Object-level Attention for Aesthetic Rating Distribution Prediction

Jingwen Hou
Nanyang Technological University
Singapore
jingwen003@e.ntu.edu.sg

Sheng Yang
Nanyang Technological University
Singapore
syang014@e.ntu.edu.sg

Weisi Lin
Nanyang Technological University
Singapore
wslin@ntu.edu.sg

ABSTRACT

We study the problem of image aesthetic assessment (IAA) and aim to automatically predict the image aesthetic quality in the form of discrete distribution, which is particularly important in IAA due to its nature of having possibly higher diversification of agreement for aesthetics. Previous works show the effectiveness of utilizing *object-agnostic* attention mechanisms to selectively concentrate on more contributive regions for IAA, e.g., attention is learned to weight *pixels* of input images when inferring aesthetic values. However, as suggested by some neuropsychology studies, the basic units of human attention are visual objects, i.e., the trace of human attention follows a series of objects. This inspires us to predict contributions of different regions at *object level* for better aesthetics evaluation. With our framework, region-of-interests (RoIs) are proposed by an object detector, and each RoI is associated with a regional feature vector. Then the contribution of each regional feature to the aesthetics prediction is adaptively determined. To the best of our knowledge, this is the first work modeling object-level attention for IAA and experimental results confirm the superiority of our framework over previous relevant methods.

CCS CONCEPTS

• **Applied computing** → **Media arts**; • **Computing methodologies** → **Computer vision**; **Neural networks**.

KEYWORDS

Image aesthetic assessment, object detection, visual attention

ACM Reference Format:
Jingwen Hou, Sheng Yang, and Weisi Lin. 2020. Object-level Attention for Aesthetic Rating Distribution Prediction. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413695

1 INTRODUCTION

Image aesthetic assessment (IAA) aims to automatically assess aesthetic value of photographs. It can be applied to various applications, e.g. image editing [35], image recommendation [21], image retrieval [27], photo management [14], etc. Three types of IAA tasks have been investigated in the previous relevant works: 1) binary classification, 2) score regression, and 3) rating distribution prediction. Binary classification [25] divides images into high aesthetic class and low aesthetic class according to their average scores, and tries to build a model to predict the binary labels. Score regression task [6], however, aims at directly predicting the average scores. Recently, a more challenging task, aesthetic rating distribution prediction (ARDP) [32], attracts increasing attention of the research community, because not only it better aligns with the uncertainty nature of IAA, but also the results of ARDP can be easily converted to the form of aesthetic scores or binary aesthetic labels. More importantly, this enables a true reflection on the nature of IAA, i.e., the possibly higher diversification of agreement among the population toward aesthetics, due to the well-known fact that beauty is in the eye of the beholder. Therefore, we focus on the ARDP task in this work.

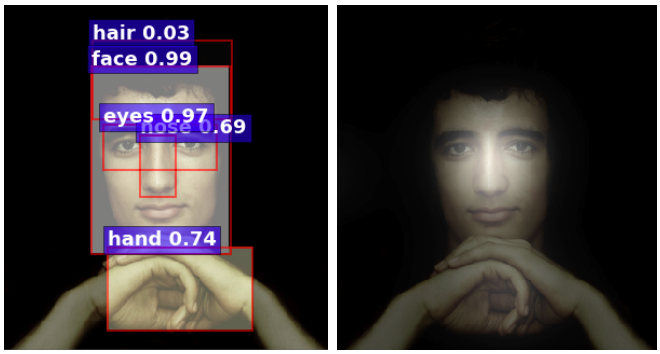


Figure 1: Left: proposed object-level attention with labels (object category and attention weights). Right: object-agnostic attention [39, 40, 48]. Both use higher brightness to denote more important regions. The object-agnostic attention roughly highlights the person's face, while the object-level attention clearly distinguishes face, eyes, and nose, and assigns them with different weights. Object-level attention allows IAA model to determine the contributions of different regions in a finer granularity.

Early deep learning based IAA methods [19, 20] comprised each input by combining a resized image with several small patches randomly cropped from the resized image. Due to the effectiveness, many later methods [10, 22, 30, 32, 48] adopted a similar strategy and utilized one or multiple patches selected from holistic images as part of the model input. However, for the human visual system, patches selected from different regions contribute differently to

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413695

overall aesthetics, which forms a natural requirement for selectively concentrating on more contributive regions when inferring aesthetic values. Recent methods address this issue by adopting attention mechanisms. Sheng *et al.* [30] adopted attention mechanism to dynamically adjust the weight on each randomly selected patch during training. Zhang *et al.* [48] proposed a two-stream model that takes a global view and a local view as input, where the local view is selected according to pixel-level spatial attention learned from the global view. Yang *et al.* [39] proposed a multi-task learning framework that explicitly predicted human fixations for weighting CNN outputs.

As some studies on neuropsychology [29] suggested, the basic units of human attention are visual objects. For example, when appreciating a portrait, the trace of human attention usually follows a series of objects: face, eyes, nose, mouth, etc, and each of these objects contribute differently to the overall aesthetics. However, none of the previous methods considers objectness when predicting attention. As a result, the attention of these object-agnostic methods can only roughly predict the contributions of different regions, while they cannot clearly distinguish the objects making up of these regions and assign the constituent object-level regions with different weights. To enhance the flexibility of the attention modeling, we introduce a framework that learns the weighting of the object-level regions subject to IAA. Fig.1 shows examples of object-level attention and object-agnostic attention. As demonstrated, the object-agnostic attention map can only roughly highlight the face region. It fails to distinguish eyes and nose and assign them with different weights. While the object-level attention map can clearly distinguish face, eyes, nose and hands, and weigh them differently. Thus, the proposed framework has more flexibility when an attentive region is composed of several potential objects. The attentive region is further broken down into object-level regions, and their contributions can be individually and adaptively determined according to the learning task.

To this end, we address the challenge of learning IAA-driven object-level attention with a bottom-up approach. We take advantage of a general object detector, which aims to detect a large diversity of objects in a fine granularity. Once a reliable object detector is built, the detected region-of-interests (RoIs) are used as the basis for modeling IAA-driven attention. Specifically, for proposing RoIs, a Faster-RCNN [28] trained on Visual Genome dataset [11] is leveraged to detect general objects in the images from the largest IAA dataset, AVA dataset [25]. Qualitative analysis shows the chosen object detector can robustly detect most objects in those images, and therefore it provides a reasonable base for verifying the proposed concept of object-level attention. Since multiple RoIs are needed to be considered for modeling object-level attention, and learning end-to-end directly from images is impractical due to the high computational cost, we tackle the problem by adopting features extracted from ImageNet pretrained network for training our IAA model. For modeling IAA-driven object-level attention, pretrained features are firstly extracted from both holistic images and RoIs. Given the extracted features, the attention-based regional feature fusion (ARFF) module in the proposed model dynamically learns the contribution of different object-level RoIs in the feed-forward process of inferring aesthetic value. Finally, regional features are weighted by predicted attention weights and fused for ARDP.

Experimental analysis on the most commonly-used AVA dataset demonstrates our framework is superior to the previous relevant methods in terms of ARDP. Apart from the improvement in flexibility of attention modeling, an extra benefit of training based on RoIs detected by Faster-RCNN is that it better maintains the semantic integrity of the selected regions. In a nutshell, our contributions can be summarized as follows:

- **A framework for learning task-specific object-level attention inspired by the related neuropsychological findings.** The proposed framework infers IAA-driven object-level attention in a bottom-up manner by firstly recognizing RoIs with a generic object detector and secondly modeling IAA task-driven attention based on RoIs. This goes one step beyond previous attention-based IAA methods and improves the flexibility of attention modeling. Experimental results confirm the impact of the proposed framework on predicting aesthetic rating distribution. To the best of our knowledge, this is the first attempt to investigate object-level attention for IAA.
- **Extensive analysis on IAA-driven object-level attention.** Since RoIs detected by a generic object detector can be treated as task-free salient regions, and the IAA-driven attention is constructed with those task-free salient regions, this work is the first attempt to bridge the gap between task-free attention and task-specific attention in the context of IAA. In addition, we also investigate the connections between human fixations and IAA-driven object-level attention.

2 RELATED WORKS

2.1 Attention-based Deep Learning Model

We firstly clarify two similar concepts, *attention model* and *attention-based deep learning model*. Attention models are a class of models that aim to predict **task-free** saliency, including human fixation prediction [8, 12, 17, 33, 40] and salient object detection [13, 34, 36, 45, 47]. While attention-based deep learning models are deep learning models that aim to enhance their representational power by predicting weights of intermediate features in a **task-specific** context. The effectiveness of attention-based deep learning models has been shown in various computer vision tasks, including image classification [7, 37], image captioning [1, 3, 18, 38], and image quality assessment [39], etc. For example, channel-wise attention [3, 7, 37] learns weighting of different channels of CNN outputs. Spatial attention [3, 37, 39] re-weights CNN outputs at different spatial locations. Although task-free attention and task-specific attention are modeled for different purposes, connection between them do exist. We notice that task-free attention has been successfully used for facilitating the prediction of task-specific attention in image captioning [1] and image quality assessment [39]. However, the connection has been rarely explored for IAA. Regarding region-of-interests (RoIs) as task-free salient regions, our framework also adopts those task-free salient regions to help predict IAA-driven attention in a finer granularity. To further investigate the connection between task-free and task-specific attention, we compare the IAA-driven attention maps with human fixation predictions. Results show task-specific attention in IAA aligns with

human fixation in most cases, while other exceptional cases suggest potential data bias in the training set.

2.2 Image Aesthetic Assessment (IAA)

IAA aims to automatically predict the aesthetic quality of any given image. Early methods adopt low-level hand-crafted features for training IAA models. For example, features for tone, colorfulness, luminance, composition, texture, sharpness, clarity, visual saliency [4, 9, 26, 31, 46] and even generic descriptors such as SIFT [43], Bag-of-Visual-Words and Fisher Vector [24] have been adopted in early works. However, such low-level hand-crafted features are hard to generalize to a large diversity of content.

With the renaissance of deep learning, CNN has been widely used in recent works [10, 20, 22, 30, 32, 35, 48] and achieves promising results. From the perspective of model inputs, most methods use image patches as a part of their model inputs. Particularly, Lu *et al.* [20] proposed a model that takes multiple small patches from one image with shared CNN columns for feature extraction and aggregates them for further prediction. Kong *et al.* [10] trained an AlexNet based model with 227×227 image crops from 256×256 rescaled images. Talebi *et al.* [32] trained VGG16, MobileNet or Inception-v2 based models with 224×224 image crops from 256×256 rescaled images. Though widely used, these methods suffer from two major drawbacks. First, single randomly selected patch usually cannot be a good representation of a holistic image because part of the information is missing. Although multiple patches may be combined as model input, since objectness is not a consideration when image patches are selected, the combination of those patches can still lack semantic integrity. Second, when multiple patches are taken, all patches are assumed to equally contribute to the final aesthetic, in contrast to the human attention mechanism that different regions have different contributions.

To overcome the aforementioned drawbacks, Ma *et al.* [22] designed a heuristic process. It learned both layout information and fine-grained details from those carefully selected patches. Most recently, the attention mechanism has been used to overcome the drawbacks. Zhang *et al.* [48] selected patches of local views according to feed-forward attention learned via IAA. Yang *et al.* [39] weighted CNN outputs with human fixation prediction. Sheng *et al.*'s work [30] adopted attention mechanism to dynamically adjust the weight on each randomly selected patches during training. Compared to previous attention-based methods, our method takes one step further. Our attention is built upon object-level RoIs, which introduces clearer boundaries between different object-level regions, and the attention can be learnt with a finer granularity.

3 OUR APPROACH

Given an image, the proposed framework aims to predict its aesthetic rating distribution (ARD) (Section 3.1). The overall design of the proposed framework is shown in Fig.2. The framework is designed to infer ARD in two separate stages. In the first stage (Section 3.2), global and regional visual features are extracted with a pretrained network from holistic images and region-of-interests (RoI). In the second stage (Section 3.3), a neural network with the attention-based regional feature fusion (ARFF) module predicts ARD from global and regional features.

3.1 Problem Formulation

The task aesthetic rating distribution prediction (ARDP) aims to predict the ARD of a given image. The raw ARD of the i-th image in the training set can be expressed as $c_i = \{c_i^j\}_{j=1}^K$, where c_i^j is the number of votes in the j-th score bucket of the i-th image and K is the number of score buckets. In practice, we use AVA dataset and $K = 10$ in this case. All raw ARD labels are normalized by being divided by the total number of votes in all buckets. Therefore, normalized ARD of the i-th image is given by:

$$p_i = \{p_i^j\}_{j=1}^K = \{c_i^j / \sum_{j=1}^K c_i^j\}_{j=1}^K, \tag{1}$$

then the training set with normalized ARD as labels can be denoted by $\{(\mathbf{I}_i, p_i)\}_{i=1}^N$, where N is the number of images in the training set. If we denote the predicted ARD by $\hat{p}_i$, then the optimization process for learning parameter θ of the neural network model for IAA can be expressed as:

$$\theta = \arg\min_\theta \sum_{i=1}^N \mathcal{L}(p_i, \hat{p}_i) \tag{2}$$

where $\mathcal{L}(\cdot)$ is the loss function. For the loss function, we follow previous works [32] and use the normalized Earth Mover Distance (EMD) loss. The normalized EMD loss is given by:

$$\mathcal{L}(p_i, \hat{p}_i) = \sqrt{\frac{1}{n} \sum_{k=1}^n |CDF_{p_i}(k) - CDF_{\hat{p}_i}(k)|^2}, \tag{3}$$

where CDF_{p_i} and $CDF_{\hat{p}_i}$ are cumulative density function for ground-truth ARD p_i and predicted ARD $\hat{p}_i$ respectively, and n is their length.

3.2 Feature Extraction

In this work, our model is designed to build upon pretrained features for higher computational efficiency. The pretrained feature should satisfy several requirements. First, considering IAA involves both low-level aspects such as image degradations and high-level aspects such as semantic information, the pretrained feature should contain both low-level and high-level information. Second, the high-level components of the pretrained feature should cover a large diversity of content so that the derivative IAA model will not work only on a specific category of content.

Based on aforementioned considerations, we adopt multi-level spatially pooled (MLSP) [6] features. As shown in Fig.3, MLSP features are extracted from InceptionResNet-v2 [2] CNN pretrained on ImageNet. To obtain an MLSP feature, firstly, the output from each convolution block is pooled into a fixed spatial size. Second, the pooled feature maps are sequentially concatenated into an MLSP feature. The original work [6] provided two pooling strategies. The first one resized each feature map into 5×5 as shown in Fig.3(a), while the second one directly applied global average pooling (GAP) to each feature map as shown in Fig.3(b). Results generated with the former and latter strategies are called Wide MLSP feature and Narrow MLSP feature, respectively.

We believe that the MLSP feature satisfies our basic requirements for the following reasons. First, the model is pretrained on

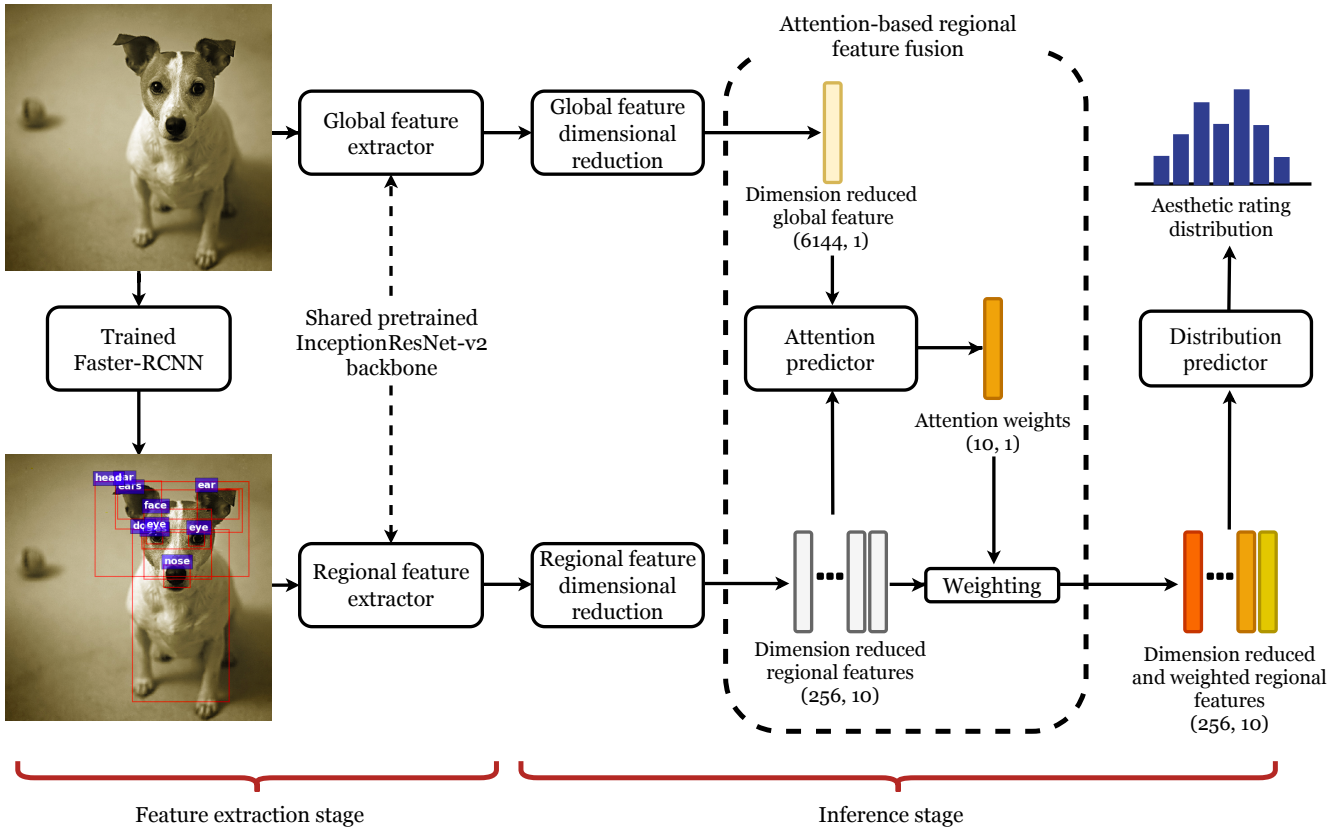


Figure 2: The diagram of the proposed framework. Dimension of features produced by each module is given as (x, y). The model infers ARD in two separate stages. In the first (feature extraction) stage, features are extracted with ImageNet pretrained InceptionResNet-v2 from a full-resolution image and its RoIs detected by Faster-RCNN. In the second (inference) stage, with the help of the attention-based regional feature fusion module, the model predicts the contributions of different RoIs, and regional features are weighted and fused according to the predicted attention weights for inferring ARD.

ImageNet subject to general classification task, which covers a large diversity of content. Second, the features are extracted from original-sized images, so that the integrity of low-level details can be better maintained. Third, the feature combines the output from each convolution block of the backbone model, and therefore the concatenated feature contains both low-level and high-level information. Therefore, our global and regional feature extractors in Fig.2 are designed to extract MLSP features from holistic images and RoIs, respectively.

Suppose the selected regions from image $\mathbf{I}$ is given as $\{\mathbf{I}^l\}_{l=1}^L$, where L is the number of selected regions, the global and regional feature extraction is represented by:

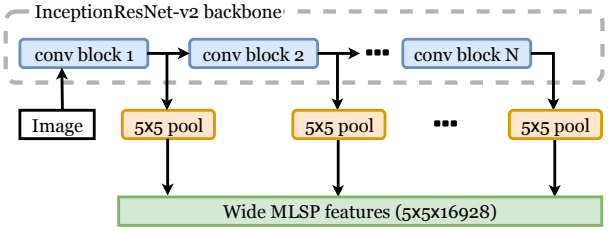
$$\mathbf{v}_g = M_{global}(\mathbf{I}) \tag{4}$$

$$v_r = \{\mathbf{v}_r^l\}_{l=1}^L = M_{regional}(\{\mathbf{I}^l\}_{l=1}^L) \tag{5}$$

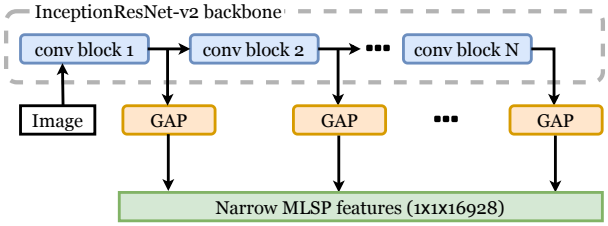
where $\mathbf{v}_g \in R^{D_g}$ represents global feature and $\mathbf{v}_r^l \in R^{D_r}$ represents MLSP feature extracted from the l-th selected region, and D_g and D_r are the length of the global and regional features, respectively.

Global feature extraction. The global features directly extracted from full-resolution images are needed to provide overall information such as layout. Then the dimension of the global feature is reduced for improving computational efficiency. After that, the dimension reduced global feature is merged with dimension reduced regional features for attention prediction as shown in Fig.2. We consider both narrow and wide MLSP settings for global feature extraction, and their performance is discussed in Section 4.2.

Regional feature extraction. Our regional features are MLSP features extracted from RoIs generated by Faster-RCNN [28], as shown in Fig.2. Faster-RCNN is a two-stage object detector, which localizes the instances of objects by bounding boxes in the first stage, and assigns semantic labels to each detected instance in the second stage. The main consideration for choosing the object detector is that it should cover a large diversity of objects, and the granularity of detection should be as fine as possible. Fine granularity is especially important for predicting object-level attentions of close-view images. Take portrait for example, we not only expect the object detector can outline the subject's face, but also his eyes, nose, etc. Although Faster-RCNN trained on MS COCO [16] or Pascal VOC [5] dataset are the most commonly-used versions, both of them still lack granularity of detection that we desired.



(a) Wide MLSP feature extraction



(b) Narrow MLSP feature extraction

Figure 3: Narrow and wide MLSP feature extraction [6].

Table 1: Specification of FCN used in the proposed model

FCN-based module name	Specification
Global feature dimensional reduction	[FC(16928, 6144), ReLU]
Regional feature dimensional reduction	[FC(16928, 256), ReLU]
Attention predictor	[FC(8704, 4096), BN, ReLU] [FC(4096,10), Sigmoid]
Distribution predictor	[FC(2560, 10), Softmax]

Therefore, we adopt the version trained on Visual Genome [11], which covers a much larger object categories than MS COCO [16] or Pascal VOC [5] (76,340 vs. 91 and 20). The effectiveness of this version has been verified on image captioning and visual question answering [1, 41, 42]. Then we apply this Faster-RCNN to all images in AVA dataset. For each image, we select the top 10 region proposals by confidence because we empirically find 10 bounding boxes are enough for covering the majority of the areas of the images (average 82% areas are covered). For the consideration of computational efficiency, we only extract narrow MLSP features from each individual region rather than wide MLSP features.

3.3 Network Architecture

The core part of the network used in the inference stage is the attention-based regional feature fusion (ARFF) module. Taking a global feature and a set of regional features as input, the ARFF module weights and passes down the regional features to the distribution predictor. Given a global feature $\mathbf{v}_g$ and a set of regional features $\{\mathbf{v}_r^l\}_{l=1}^L$, the dimensions of features are first reduced with the global and regional feature dimensional reduction (FDR) modules respectively. When using the wide MLSP setting for global

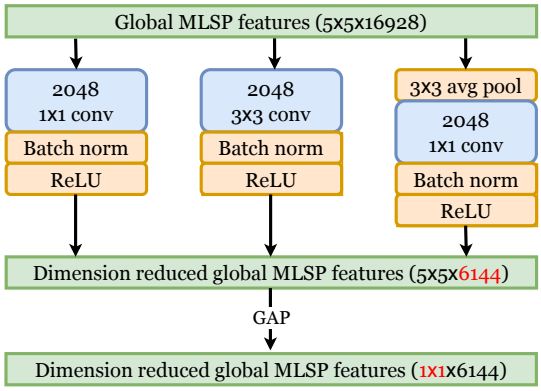


Figure 4: CNN-based global feature dimensional reduction (FDR) module for wide MLSP features.

feature extraction, the architecture of the FDR module is shown in Fig.4. It is constructed with 3 different convolution modules, where each module reduces the channel dimension of wide MLSP features from 16928 to 2048 while maintaining its size of spatial dimension. Therefore, by concatenating the outputs of the 3 convolution modules, the $5 \times 5 \times 16928$ feature is downsized to $5 \times 5 \times 6144$. Then its spatial dimension is further reduced by global average pooling (GAP), which generates the final $1 \times 1 \times 6144$ dimension reduced global feature. When adopting the narrow MLSP setting, the FDR modules for global features or regional features are fully-connected network (FCN) based. The specification of FCN-based FDR modules are shown in Table 1. After dimensional reduction, the size of a global feature is reduced to 6144 and the size of each regional feature is reduced to 256.

Since there are L RoIs for each image in practice, there are L corresponding regional features. Suppose $v = \{\mathbf{v}^l\}_{l=1}^{L+1}$ is the set of dimension reduced global and regional features, where $\mathbf{v}^{L+1} \in \mathbf{R}^{6144}$ is the dimension reduced global feature and the remaining are dimension reduced regional features $\mathbf{v}^l \in \mathbf{R}^{256}$ and $l \in [1, L]$. Since we take $L = 10$ in practice, all of the features are concatenated into one 8704-length feature vector and passed to the attention predictor. The target of the attention predictor $f_{attention}(\cdot)$ is to predict the contribution of each regional feature. Therefore, we implement the attention predictor $f_{attention}(\cdot)$ with an FCN that takes an 8704-length input and produces a 10-length attention vector $\mathbf{a}$. The specification of the FCN-based attention predictor is presented in Table 1. Because the attention predictor adopts sigmoid activation at its output layer, the scale of predicted weights is $(0, 1)$. The inference of the attention weights can be summarized as:

$$\mathbf{a} = f_{attention}(\overset{L+1}{\underset{l=1}{\oplus}} \mathbf{v}^l), \tag{6}$$

where $\oplus$ represents the concatenation operation and $\mathbf{a} \in \mathbf{R}^{\mathbf{L}}$ is the attention vector predicted by the attention predictor $f_{attention}(\cdot)$. Finally, each of the regional feature is weighted by applying corresponding attention weight:

$$\widetilde{v} = \{\widetilde{\mathbf{v}}^l\}_{l=1}^{L} = \{\mathbf{a}_l \cdot \mathbf{v}^l\}_{l=1}^{L}, \tag{7}$$

Model name	Attention predictor	SRCC (mean)	PLCC (mean)	SRCC (std. dev)	PLCC (std. dev)	Accuracy
Model 1	-	0.652	0.654	0.225	0.233	78.10%
Model 2	Regional	0.675	0.678	0.270	0.279	78.98%
Model 3	Regional + narrow global	0.735	0.738	0.338	0.347	81.11%
Model 4	Regional + wide global	**0.751**	**0.753**	**0.353**	**0.363**	**81.67%**

Table 2: Ablation study. Best results are shown in bold face.

Method	SRCC (mean)	PLCC (mean)	SRCC (std. dev)	PLCC (std. dev)	Accuracy
Talebi et al. (TIP 2018) [32]	0.636	0.612	0.233	0.218	81.51 %
Zhang et al. (TMM 2019) [48]	0.690	0.704	-	-	**81.81%**
Hosu et al. (CVPR 2019) [6]	**0.740**	**0.742**	**0.333**	**0.344**	80.97%
Li et al. (TIP 2020) [15]	0.677	-	-	-	**83.70%**
Zeng et al. (TIP 2020) [44]	0.719	0.720	0.241	0.247	80.81%
Ours	**0.751**	**0.753**	**0.353**	**0.363**	81.67%

Table 3: Peer comparison. Top 2 results on each metric are shown in bold face.

where $\widetilde{v}$ is the set of weighted regional features. Finally, the weighted regional features are further concatenated and passed to the FCN-based distribution predictor (Table 1) to infer the final ARD.

4 EXPERIMENT AND ANALYSIS

4.1 Experimental Setup

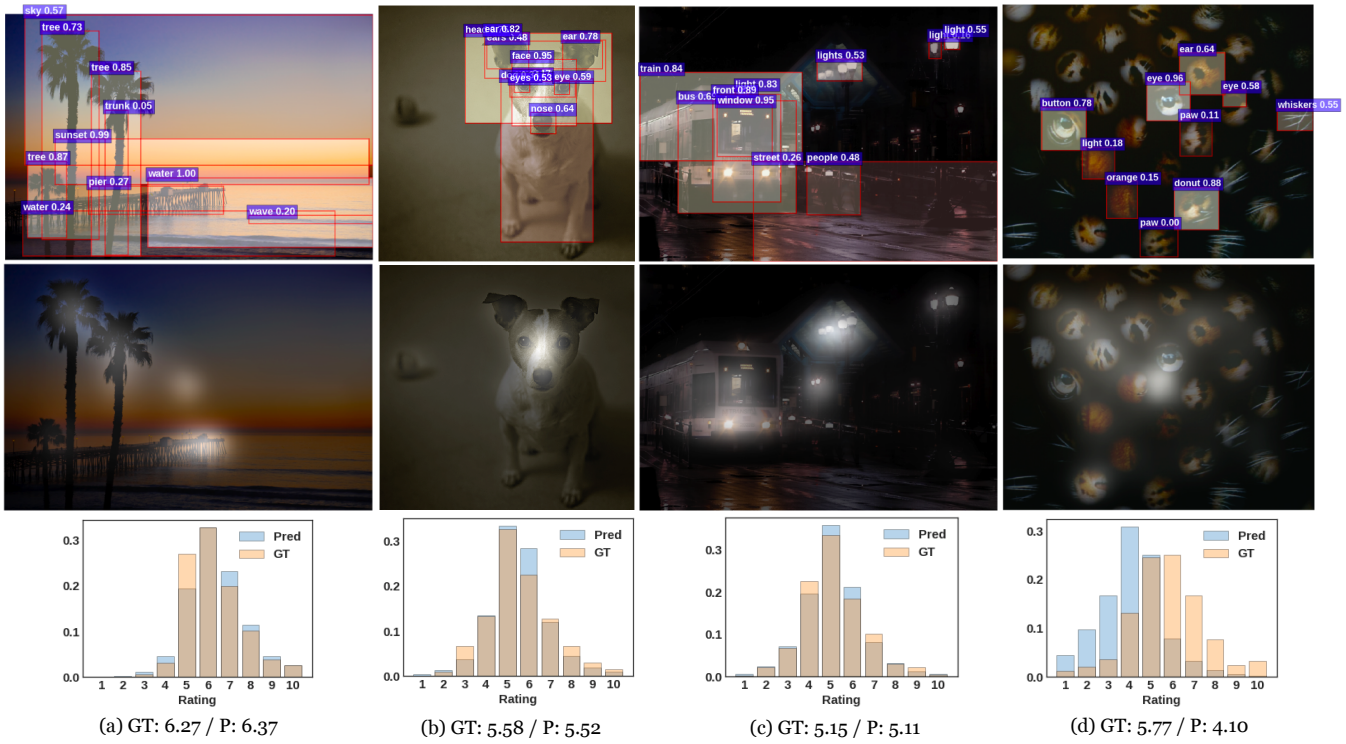
Our experiment is conducted on the official split of AVA dataset as previous works [6, 15, 20, 23, 44, 48]. It consists of $\sim 250k$ images, and officially divided into a training set with $\sim 230k$ and a testing set with $\sim 20k$ images. It provides ground truth ratings for each image in forms of raw rating distribution and average scores on a scale of $1 \sim 10$. Since some images are not available, there are 235,574 images for training and 19,928 images for testing in practice. The evaluation on ARDP task follows [32]. Ground truth ARD and predicted ARD results are converted to average scores and standard deviations. Given a normalized ARD $\{p^j\}_{j=1}^{10}$, the average score is computed as $\mu = \sum_{j=1}^{10} j \cdot p^j$ and the standard deviation is computed as $\sigma = \sqrt{\sum_{j=1}^{10} (j-\mu)^2 \cdot p^j}$. We adopt two commonly-used metrics, Spearman rank order coefficient (SRCC) and Pearson linear correlation coefficient (PLCC), for evaluating the goodness of fitting of average scores and standard deviations. We also convert ARD to binary labels with 5 as the cut-off threshold as previous works and then accuracy is computed.

4.2 Ablation Study

We first conduct an ablation study to investigate the effectiveness of the attention-based regional feature fusion (ARFF) module in different settings. We set up 4 different models and the detailed settings of the evaluated model are listed below:

- **Model 1**: Model 1 is the baseline model without the ARFF module. Taking regional features as model inputs, the model directly concatenates all dimension reduced regional features and finally passes them for ARD prediction. Therefore, all regional features are treated equally in this setting.
- **Model 2**: Compared to Model 1, Model 2 also only takes regional features as model input, while ARFF module is added to model attention weights for weighting regional features. Specifically, the attention predictor learns attention weights from dimension reduced and concatenated regional features.
- **Model 3**: Compared to Model 2, in addition to regional features, Model 3 also takes global features as inputs, while it adopts the narrow MLSP setting as mentioned in Sec.3.2. The attention predictor learns attention weights from both dimension reduced global and regional features for weighting regional features.
- **Model 4**: Compared to Model 3, Model 4 adopts the wide MLSP setting for global feature extraction instead of taking the narrow MLSP setting. The dimension of global features is reduced by the CNN-based global FDR module.

All models are trained on the official training set for 10 epochs with Adam optimizer and then evaluated on the testing set. The learning rate is set to 3e-5 for the first 2 epochs, and divided by 10 every 3 epochs. The results of the ablation study are shown in Table 2. We can observe all measurements are improved as the model setting is upgraded. Treating Model 1 as the baseline, the last three models with the ARFF module achieve higher performance in all selected measurements. This confirms that the proposed attention module can effectively improve the representational power of the deep IAA model by weighting object-level regional features in a learnable process. By comparing Model 2 with the last two models with global features for attention prediction, we can observe a prominent performance gain when global features are cooperated. This indicates global features can provide valuable information for guiding the attention prediction. We believe this phenomenon also aligns with the nature of human vision that it is hard for humans to judge which region is more attractive without looking at the overall structure and layout of the image. By comparing the last two models, we can see adopting the wide MLSP setting can further improve the performance, which demonstrates spatially pooling MLSP features directly with global average pooling can lose necessary information and leaving more spatial information to the learning process can effectively increase the model performance.



(a) GT: 6.27 / P: 6.37 (b) GT: 5.58 / P: 5.52 (c) GT: 5.15 / P: 5.11 (d) GT: 5.77 / P: 4.10

Figure 5: Three representative successful cases ((a)-(c)) and a typical failed case ((d)). Row 1: region-of-interests with projected IAA-driven object-level attention; numbers in the labels are attention weights, and higher brightness indicates a higher attention level. Row 2: task-free pixel-level fixation maps which also use higher brightness for a higher attention level. Row 3: histograms for ground-truth (GT) and predicted (Pred) aesthetic rating distribution and the corresponding average scores.

4.3 Peer Comparison

We have also compared our model with previous relevant works. We choose the results of Model 4 from ablation studies to compare with 5 recent relevant works. For the 4 methods by Talebi *et al.* [32], Zhang *et al.* [48], Li *et al.* [15] and Zeng *et al.* [44], we refer to the results from their original papers. For the work by Hosu *et al.* [6], since the original work is trained subject to score regression, we have modified the model by replacing its output layer with a 10-way softmax layer and re-trained it with normalized EMD loss (Eq.(3)) on the ARDP task with the AVA trainset. The evaluation results on the AVA testset of the altered version is reported. All results are presented in Table 3 and the top 2 results on each measurement are shown in **bold face**.

As shown, we can observe our model outperforms all selected methods in terms of SRCC and PLCC of average scores and standard deviations. This indicates our model is superior to other selected models in ARDP. To be specific, compared to Zhang *et al.*'s model [48] which adopts pixel-level feed-forward attention for selecting the most important regions, our model is superior in terms of SRCC and PLCC of average scores. We interpret the superiority of our model by the effectiveness of object-level attention. With the help of object-level attention, the model can learn attention with finer granularity, and semantic integrity of selected regions are better preserved. We also notice that our accuracy is not the best among all presented methods. We argue that accuracy is not a suitable metric for the ARDP task, because the binary classification results are very sensitive to predictions around the cut-off threshold, while insensitive to predictions far away from the threshold. For example, the scores 5.01 and 9 can both fall into the high aesthetic class when we set 5 as the cut-off threshold although these are two distinct scores, while the scores 5.01 and 4.99 can fall into two different classes although these two scores are very close.

4.4 Further Analysis

We have chosen three representative successful cases for demonstrating the effectiveness of our framework, including a long shot (Fig.5(a)), a close-up shot (Fig.5(b)), and a medium shot (Fig.5(c)). We also present a typical failed case of our framework (Fig.5(d)). For each example, we present its object-level attention map (row 1 in Fig.5), fixation map (row 2 in Fig.5) and histograms of predicted ARD (row 3 in Fig.5). For object-level attention maps, the number in each bounding box label reflects the attention weight of the corresponding region. The attention weight is on a scale of $0 \sim 1$, and a higher value means stronger attention and higher contribution to the final ARD prediction. The attention weights are projected to the corresponding bounded regions in ascending order so that low-attention regions will be covered by high-attention regions when there is any overlapping. The fixation maps are generated with Yang *et al.*'s work [40].

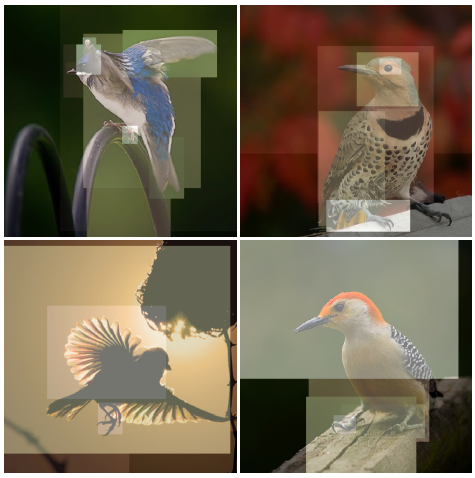


Figure 6: IAA-driven object-level attention maps (labels are omitted) that do not fully align with human attention. Feets of all the birds are highly attentive to the learned model, which implies a potential data bias.

Fig.5(a) shows a long shot taken in the distance for a beautiful scene of seaside in the sunset. Long shots such as landscape or cityscape typically have unbounded regions like sky and water. As shown, the object detector recognizes and outlines the components of the scene, including trees, pier, sky, sunset, water, and wave. This implies although some objects do not have a clear boundary (sunset, sky, water and wave), the object detector can still roughly outlines the region. And the objects inside those regions such as trees and the pier can be further detected. Comparing to the fixation prediction, our model can predict the attention weight of each individual object and learn attentive regions similar to the fixation map (trees, sunset, and water around the pier receive higher attention).

Fig.5(b) shows a close-up shot of a dog. For close-up shots like portrait, the photographer usually wants to show a close look at the subject's facial features and each has a different contribution to the overall aesthetics. As the example shows, our model can account for the aesthetic contribution of each individual facial feature, whereas the fixation prediction can only roughly highlight the face region.

Medium shots are very similar to the sight of the human and are widely used for street photographs. One common characteristic of street shots is that they are usually taken without any plan, because things on the street are constantly changing. This means street photographers merely have a chance to carefully control the shutter speed and the use of light, and therefore distortions like under-exposure and motion blur are commonly seen in street shots. Fig.5(c) presents a typical example of a medium shot of street. As shown, the photograph is obviously under-exposed and things on the streets can be hardly seen. Nevertheless, our framework can still recognize the pedestrians and the lights in the dark and predict corresponding attention weights. The resulting attentive regions are similar to the fixation map, while our model can analyze the attention in finer granularity.

Although the chosen object detector works well in the above cases, it cannot work so well when the subject is extremely abstract or obscured. Fig.5(d) shows a puma whose face is obscured by the cage. In this case, the object detector in use fails to recognize the overall face of the puma, and only detects its eyes and whiskers. Since only partial information is considered for the ARDP, the result is not accurate. Thus, we believe development of a better object detector is important, but it is beyond the scope of this research. However, such failed cases are rare in the AVA dataset. There are 17,848 (out of 255,502; <7%) images labelled as abstract ones by AVA dataset. Dealing with abstract images, the chosen object detector can still define objects by their shapes or forms in most cases, e.g. line, circle, dot, rim, light, shadow, reflection, etc. We have evaluated the proportion of area covered by bboxes for each image to discover those potentially 'failed detected cases'. If we take 0.3 as the threshold for determining a failed detected case (i.e., less then 30% of area is covered by bboxes), 245 such cases are founded in abstract images.

Although in most cases the predicted IAA-driven object-level attention roughly aligns with task-free fixation, we do observe some cases where IAA-driven attention is paid to some minor regions. As shown in Fig.6, one of the phenomena is that the model tends to pay too much attention to the birds' feet. We interpret this phenomenon by data bias. We observe that AVA trainset contains a large amount of close-up shots of birds (roughly 2,317 images) and most of them are of high aesthetics (1,885 out of 2,317 images of birds have average scores larger than 5). The model tends to memorize that close-up shots of birds are of high aesthetics. Thus, when making predictions of high aesthetic images, the model tends to find whether an image is a close-up shot of a bird. Since birds can be easily recognized by their feet, the model then tends to focus on the birds' feet. Therefore, the model learns the data bias, and such bias is reflected on the IAA-driven object-level attention maps.

5 CONCLUSION

In this work, we have proposed a neuropsychologically-inspired object-level attention-based framework for aesthetic rating distribution prediction (ARDP). The proposed framework dynamically learns contribution of features extracted from object-level regions defined by a generic object detector. This allows our framework to have more flexibility in attention modeling in the cases when an attentive region is composed of several constituent objects. Extensive experimental analysis over the most commonly-used AVA dataset demonstrates our model is superior to previous relevant methods and ablation study confirms the effectiveness of the use of object-level attention. Qualitative analyses comparing IAA-driven object-level attention with task-free pixel-level fixation shows IAA-driven object-level attention roughly aligns with task-free pixel-level fixation, while the former one can model attention with a finer granularity. Apart from its effectiveness, IAA-driven object-level attention can also serve as a tool for interpreting our deep IAA model. To the best of our knowledge, this is the first attempt to model object-level attention for IAA and we believe this opens a new research direction on object-level attention based IAA model.

ACKNOWLEDGMENTS

This work is supported in part by Singapore Ministry of Education Tier-2 Fund MOE2016-T2-2-057(S).

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[2] Federico Baldassarre, Diego González Morín, and Lucas Rodés-Guirao. 2017. Deep koalarization: Image colorization using cnns and inception-resnet-v2. *arXiv preprint arXiv:1712.03400* (2017).

[3] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5659–5667.

[4] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *European conference on computer vision*. Springer, 288–301.

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[6] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. 2019. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9375–9383.

[7] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[8] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 262–270.

[9] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, Vol. 1. IEEE, 419–426.

[10] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *European Conference on Computer Vision*. Springer, 662–679.

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[12] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. 2017. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing* 26, 9 (2017), 4446–4456.

[13] Jason Kuen, Zhenhua Wang, and Gang Wang. 2016. Recurrent attentional networks for saliency detection. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*. 3668–3677.

[14] Congcong Li, Alexander C Loui, and Tsuhan Chen. 2010. Towards aesthetics: A photo quality assessment and photo selection system. In *Proceedings of the 18th ACM international conference on Multimedia*. 827–830.

[15] Leida Li, Hancheng Zhu, Sicheng Zhao, Guiguang Ding, and Weisi Lin. 2020. Personality-Assisted Multi-Task Learning for Generic and Personalized Image Aesthetics Assessment. *IEEE Transactions on Image Processing* 29 (2020), 3898–3910.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[17] Nian Liu, Junwei Han, Tianming Liu, and Xuelong Li. 2016. Learning to predict eye fixations via multiresolution convolutional neural networks. *IEEE transactions on neural networks and learning systems* 29, 2 (2016), 392–404.

[18] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 375–383.

[19] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. 2014. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*. 457–466.

[20] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang. 2015. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 990–998.

[21] Shuang Ma, Yangyu Fan, and Chang Wen Chen. 2014. Pose Maker: A Pose Recommendation System for Person in the Landscape Photographing. In *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, Florida, USA) *(MM '14)*. Association for Computing Machinery, New York, NY, USA, 1053–1056. https://doi.org/10.1145/2647868.2655053

[22] Shuang Ma, Jing Liu, and Chang Wen Chen. 2017. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4535–4544.

[23] Long Mai, Hailin Jin, and Feng Liu. 2016. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 497–506.

[24] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. 2011. Assessing the aesthetic quality of photographs using generic image descriptors. In *2011 international conference on computer vision*. IEEE, 1784–1791.

[25] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2408–2415.

[26] Masashi Nishiyama, Takahiro Okabe, Imari Sato, and Yoichi Sato. 2011. Aesthetic quality classification of photographs based on color harmony. In *CVPR 2011*. IEEE, 33–40.

[27] Pere Obrador, Xavier Anguera, Rodrigo de Oliveira, and Nuria Oliver. 2009. The role of tags and image aesthetics in social image search. In *Proceedings of the first SIGMM workshop on Social media*. 65–72.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[29] Brian J Scholl. 2001. Objects and attention: The state of the art. *Cognition* 80, 1-2 (2001), 1–46.

[30] Kekai Sheng, Weiming Dong, Chongyang Ma, Xing Mei, Feiyue Huang, and Bao-Gang Hu. 2018. Attention-based multi-patch aggregation for image aesthetic assessment. In *Proceedings of the 26th ACM international conference on Multimedia*. 879–886.

[31] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Shaohui Liu. 2009. Photo assessment based on computational visual attention model. In *Proceedings of the 17th ACM international conference on Multimedia*. 541–544.

[32] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[33] Wenguan Wang and Jianbing Shen. 2017. Deep visual attention prediction. *IEEE Transactions on Image Processing* 27, 5 (2017), 2368–2378.

[34] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. 2019. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5968–5977.

[35] Wenguan Wang, Jianbing Shen, and Haibin Ling. 2018. A deep network solution for attention and aesthetics aware photo cropping. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1531–1544.

[36] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. 2019. Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1448–1457.

[37] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19.

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[39] Sheng Yang, Qiuping Jiang, Weisi Lin, and Yongtao Wang. 2019. SGDNet: An End-to-End Saliency-Guided Deep Neural Network for No-Reference Image Quality Assessment. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1383–1391.

[40] Sheng Yang, Guosheng Lin, Qiuping Jiang, and Weisi Lin. 2019. A dilated inception network for visual saliency prediction. *IEEE Transactions on Multimedia* (2019).

[41] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[42] Xu Yang, Hanwang Zhang, and Jianfei Cai. 2020. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923* (2020).

[43] Mei-Chen Yeh and Yu-Chen Cheng. 2012. Relative features for photo quality assessment. In *2012 19th IEEE International Conference on Image Processing*. IEEE, 2861–2864.

[44] Hui Zeng, Zisheng Cao, Lei Zhang, and Alan C Bovik. 2019. A Unified Probabilistic Formulation of Image Aesthetic Assessment. *IEEE Transactions on Image Processing* 29 (2019), 1548–1561.

[45] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. 2018. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1741–1750.

[46] Luming Zhang, Yue Gao, Roger Zimmermann, Qi Tian, and Xuelong Li. 2014. Fusion of multichannel local and global structural cues for photo aesthetics evaluation. *IEEE Transactions on Image Processing* 23, 3 (2014), 1419–1429.

[47] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. 2017. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE International Conference on computer vision*. 212–221.

[48] Xiaodan Zhang, Xinbo Gao, Wen Lu, and Lihuo He. 2019. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. *IEEE Transactions on Multimedia* 21, 11 (2019), 2815–2826.