., visual balance, rule of thirds, and diagonal dominance [36, 46, 49], while recent efforts focus on using CNN for learning aesthetic representations [3, 7, 42]. As aesthetics can be influenced by the transformations applied to the input, several existing CNN methods try to overcome the limitations by designing multi-column architectures that take multiple patches as inputs and aggregate their contributions to the aesthetics score [26–29]. Kao *et al*. [15] propose to learn aesthetic features by dividing the images into different categories and training associated networks capturing different information in terms of scene, object, and texture. In addition, MNA-CNN [29] proposes to preserve compositions by feeding the original image into the network once at a time, which is not mini-batch compatible since images with different aspect ratios cannot be concatenated into batches. Ma *et al*. [28] propose A-Lamp to crop salient patches from the original image without any transformation, and then build an attribute relation graph over these regions to preserve the spatial layout of the image. As A-Lamp requires a manually designed aggregation structure, Liu *et al*. [25] propose to use GCN to model the mutual dependencies of the local regions. However, these methods do not explicitly build aesthetics-related graph for modeling complex relations among image regions which is an important cue for aesthetics assessment.

**Graph-based Representation Learning** Graph-based methods have shown to be an efficient approach to relation reasoning. Early efforts including CRFs (Conditional Random Fields) [2] and random walk networks [1] are proposed based on the graph model for effective image segmentation. Recently, a great deal of research on generalizing convolution to graph-based data has emerged [5, 17, 24, 34]. Among these methods, the Graph Convolution Network (GCN) [17] serves as a simplified model with a 1-st approximation of the Chebyshev expansion, which restricts the convolution to operating locally. Wang *et al*. [45] propose to capture relations between regions detected by an object detector via GCN, while Chen *et al*. [4] propose a generic trainable module Global Reasoning unit for reasoning between disjoint and distant regions. In this paper, we exploit the reasoning power of GCN to build an aesthetics-related graph for relation reasoning in both coordinate and latent spaces.

## 3. Methodology

In this section, we first provide an overview of the proposed hierarchical layout-aware graph convolutional network. Then we present the core LA-GCN module and extend the proposed module to a hierarchical architecture for learning aesthetic representations on both coordinate and latent spaces. Finally, we give the problem formulation of unified aesthetic assessment task and loss function.

[1] https://www.dpchallenge.com/image.php?IMAGE_ID=263833

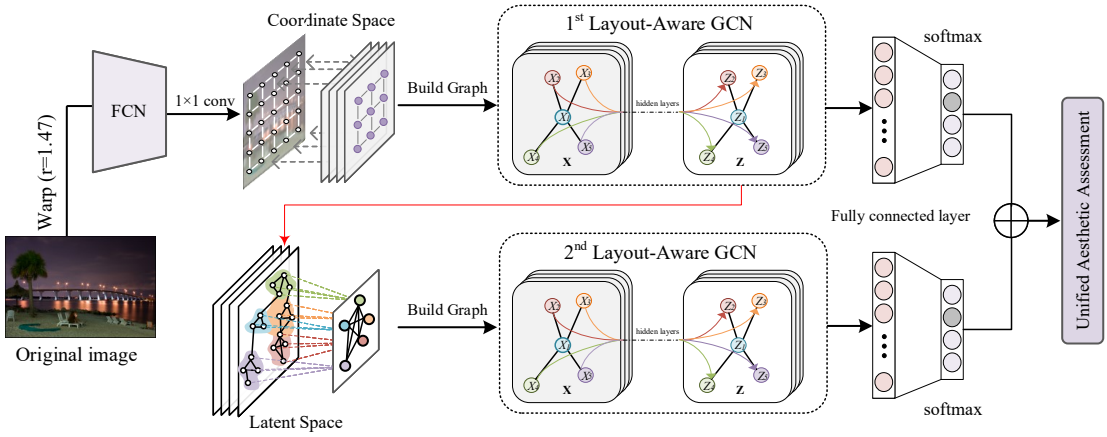


Figure 2. Illustration of the proposed HLA-GCN framework for unified aesthetics assessment. It takes a warped image associated with its aspect ratio as input, which is passed through the fully convolutional layers of CNN. Each channel-wise vector of the obtained feature maps is viewed as a node representation of the corresponding spatial region. The proposed $1^{\text{st}}$ LA-GCN module is followed to construct the aesthetics graph and perform graph reasoning over all the spatial locations. Then the spatial nodes are aggregated by projecting node representations from the coordinate space to the latent space. Based on the aggregated nodes, the $2^{\text{nd}}$ LA-GCN module constructs a fully-connected graph and performs graph reasoning over significant regions. Finally, the features of both spatial and aggregated nodes are fused as a hierarchical representation for aesthetics prediction.

### 3.1. Overview

Motivated by overcoming the inherent limitations of convolution operations, the proposed HLA-GCN aims to model a hierarchical image layout by performing graph reasoning on both coordinate and latent spaces. The architecture of the proposed framework is shown in Fig. 2. Given the input aesthetic image, we view the extracted feature maps as representations of nodes throughout all spatial locations. The proposed framework performs graph reasoning over all the spatial nodes and aggregated significant nodes via two LA-GCN modules. Considering both content and spatial relations, the $1^{\text{st}}$ LA-GCN constructs an aesthetics-related graph and applies graph convolution on all the spatial nodes. By performing graph reasoning in the coordinate space, the proposed framework can model the overall image layout of different visual elements. Furthermore, we propose a node aggregation strategy to aggregate significant nodes in a latent space, where a set of disjoint regions with similar semantics can be projected onto a single representation. The $2^{\text{nd}}$ LA-GCN is employed to perform relation reasoning over the graph constructed in the latent space. Finally, both spatial and aggregated nodes are fused as a hierarchical representation for unified aesthetic assessment.

### 3.2. Layout-aware GCN Module

Given the aesthetic image, we first extract the FCN feature map of shape $W \times H \times C$, where $W$, $H$, and $C$ denote the spatial size, *i.e.*, width and height, and the channel-wise dimension, respectively. As the spatial information is preserved after FCN, we can view the feature map as node representations $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L] \in \mathbb{R}^{L \times C}$ throughout the entire spatial grid, where $L = W \times H$ denotes the total number of spatial locations. To model relations between regions, we construct an aesthetics-related graph based on the node representations. Then by performing graph convolution on the graph, feature maps can be refined by message passing among nodes, resulting in layout-aware visual representations.

**Graph Building** Considering $L$ nodes associated with the node representations $\mathbf{X}$, we first construct an undirected fully connected graph $G_c = (\mathcal{V}_c, \mathcal{E}_c, A_c)$ in the coordinate space. Here, $G_c$ is constructed by its nodes $\mathcal{V}_c$, the set of edges connecting nodes $\mathcal{E}_c$ and adjacent matrix $A_c$ describing the edge weights. The adjacent matrix is defined according to two main types of pair-wise relations, *i.e.*, content similarity, and spatial relations, as follows:

$$\mathbf{A}_c = \mathbf{A}^{\text{sim}} + \mathbf{A}^{\text{spa}} \tag{1}$$

Specifically, we first measure the content relations between region nodes using cosine similarity as follows:

$$sim(\mathbf{x_i}, \mathbf{x_j}) = \frac{< \phi(\mathbf{x_i}), \phi'(\mathbf{x_j}) >}{\|\phi(\mathbf{x_i})\| \|\phi'(\mathbf{x_j})\|}, \tag{2}$$

where $\|\cdot\|$ represents the $\ell_2$-norm and $< \cdot\ , \cdot >$ denotes the inner product. Here, $\phi(\mathbf{x_i}) = \omega \mathbf{x}_i$ and $\phi'(\mathbf{x_j}) = \omega' \mathbf{x}_j$

are two linear transformations applied for increasing generalization ability [25, 45], and $\omega, \omega' \in \mathbb{R}^{C\times C}$ are learnable parameters that can be optimized via back propagation. After computing the pairwise similarity matrix, we obtain the affinity matrix $\mathbf{A}^{\text{sim}} \in \mathbb{R}^{L\times L}$ and adopt the softmax function for normalization on each row of the matrix by:

$$\mathbf{A}^{\text{sim}}_{ij} = \frac{\exp\left(sim\left(\mathbf{x}_i, \mathbf{x}_j\right)\right)}{\sum_{k=1}^{L}\exp\left(sim\left(\mathbf{x}_i, \mathbf{x}_k\right)\right)}, \tag{3}$$

where $\exp(\cdot)$ is the exponential function. In addition to visual contents, image aesthetics can also be affected by the spatial composition of visual elements [28, 43]. As stated in Sec. 1, common preprocessing operations (*e.g.*, cropping, warping) alter image composition and may introduce label noise for representation learning. To address the problem, we adopt layout-preserving warping that preserves the aspect ratio information of the original image, as shown in Fig. 3 (a). As most current CNN models only take square images as input (*i.e.*, $W = H$), the aspect ratio is computed by $r = \frac{w}{h}$. We embed the aspect ratio information by considering spatial relations between pairwise nodes in the graph. To be specific, for the $i$-th node and $j$-th node, we denote their coordinates in the original feature maps as $(s_i^x, s_i^y)$ and $(s_j^x, s_j^y)$, where $1 \leqslant s_i^x, s_j^x \leqslant W$ and $1 \leqslant s_i^y, s_j^y \leqslant H$. The spatial relation between region nodes is then computed by the aspect-ratio-embedded distance:

$$dis(i,j) = \begin{cases} \sqrt{(\Delta s_x \cdot r)^2 + \Delta s_y^2}, & \text{if } r < 1 \\ \sqrt{\Delta s_x^2 + (\frac{\Delta s_y}{r})^2}, & \text{if } r \geqslant 1 \end{cases} \tag{4}$$

where $\Delta s_x = s_i^x - s_j^x$ and $\Delta s_y = s_i^y - s_j^y$. Similarly, the affinity matrix $\mathbf{A}^{\text{dis}} \in \mathbb{R}^{L\times L}$ can be obtained with softmax normalization:

$$\mathbf{A}^{\text{spa}}_{ij} = \frac{\exp\left(-dis\left(i,j\right)\right)}{\sum_{k=1}^{L}\exp\left(-dis\left(i,k\right)\right)}. \tag{5}$$

Intuitively, the spatial relation between nodes shares a similar idea as the receptive field where the features of neighboring nodes are utilized during convolution. Different from convolution considering the receptive field with fixed size (*e.g.*, $3 \times 3$), our spatial relation provides a more flexible way that incorporates 'dynamic receptive field' by adjusting the weights of neighbors for images with different aspect ratios. Fig. 3 (b) shows the weights of adjacent matrix by taking the node in the green box as the anchor. As can be seen, the content similarity matrix enables the network to exploit the semantically related regions, while the spatial relation matrix considers a more flexible receptive field according to the original image's aspect ratio. Combining these aesthetics-related attributes, we can obtain the graph $A_c$ in Eqn. (1).

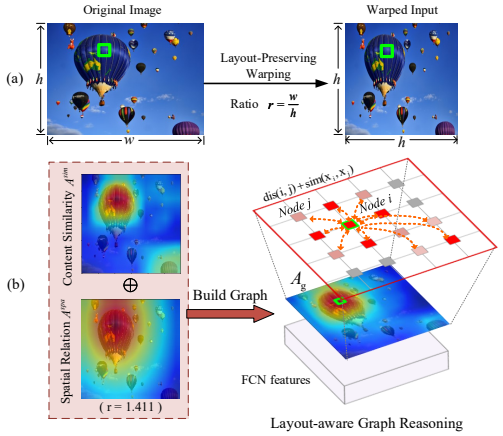


Figure 3. Illustration of layout-aware graph reasoning. For the warped input, we preserve the image aspect ratio by encoding it to the spatial relations, and capturing the semantic information through content similarity. Then the aesthetic graph can be built to guide message passing between the anchor node (denoted in the green bounding box) and the neighbors. Best viewed in color.

**Graph Reasoning** After the aesthetics-related graph is constructed, we then perform reasoning on the graph via graph convolutions. Compared with standard convolutions that have the intrinsic limitation on a regular receptive field, the graph convolutions are able to pass messages among neighbors of each node based on defined relations [4, 45]. Formally, the proposed LA-GCN can be formulated as follows,

$$\mathbf{H}_c^{(m+1)} = \sigma\left(\tilde{\mathbf{A}}_c^{(m)}\mathbf{H}_c^{(m)}\Theta_c^{(m)})\right), \tag{6}$$

where $\mathbf{H}_c^{(m)}$ is the activation in the $m$-th layer, and $\mathbf{H}_c^{(0)} = \mathbf{X}$, $\Theta_c^{(m)} \in \mathbb{R}^{C\times C}$ is the trainable weight matrix of the $m$-th layer. Note that we add a self-loop to each node in the graph following [4, 23], thus the adjacency matrix of the graph is $\tilde{\mathbf{A}}_c^{(m)} = \mathbf{A}_c^{(m)} + \mathbf{I}$, where $\mathbf{A}_c^{(m)}$ is computed by Eqn. (1) using the feature map of the current layer $\mathbf{H}_c^{(m)}$. The identity matrix $\mathbf{I}$ serves as a shortcut connection alleviating the vanishing gradient problem during training, which leads to stable updating during graph message passing. After each layer of graph convolution except for the last layer, we use ReLU as the activation function $\sigma(\cdot)$ on the output.

### 3.3. Hierarchical Prediction Architecture

The proposed 1st LA-GCN incorporates the holistic layout attributes via an aesthetics-related graph in the coordinate space $\Omega$. To describe image layouts effectively, we further aggregate graph nodes to high-level attributes in the latent space $\mathcal{H}$, where each node is aggregated from a set of disjoint regions in $\Omega$. Specifically, given the input representation $\mathbf{X} \in \mathbb{R}^{L\times C}$ we first project the original features

to $\mathbf{Z} \in \mathbb{R}^{K \times C}$, where $K$ is the number of aggregated nodes in the latent space. Similar to [4], we formulate the projection function as a linear combination (*i.e.*, weighted global pooling) of original features such that the new features can aggregate information from multiple regions. In particular, the feature $\mathbf{z}_i \in \mathbb{R}^{1 \times C}$ of the projected node can be denoted as: $\mathbf{z}_i = \mathbf{b}_i \mathbf{X} = \sum_{j=1}^{L} b_{ij} \mathbf{x}_j$, where the weights $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_K] \in \mathbb{R}^{K \times L}$ are learnable during end-to-end training. We note that the above equation gives a more generic formulation than existing methods [28, 41] that require additional object bounding boxes.

Based on these $K$ aggregated nodes, we further construct the graph $G_l = (\mathcal{V}_l, \mathcal{E}_l, A_l)$ and perform the graph reasoning with the 2nd LA-GCN, denoted as:

$$\mathbf{H}_l^{(m+1)} = \sigma\left(\tilde{\mathbf{A}}_l^{(m)} \mathbf{H}_l^{(m)} \Theta_l^{(m)}\right), \tag{7}$$

where $\mathbf{H}_l^{(m)}$ is the refined representation in the $m$-th layer, and the initial node representation $\mathbf{H}_l^{(0)} = \mathbf{Z}$. We denote the trainable weights of LA-GCN as $\Theta_l^{(m)} \in \mathbb{R}^{C \times C}$. Similarly, $\tilde{\mathbf{A}}_l^{(m)} = \mathbf{A}_l^{(m)} + \mathbf{I}$. Note that $\mathbf{A}_l^{(m)}$ only considers the content similarity in Eqn. (3) due to the non-Euclidean geometry of the latent space, which is also computed using the feature map of the current layer $\mathbf{H}_l^{(m)}$. To make the module compatible with the 1st LA-GCN module, we map the output features back of the 2nd LA-GCN to the original coordinate space using reverse projection without sacrificing efficiency or effectiveness [4]. Assuming that the output from the 2nd LA-GCN is $\mathbf{H}_l \in \mathbb{R}^{K \times C}$, the representations projected in the coordinate space can be obtained by $\widehat{\mathbf{H}_l} = \mathbf{B}^\top \mathbf{H}_l \in \mathbb{R}^{L \times C}$. Given the outputs $\mathbf{H}_c$ and $\widehat{\mathbf{H}_l}$ from two modules, the global average pooling and fully connected layers are added to each output, mapping the representations to score distributions $\hat{\mathbf{p}}_c, \hat{\mathbf{p}}_l \in \mathbb{R}^S$. Thus, the final predicted distributions can be denoted as: $\hat{\mathbf{p}} = \frac{\sigma(\hat{\mathbf{p}}_c) + \sigma(\hat{\mathbf{p}}_l)}{2}$.

### 3.4. Problem Formulation

Generally, there are three kinds of aesthetic labels (*i.e.*, the mean score, binary class, and distribution), which are treated as different IAA problems in previous methods [27, 28, 41]. Following a recent trend [47], our objective is to learn a deep network that allows solving all three IAA tasks using a single model trained for the distribution prediction task, which can provide more supervisions about image aesthetics.

Formally, assuming that we have $N$ sample images, raw annotations are first collected in the form of score histograms and then $\ell_1$-normalized as the aesthetic score distributions. The ground truth distribution can be denoted as $\mathbf{p} = [p^{(1)}, p^{(2)}, \cdots, p^{(S)}]$, where $\sum_{s=1}^{S} p^{(s)} = 1$. In AVA [33], the total number of score bins $S$ is 10, and thus each vote ranges from 1 to 10. Following the previous methods [3, 42], our framework is optimized to predict the aesthetic score distribution by minimizing the EMD (Earth Mover's Distance) loss [42] with the ordered distribution distance as follows:

$$\mathcal{L}(\mathbf{p}, \hat{\mathbf{p}}) = \left(\frac{1}{S} \sum_{k=1}^{S} |CDF_{\mathbf{p}}(k) - CDF_{\hat{\mathbf{p}}}(k)|^{\hat{r}}\right)^{1/\hat{r}}, \tag{8}$$

where $CDF_{\mathbf{p}}(k) = \sum_{i=1}^{k} p^{(i)}$ is the cumulative distribution function. During the inference phase, our framework directly predicts the distribution $\hat{\mathbf{p}}$, and then aesthetic score $\hat{\mu}$ and binary class $\hat{c}$ can be inferred accordingly. Mean aesthetic scores are formulated as the average rating scores $\hat{\mu} = \sum_{s=1}^{S} s \times \hat{p}^{(s)}$, and binary class labels $\hat{c} \in \{0, 1\}$ can be assigned by thresholding the average score with $\hat{c} = \mathbf{1}(\hat{\mu} \geqslant \lceil \frac{S}{2} \rceil)$, where $\mathbf{1}(\cdot)$ the indicator function.

## 4. Experiments

In this section, we evaluate the proposed method against state-of-the-art algorithms to demonstrate the effectiveness of HLA-GCN for three IAA tasks.

### 4.1. Datasets

**AVA dataset** [33] is the largest publicly available benchmark in the research field of IAA, which contains more than $25k$ images collected from $DPChallenge$. Each image is rated by an average of 210 users, and the ratings range from 1 to 10 with 10 being the highest aesthetic score. The obtained score histogram is $\ell_1$-normalized to generate our target training and testing distributions. For the binary aesthetics task, images with average scores smaller than 5 are labeled with low-level aesthetics, and the other images are labeled as high-level aesthetics following the same routine as [25, 29, 42]. We employ the training/test split provided by [22] and randomly sample $2,000$ images from the training set for validation. Since some images are not available, our experiments use a total of $235,503$ images for training/validation and $19,997$ images for testing.

**AADB dataset** includes 10,000 photographic images collected from $Flickr$. Each image is annotated by, on average, five people with integer scores ranging from 1 to 5. Following the previous work [18, 19, 21], we use the standard split with 8,500 images for training, 500 images from validation, and 1,000 images for testing. We normalize the score histogram to generate distributions and discard about 100 images that have only one rating following [47].

### 4.2. Experimental Setup

**Training and Implementation** We use ResNet-50 and ResNet-101 [9] pretrained on ImageNet [6] as the backbone of our proposed framework. The entire model is optimized by minimizing the EMD loss with $\hat{r} = 2$ using

Table 1. Comparison with the SOTA methods on AVA. Note that † and ‡ denote that multi-patches and additional annotations are used for aesthetics assessment, respectively. We also show the backbone and input image size of each method. Here, 'Resize$(\cdot)$' denotes that the smaller image dimension of input is resized to a specified dimension while maintaining the original aspect ratio. The results were quoted from their original papers, and '-' denotes the unreported metrics. The corresponding training and testing splits are also shown.

| Methods | Network | Image Size | Split | Classification | Score Regression | | | Distribution | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Accuracy ↑ | SRCC ↑ | LCC ↑ | MSE ↓ | $EMD_1$ ↓ | $EMD_2$ ↓ |
| DMA-Net [27]† | AlexNet | 227 × 227 | from [33] | 75.4 % | - | - | - | - | - |
| MNA-CNN [29]† | VGG16 | 224 × 224 | from [33] | 77.1 % | - | - | - | - | - |
| Zeng *et al.* [47] | ResNet-101 | 384 × 384 | from [33] | 80.8 % | 0.719 | 0.720 | 0.275 | - | 0.065 |
| APM [32] | ResNet-101 | Resize(500) | from [33] | 80.3 % | 0.709 | | 0.279 | - | **0.061** |
| A-Lamp [28]† | VGG16 | 224 × 224 | from [33] | 82.5 % | - | - | - | - | - |
| $MP_{ada}$ [41]† | ResNet-18 | 224 × 224 | from [33] | 83.0 % | - | - | - | - | - |
| RGNet [25] | ResNet-101 | 300 × 300 | from [33] | 82.5 % | - | - | - | - | - |
| Hosu *et al.* [10] | InceptionResNet | Full resolution | from [33] | 81.7 % | **0.756** | **0.757** | - | - | - |
| NIMA [42] | Inception-v2 | 299 × 299 | random | 81.5 % | 0.612 | 0.636 | - | 0.050 | - |
| AFDC [3] | ResNet-50 | 320 × 320 | random | 83.0 % | 0.649 | 0.671 | 0.271 | 0.045 | - |
| PA_IAA [22]‡ | Inception-v3 | 299 × 299 | from [22] | 83.7 % | 0.677 | - | - | 0.047 | - |
| PA_IAA [22]‡ | DenseNet-121 | 224 × 224 | from [22] | 82.9 % | 0.666 | - | - | 0.049 | - |
| HLA-GCN | ResNet-50 | 300 × 300 | from [22] | **84.1%** | 0.656 | 0.678 | **0.264** | **0.045** | 0.065 |
| HLA-GCN | ResNet-101 | 300 × 300 | from [22] | **84.6%** | 0.665 | 0.687 | **0.255** | **0.043** | 0.063 |

Table 2. Comparison results on AADB. Our HLA-GCN is based on the ResNet-50. Note that '-' denotes the unreported metrics.

| Methods | SRCC | LCC | MSE | $EMD_1$ | $EMD_2$ |
|---|---|---|---|---|---|
| Reg-Net [19] | 0.678 | - | 0.1268 | - | - |
| NIMA [42] | 0.708 | - | - | - | - |
| RGNet [25] | 0.710 | - | - | - | - |
| PAC-Net [18] | 0.837 | - | - | - | - |
| Lee *et al.* [21] | 0.879 | - | 0.1141 | - | - |
| HLA-GCN | 0.899 | 0.9037 | 0.0980 | 0.0842 | 0.1093 |

Table 3. Inference speed comparison between PyTorch and Jittor. The boldface denotes the faster framework. Inference speed is the average results running on a single GPU with a batch size of 32.

| Platform | Time per image (ms/im) | Iterations per second (it/s) |
|---|---|---|
| PyTorch [38] | 3.8704 | 8.0742 |
| Jittor [12] | **3.6211** | **8.6300** |

stochastic gradient descent (SGD) following [42]. During training, we employ the layout-preserving warping introduced in Sec. 3.2, which rescales each image to the size of $300 \times 300$ associated with the original aspect ratio. To avoid over-fitting, common data augmentations [25, 42] are adopted on the rescaled images as preprocessing, including randomly flipping training images horizontally and scaling images $1.05$ times followed by random cropping. The size of output feature maps after FCN is $10 \times 10$, and we first add a $1 \times 1$ convolutional layer to reduce the channel-wise dimension to $1024$, which is followed by the $1^{\text{st}}$ LA-GCN with $3$ graph convolution blocks and the $2^{\text{nd}}$ LA-GCN with $1$ block for representation learning. In practice, we set the number of the aggregated nodes $K$ to be $\frac{1}{4}$ of the number of total nodes following [4], *i.e.*, $K = 25$. We train the network on a machine with four NVIDIA GeForce GTX 1080 Ti and use a mini-batch size of $100$ images running for a total of $20$ epochs. The initial learning rate is set to $0.01$ for the first $6$ epochs and dampened to $0.001$ for the rest epochs and we use the default weight decay of $5e^{-4}$ with a momentum of $0.9$.

We implement our proposed framework on two different deep learning platforms: PyTorch [38] and Jittor [12]. Experiments (as shown in Tab. 3) show that the implementation using Jittor yields faster inference than that using PyTorch, thanks to the powerful just-in-time compiler of Jittor.

**Evaluation** We employ six commonly used metrics to evaluate three IAA tasks following [42]. Accuracy is reported for binary aesthetic quality classification. For aesthetic score regression, we report Spearman's Rank Correlation Coefficient (SRCC), Linear Correlation Coefficient (LCC) and Mean Squared Error (MSE). For the distribution prediction, EMD with $\hat{r} = 1$ and $\hat{r} = 2$ are both reported, denoted by $EMD_1$ and $EMD_2$, respectively.

### 4.3. Comparison with State-of-the-Art Results

Tab. 1 and Tab. 2 show the results of the state-of-the-art aesthetics prediction models on both AVA and AADB datasets. As seen, our proposed HLA-GCN consistently performs favorably against the state-of-the-art methods for the unified aesthetics assessment. For the binary classification task, compared with multiple-patches methods [27–29, 41], our proposed framework achieves better accuracy and alleviates efforts to aggregate sampling prediction by learning discriminative features directly from the complete images. For example, A-Lamp [28] focuses on capturing the spatial layout of images via a complicated path sampling strategy and manually designed aggregation structure, which crops 50 groups of patches from the original image. Our end-to-end framework is much more efficient without feeding multiple cropping patches and outperforms such a

Table 4. Results on AVA [33] using different variants of the proposed model. Note that all models are based on ResNet-50 and optimized using EMD loss with an input size of 300 × 300. Training speed is the average results running on a single GPU with a batch size of 16.

| Models | #Params | Speed | Classification | Score Regression | | | Distribution | |
|---|---|---|---|---|---|---|---|---|
| | | | Accuracy ↑ | SRCC ↑ | LCC ↑ | MSE ↓ | $EMD_1$ ↓ | $EMD_2$ ↓ |
| Baseline | 23.5M | 6.82 | 81.95 % | 0.6050 | 0.6290 | 0.303 | 0.0470 | 0.0681 |
| + GCN [17] | 36.1M | 6.09 | 83.60 % | 0.6300 | 0.6370 | 0.283 | 0.0460 | 0.0664 |
| + GloRe [4] | 25.8M | 5.05 | 83.46 % | 0.6150 | 0.6368 | 0.280 | 0.0459 | 0.0660 |
| + LAGCN | 35.0M | 6.15 | 83.96 % | 0.6480 | 0.6630 | 0.270 | 0.0452 | 0.0648 |
| + LAGCN (w/o s.) | 35.0M | 6.22 | 83.67 % | 0.6320 | 0.6530 | 0.274 | 0.0456 | 0.0665 |
| + LAGCN (w/o c.) | 28.8M | 6.35 | 83.56 % | 0.6240 | 0.6400 | 0.285 | 0.0464 | 0.0671 |
| + $GCN^{\times 2}$ | 48.8M | 5.63 | 83.95 % | 0.6402 | 0.6620 | 0.269 | 0.0454 | 0.0654 |
| + $LAGCN^{\times 2}$ | 44.6M | 5.33 | 83.98 % | 0.6489 | 0.6704 | 0.266 | 0.0448 | 0.0648 |
| + HLA-GCN (Ours) | 38.2M | 5.83 | 84.10 % | 0.6555 | 0.6776 | 0.264 | 0.0451 | 0.0646 |

method by 2.1%. For the score regression and distribution prediction tasks, APM [32] and Hosu *et al*. [10] achieve the best results on SRCC and LCC since they explicitly keep the original image ratio for each image, while slows down the training phase significantly. Compared with the methods that supports mini-batch training on images with different aspect ratios [3, 42], our framework incorporating layout information shows a consistent improvement on all three tasks. On the AADB dataset, our method improves the performance by a large margin, illustrating that our proposed HLA-GCN can learn more discriminative and accurate aesthetic representations, resulting in better generalization ability for unified aesthetics assessment.

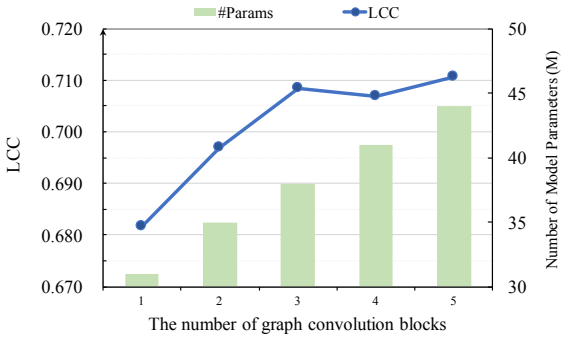


Figure 4. Performance of our proposed HLA-GCN on the AVA validation set using different number of graph convolution blocks. The LCC results and number of model parameters are shown.

### 4.4. Ablation Study

To illustrate the effect of individual components, we conduct ablation studies by analyzing the following variants based on the ResNet-50 NIMA model. (1) **GCN**: this model adds typical GCN [17] after FCN with the same number of blocks (*i.e*., 3) as our first LA-GCN; (2) **GloRe**: this model adds Global Reasoning Unit (GloRe) [4] after FCN projecting the same number of nodes (*i.e*., 25) as our second LA-GCN; (3) **LAGCN**: this model is a truncated version of our HLA-GCN, in which the second LA-GCN and the fusion are removed. (4) **LAGCN (w/o s.)**: this model follows LAGCN while discarding the spatial relation term in Eqn. (1); (5) **LAGCN (w/o c.)**: this model follows LAGCN while discarding the content similarity term in Eqn. (1); (6) **$GCN^{\times 2}$**: this model replaces the LAGCNs in our HLA-GCN with two typical GCNs; (7) **$LAGCN^{\times 2}$**: this model replaces the second LA-GCN module in our HLA-GCN with the first LA-GCN module. (8) **HLA-GCN**: this is our proposed framework including all the components. All models are fine-tuned based on ResNet-50 and optimized using EMD loss with an input size of 300 × 300. Tab. 4 shows the ablation study results and computation. The training speed, *i.e*., images per second, is reported by averaging timing results on a single GPU with a batch size of 16.

**Graph Reasoning Module** We first study the case when only a single graph reasoning module is added to the baseline model. We compare the proposed $1^{st}$ LA-GCN with typical GCN [17] employed in [25] and Global Reasoning Unit [4]. The results show that all graph-based models lead to better representations than the baseline CNN, while our proposed LA-GCN module further improves results by incorporating aesthetics-related attributes in graph reasoning.

**Choice of Adjacency Matrix** To verify the effect of the constructed aesthetics-related graph, we directly discard the spatial relation or the content similarity in Eqn. (1) and see how the model performs. We find that SRCC will drop sharply from 0.6480 to 0.6240 without the content similarity while still outperforms the baseline model showing the advance of 'dynamic receptive field'. When the spatial relation matrix is discarded, the performance also declines. Such results demonstrate that both the spatial information and semantic-aware content are significant attributes and complementary for aesthetics assessment.

**Leveraging Multiple Representations** To verify that the improvement of HLA-GCN is not just because it has more parameters to the network backbone, we further compare

Figure 5. Aesthetic quality prediction results. The ground truth (GT) and predicted (Pre) scores are shown underneath each image. The specific prediction of the first and second LA-GCNs are also shown in brackets in order. The quality is highlighted in different colors, red and blue denote low-level and high-level aesthetics, respectively. The third row shows several typical failure cases.

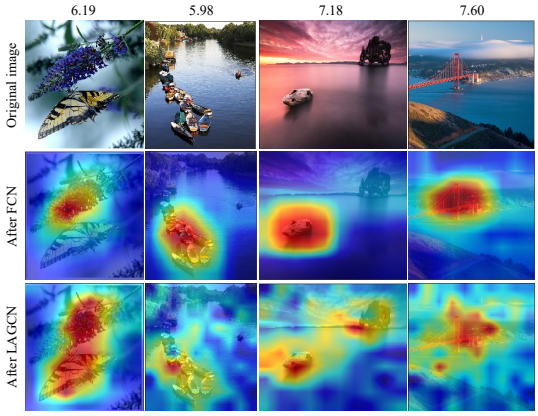


Figure 6. Qualitative results of Class Activation Maps. Given the input image, we extract FCN features and refined features after the first LA-GCN using our proposed model, and visualize the feature response via CAM. The ground truth aesthetic score is also given above each input image. Best viewed in color.

with different double-subnet models. The results show that although stacking GCNs has more parameters, our proposed framework with dedicated design is more effective for learning aesthetic representations. In addition, introducing two LAGCNs in the coordinate space is less effective, mainly due to the overfitting problem.

**Graph Convolution Layers** Fig. 4 shows the validation results of our proposed method using different numbers of graph convolution blocks in the $1^{st}$ LA-GCN. It shows that with the increase of block number from 1 to 3, the performance is boosted, while further increasing the block number leads to no significant improvement. Therefore, we choose to use three blocks of graph convolution for the $1^{st}$ LA-GCN making a tradeoff between performance and model size. In addition, for the $2^{nd}$ LA-GCN, we found no significant difference when using more blocks and thus use one block of graph convolution.

### 4.5. Model Interpretation

**Prediction Results** We first show prediction results from the test set of AVA using our proposed HLA-GCN. In Fig. 5, the first row shows the examples that are predicted with low assessment errors. We find that differences between most low-aesthetic and high-aesthetic quality primarily lie in the harmony of the entire image with a clear semantic meaning. When the images are presented with a simple scene, for example, images with clear foreground and background shown in the second row, the prediction of aesthetics can be more ambiguous and difficult. Due to the concise appearance, these images are misclassified to high-aesthetic quality by the first LA-GCN module, while the second LA-GCN module lessens the confidence mainly due to little interaction in the latent space. Besides, we also show several failure cases in the last row, where images that exhibit large prediction errors tend to require a more abstract high-level understanding of semantics.

**Class Activation Maps** To visualize aesthetics-specific activation within the model, we directly extract Class Activation Maps (CAM) [48] from our fine-tuned HLA-GCN. The activation of feature maps after FCN and the first LA-GCN module are shown in Fig. 6. As can be seen, the attended region (highlighted in red) is able to cover highly correlated objects in the scene after graph convolution, which illustrates that our proposed module refines the FCN features by incorporating image layout information in the network.

## 5. Conclusion

In this paper, we present HLA-GCN, an end-to-end graph-based representation learning framework for image aesthetics assessment. Our proposed method builds a graph representing visual elements and their aesthetics-related attributes, including aspect-ratio-embedded spatial information and semantic-aware contents. By performing graph convolutions, the interactions over the aesthetics-related graph are modeled in both the coordinate space and latent space, leading to the layout-aware hierarchical representation. Extensive evaluations show that our proposed model achieves state-of-the-art performance on the benchmark visual aesthetics datasets.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (No. 61822204, 61521002), and the 2020 Tencent Rhino-Bird Elite Training Program.

## References


[1] Gedas Bertasius, Lorenzo Torresani, Stella X. Yu, and Jianbo Shi. Convolutional random walk networks for semantic image segmentation. In *CVPR*, 2017. 2

[2] Siddhartha Chandra, Nicolas Usunier, and Iasonas Kokkinos. Dense and low-rank Gaussian CRFs using deep embeddings. In *ICCV*, 2017. 2

[3] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. Adaptive fractional dilated convolution network for image aesthetics assessment. In *CVPR*, 2020. 1, 2, 5, 6, 7

[4] Yunpeng Chen, Marcus Rohrbach, Zhicheng Yan, Shuicheng Yan, Jiashi Feng, and Yannis Kalantidis. Graph-based global reasoning networks. In *CVPR*, 2019. 2, 4, 5, 6, 7

[5] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NIPS*, 2016. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[7] Yubin Deng, Chen Change Loy, and Xiaoou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Proc. Mag.*, 34(4):80–106, 2017. 1, 2

[8] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, 2011. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[10] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *CVPR*, 2019. 1, 6, 7

[11] Jingwen Hou, Sheng Yang, and Weisi Lin. Object-level attention for aesthetic rating distribution prediction. In *ACM MM*, 2020. 1

[12] Shi-Min Hu, Dun Liang, Guo-Ye Yang, Guo-Wei Yang, and Wen-Yang Zhou. Jittor: a novel deep learning framework with meta-operators and unified graph execution. *Information Sciences*, 63(222103):1–21, 2020. 6

[13] Jingjia Huang, Zhangheng Li, Nannan Li, Shan Liu, and Ge Li. Attpool: Towards hierarchical feature representation in graph convolutional networks via attention mechanism. In *ICCV*, 2019. 2

[14] Dhiraj Joshi, Ritendra Datta, Elena A. Fedorovskaya, Quang-Tuan Luong, James Ze Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *IEEE Signal Proc. Mag.*, 28(5):94–115, 2011. 1

[15] Yueying Kao, Kaiqi Huang, and Steve J. Maybank. Hierarchical aesthetic quality assessment using deep convolutional neural networks. *Signal Process. Image Commun.*, 47:500–510, 2016. 2

[16] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006. 1

[17] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017. 2, 7

[18] Keunsoo Ko, Jun-Tae Lee, and Chang-Su Kim. Pac-net: Pairwise aesthetic comparison network for image aesthetic assessment. In *ICIP*, 2018. 5, 6

[19] Shu Kong, Xiaohui Shen, Zhe L. Lin, Radomír Mech, and Charless C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 5, 6

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *CVPR*, 2012. 1

[21] Jun-Tae Lee and Chang-Su Kim. Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization. In *ICCV*, 2019. 5, 6

[22] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. Personality-assisted multi-task learning for generic and personalized image aesthetics assessment. *IEEE Trans. Image Process.*, 29:3898–3910, 2020. 5, 6

[23] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *AAAI*, 2018. 4

[24] Ruoyu Li, Sheng Wang, Feiyun Zhu, and Junzhou Huang. Adaptive graph convolutional neural networks. In *AAAI*, 2018. 2

[25] Dong Liu, Rohit Puri, Nagendra Kamath, and Subhabrata Bhattacharya. Composition-aware image aesthetics assessment. In *WACV*, 2020. 1, 2, 4, 5, 6, 7

[26] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Zijun Wang. RAPID: rating pictorial aesthetics using deep learning. In *ACM MM*, pages 457–466, 2014. 2

[27] Xin Lu, Zhe Lin, Xiaohui Shen, Radomír Mech, and James Zijun Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, 2015. 2, 5, 6

[28] Shuang Ma, Jing Liu, and Chang Wen Chen. A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *CVPR*, 2017. 1, 2, 4, 5, 6

[29] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *CVPR*, 2016. 1, 2, 5, 6

[30] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *ICCV*, 2011. 1

[31] Eftichia Mavridaki and Vasileios Mezaris. A comprehensive aesthetic quality assessment method for natural images using basic rules of photography. In *ICIP*, 2015. 1

[32] Naila Murray and Albert Gordo. A deep architecture for unified aesthetic prediction. *CoRR*, abs/1708.04890, 2017. 6, 7

[33] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012. 5, 6, 7

[34] Mathias Niepert, Mohamed Ahmed, and Konstantin Kutzkov. Learning convolutional neural networks for graphs. In *ICML*, pages 2014–2023, 2016. 2

[35] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. Aesthetic quality classification of photographs based on color harmony. In *CVPR*, 2011. 1

[36] Pere Obrador, Ludwig Schmidt-Hackenberg, and Nuria Oliver. The role of image composition in image aesthetics. In *ICIP*, 2010. 2

[37] Bowen Pan, Shangfei Wang, and Qisheng Jiang. Image aesthetic assessment assisted by attributes through adversarial learning. In *AAAI*, 2019. 1

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, 2019. 6

[39] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009. 2

[40] Kekai Sheng, Weiming Dong, Menglei Chai, Guohui Wang, Peng Zhou, Feiyue Huang, Bao-Gang Hu, Rongrong Ji, and Chongyang Ma. Revisiting image aesthetic assessment via self-supervised feature learning. In *AAAI*, 2020. 1

[41] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. Attention-based multi-patch aggregation for image aesthetic assessment. In *ACM MM*, 2018. 1, 5, 6

[42] Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Trans. Image Process.*, 27(8):3998–4011, 2018. 2, 5, 6, 7

[43] Yi Tu, Li Niu, Weijie Zhao, Dawei Cheng, and Liqing Zhang. Image cropping with composition and saliency aware aesthetic score map. In *AAAI*, 2020. 4

[44] Wenguan Wang, Jianbing Shen, and Haibin Ling. A deep network solution for attention and aesthetics aware photo cropping. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1531–1544, 2019. 1

[45] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 2, 4

[46] Lei Yao, Poonam Suryanarayan, Mu Qiao, James Z. Wang, and Jia Li. OSCAR: on-site composition and aesthetics feedback through exemplars for photographers. *Int. J. Comput. Vis.*, 96(3):353–383, 2012. 1, 2

[47] Hui Zeng, Zisheng Cao, Lei Zhang, and Alan C Bovik. A unified probabilistic formulation of image aesthetic assessment. *IEEE Trans. Image Process.*, 29:1548–1561, 2019. 2, 5, 6

[48] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 8

[49] Zihan Zhou, Siqiong He, Jia Li, and James Ze Wang. Modeling perspective effects in photographic composition. In Xiaofang Zhou, Alan F. Smeaton, Qi Tian, Dick C. A. Bulterman, Heng Tao Shen, Ketan Mayer-Patel, and Shuicheng Yan, editors, *ACM MM*, 2015. 2