Attention-based Multi-Patch Aggregation for Image Aesthetic Assessment

Kekai Sheng
NLPR, Institute of Automation,
Chinese Academy of Sciences &
University of Chinese Academy of Sciences
shengkekai2014@ia.ac.cn

Weiming Dong*
NLPR, Institute of Automation,
Chinese Academy of Sciences
weiming.dong@ia.ac.cn

Chongyang Ma
Snap Inc.
cma@snap.com

Xing Mei
Snap Inc.
xing.mei@snap.com

Feiyue Huang
Youtu Lab, Tencent
garyhuang@tencent.com

Bao-Gang Hu
NLPR, Institute of Automation,
Chinese Academy of Sciences
hubg@nlpr.ia.ac.cn

ABSTRACT

Aggregation structures with explicit information, such as image attributes and scene semantics, are effective and popular for intelligent systems for assessing aesthetics of visual data. However, useful information may not be available due to the high cost of manual annotation and expert design. In this paper, we present a novel multi-patch (MP) aggregation method for image aesthetic assessment. Different from state-of-the-art methods, which augment an MP aggregation network with various visual attributes, we train the model in an end-to-end manner with aesthetic labels only (i.e., aesthetically positive or negative). We achieve the goal by resorting to an attention-based mechanism that adaptively adjusts the weight of each patch during the training process to improve learning efficiency. In addition, we propose a set of objectives with three typical attention mechanisms (i.e., average, minimum, and adaptive) and evaluate their effectiveness on the Aesthetic Visual Analysis (AVA) benchmark. Numerical results show that our approach outperforms existing methods by a large margin. We further verify the effectiveness of the proposed attention-based objectives via ablation studies and shed light on the design of aesthetic assessment systems.

CCS CONCEPTS

• **Computing methodologies → Computational photography**; *Neural networks*;

KEYWORDS

Image aesthetic assessment, attention mechanism, multi-patch aggregation, convolutional neural network

ACM Reference Format:
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 *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3240508.3240554

1 INTRODUCTION

As the volume of visual data grows exponentially each year, the capability of assessing image aesthetics becomes crucial for various applications such as photo enhancement, image stream ranking, and album thumbnail composition [1, 3, 4]. The aesthetic assessment process of the human visual system involves numerous factors such as lighting, contrast, composition, and texture [7, 26]. Some of these factors belong to holistic scene information, whereas others are fine-grained image details. Designing an artificial intelligent system that accommodates all these factors is a challenging task.

Many studies have focused on this problem in the last decade. Although some early methods only consider global factors [5, 13, 22], most recent approaches propose combining holistic scene information and fine-grained details in a multi-patch (MP) aggregation network [6, 12, 23, 34]. A common practice is to leverage explicit information, such as image attributes [34, 36], scene semantics [19], and intrinsic components [6, 23] in the network design. Using explicit information encodes various complementary visual cues and can significantly outperform alternative methods that only rely on aesthetic labels [22]. However, explicit information might not be always available due to the high cost of manual annotation and the expert knowledge required for feature design. As images with aesthetic labels become available at a large scale from online photography communities, we revisit the problem of image aesthetic assessment and explore how to learn an effective aesthetic-aware model in an end-to-end manner with aesthetic labels only.

Learning with aesthetic labels only is challenging because the labels may not provide sufficient signals for training and can lead to poor assessment results. In the absence of explicit information, deciding which image patches are useful in making the correct prediction is difficult. Therefore, image patches are usually considered equally important in the training stage and the inference time for

[*] Corresponding author

MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240554

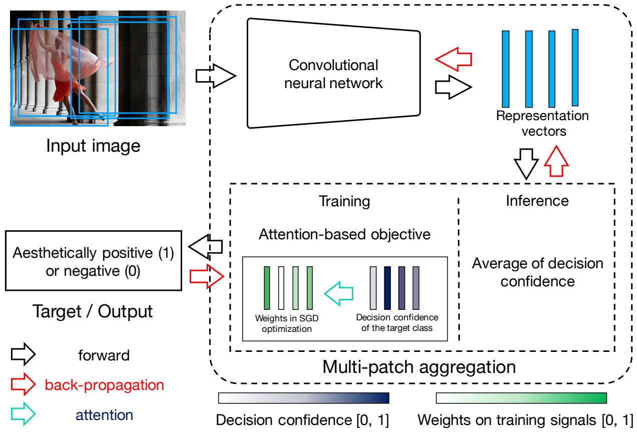
Input image
Convolutional neural network
Representation vectors
Training
Inference
Attention-based objective
Average of decision confidence
Weights in SGD optimization
Decision confidence of the target class
Aesthetically positive (1) or negative (0)
Target / Output
forward
back-propagation
attention
Multi-patch aggregation
Decision confidence [0, 1]
Weights on training signals [0, 1]

Figure 1: System overview. We use an attention-based objective to enhance training signals by assigning relatively larger weights to misclassified image patches.

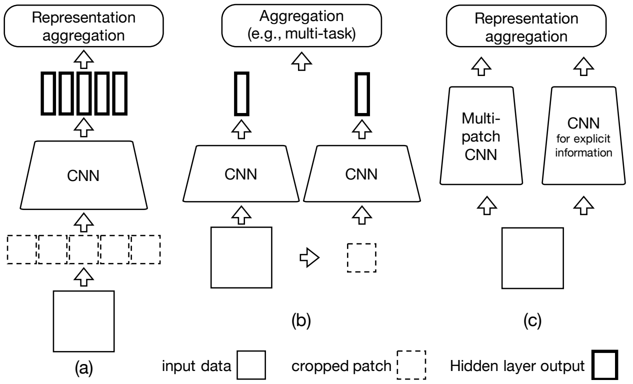
Representation aggregation
Aggregation (e.g., multi-task)
Representation aggregation
CNN
CNN
CNN
Multi-patch CNN
CNN for explicit information
(a)
(b)
(c)
input data
cropped patch
Hidden layer output

Figure 2: Typical aggregation-based architectures for image aesthetic assessment: (a) MP aggregation; (b) multi-column aggregation; (c) aggregation with explicit information.

most previous methods. Recently, Ma et al. [23] proposed an effective patch selection module to select useful patches heuristically during the training stage and showed that patch selection alone improved the aesthetic assessment performance by 7%. However, their heuristic patch selection was indirectly learned from the training data and might not fully leverage meaningful information from aesthetic labels. We notice that the scheme of patch selection shares an idea similar with attention mechanisms [2, 10, 30], in which human visual attention is not distributed evenly within an image. Recent successes in visual analysis [18, 33, 35] have demonstrated that a well designed attention-based module can significantly improve the performance of a learning system.

Motivated by patch selection and attention mechanism, we propose a simple yet effective solution for image aesthetic assessment, as shown in Figure 1. The key ingredient is an attention-based objective that strengthens training signals by assigning large weights to patches on which the current model has made incorrect predictions. In this manner, we improve the learning efficiency and eventually achieve better assessment results compared with existing approaches that consider each patch with equal weight.

For comparison purpose, we present and evaluate three typical attention mechanisms (i.e., *average*, *minimum*, and *adaptive*). In comparison with the heuristic patch selection scheme [23], our method simplifies the design of the network architecture, and more importantly, enables an end-to-end way to train an assessment model with aesthetic labels only. To the best of our knowledge, attention mechanism has not been explored for image aesthetic assessment. Our quantitative results demonstrate that our proposed solution outperforms state-of-the-art methods on the large-scale Aesthetic Visual Analysis (AVA) benchmark [25]. We also conduct ablation studies and provide additional visualizations to analyze our learned models.

2 RELATED WORK

The estimation of image styles, aesthetics, and quality has been actively investigated over the past few decades. Early studies started from distinguishing snapshots from professional photographs by modeling well-established photographic rules based on low-level handcrafted features [5, 13, 16, 32]. Recently, machine learning based approaches have been successfully applied to various computer vision tasks [9, 18, 33]. Deep learning methods, such as deep convolutional neural network (CNN) and deep belief network, have been successfully applied to photo aesthetic assessment task with promising results. In Figure 2, we divide recent deep learning based aesthetic assessment methods into three categories based on different aggregation structures.

MP aggregation (Figure 2a) concatenates vector representations extracted from multiple patches of the input image for aesthetic assessment. Typical examples include deep multi-patch aggregation network (DMA-Net) [22], multi-net adaptive spatial pooling CNN (MNA-CNN) [19], and MP subnet with an effective patch selection scheme (New-MP-Net) [23].

Multi-column aggregation (Figure 2b) focuses on boosting training signals of aesthetic modeling with additional task-related explicit information, such as multi-column for various attribute modeling [15, 36], brain-inspired deep networks (BDN) [34], two-column CNN for rating pictorial aesthetics (RAPID) [21], aesthetic-attention net (AA-Net) [35], multi-task CNN (MTCNN) with semantic prediction [11], aesthetic quality regression with simultaneous image categorization (A&C CNN) [12], and two-column deep aesthetic net (DAN) with triplet pre-training and category prediction [6].

Representation aggregation with explicit information (Figure 2c) is built on an MP aggregation module and uses explicit information as complementary visual cues for good results. Common examples of explicit information include object instances (e.g., DMA-Net with an object-oriented model [22]), scene semantic (e.g., MNA-CNN with scene-aware aggregation [19]), and expert-designed photographic attributes (e.g., depth of field and color harmonization [23]).

Although explicit information provides meaningful cues for image aesthetic assessment, useful information may not always be available due to the high cost of manual annotation and expert knowledge in design. Therefore, compared with methods having explicit information, training a CNN with aesthetic labels alone is useful if it achieves similar or even better assessment results.

3 APPROACH

3.1 Problem Statement

We denote N pairs of input image I and its corresponding ground-truth aesthetic label $\hat{y}$ as our dataset $\{I_i, \hat{y}_i\}_{i=1}^{N}$. Here, $\hat{y}_i = 1$ means that the image I_i is aesthetically positive, whereas $\hat{y}_i = 0$ denotes an aesthetically negative image. Given the dataset, the problem of learning image aesthetic assessment can be formulated as follows:

$$\underset{\theta}{argmax} \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} Pr(\tilde{y} = \hat{y} \,|\, p, \theta) \tag{1}$$

where $\mathcal{P}$ is a set of square patches $\{p\}$ cropped from images in the dataset, $\tilde{y}$ denotes the predicted aesthetic label, and θ refers to all the network parameters that need to be learned for image aesthetic assessment task. In Equation 1, $Pr(\tilde{y} = \hat{y} \,|\, p, \theta)$ represents the probability that the patch is correctly predicted as the ground-truth label and is computed as the output of the last $softmax$ layer in our network.

Directly optimizing Equation 1 is computationally expensive and might lead to unwanted artifacts, such as overfitting, especially when only aesthetic labels are available. This issue can be linked to the fact that a large batch size usually introduces detrimental effects (e.g., sharp local minima) during the training process via mini-batch stochastic gradient descent (SGD) [14]. Therefore, it is desirable to design an objective function for efficient and effective learning of aesthetic-aware image representations.

3.2 Attention-based Objective Functions

Inspired by patch selection [23] and attention mechanisms [2, 10, 30], we propose assigning different weights to different image patches for effective learning of the aesthetic assessment model. For comparison purpose, we propose three different MP weight assignment schemes, namely, MP_{avg}, MP_{min}, and MP_{ada}.

MP_{avg} scheme. Recall the Jensen's inequality that: given a real-valued concave function f and a set of points $\{x\}$ in a domain S, Jensen's inequality can be stated as follows:

$$f\Big(\frac{1}{|S|} \sum_{x \in S} x\Big) \geq \frac{1}{|S|} \sum_{x \in S} f(x) \tag{2}$$

where the equality holds if and only if $x_i = x_j (\forall x_i \in S)$ or f is linear. On the basis of Jensen's inequality, MP_{avg} can be proposed as an efficient relaxation of the original objective in Equation 1, as shown below:

$$\log\Big(\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} Pr(\tilde{y} = \hat{y} \,|\, p, \theta)\Big) \geq \underbrace{\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \log\Big(Pr(\tilde{y} = \hat{y} \,|\, p, \theta)\Big)}_{MP_{avg}} \tag{3}$$

$$\frac{\partial MP_{avg}}{\partial \theta} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \underbrace{\frac{1}{Pr(\tilde{y} = \hat{y} \,|\, p, \theta)}}_{weights} \cdot \frac{\partial Pr(\tilde{y} = \hat{y} \,|\, p, \theta)}{\partial \theta} \tag{4}$$

If we sample one patch from each image, the MP_{avg} scheme will share a training pipeline similar to the training of a common image classification model. Therefore, this scheme can be trained efficiently compared with the existing MP aggregation models.

MP_{min} scheme. In many machine learning algorithms, another typical attention mechanism is to focus on improving results at data points with moderate confidences, such as hinge loss and hard example mining [20, 28]. Inspired by these prior methods, we propose the MP_{min} scheme as another relaxation of Equation 1:

$$\begin{aligned} \log\Big(\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} Pr(\tilde{y} = \hat{y} \,|\, p, \theta)\Big) &\geq \min_{p \in \mathcal{P}} \frac{1}{|\mathcal{P}|} \log\Big(Pr(\tilde{y} = \hat{y} \,|\, p, \theta)\Big) \\ &= \underbrace{\frac{1}{|\mathcal{P}|} \log\Big(Pr(\tilde{y} = \hat{y} \,|\, p^m, \theta)\Big)}_{MP_{min}} \end{aligned} \tag{5}$$

where

$$p^m = \underset{p \in \mathcal{P}}{argmin}\, Pr(\tilde{y} = \hat{y} \,|\, p, \theta)$$

$$\begin{aligned} \frac{\partial MP_{min}}{\partial \theta} &= \frac{1}{|\mathcal{P}|} \frac{1}{Pr(\tilde{y} = \hat{y} \,|\, p^m, \theta)} \cdot \frac{\partial Pr(\tilde{y} = \hat{y} \,|\, p^m, \theta)}{\partial \theta} \\ &= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{\mathbb{I}(p = p^m)}{Pr(\tilde{y} = \hat{y} \,|\, p, \theta)} \cdot \frac{\partial Pr(\tilde{y} = \hat{y} \,|\, p, \theta)}{\partial \theta} \end{aligned} \tag{6}$$

where $\mathbb{I}(\cdot)$ equals to 1 if p is p^m, and 0 otherwise. As shown in Equation 5, the MP_{min} scheme only considers image patches with the lowest prediction confidence to search for meaningful visual cues, while ignoring other patches from the same image. In practice, we implement a softer version of MP_{min} to avoid a potentially unstable training process. Specifically, the possibility that p is selected in the SGD process is proportional to $1 - Pr(\tilde{y} = \hat{y} \,|\, p, \theta)$.

MP_{ada} scheme. To take advantage of patch selection in the training stage in an end-to-end manner, we design MP_{ada} to assign adaptively larger weights to *meaningful* training instances, i.e., patches on which the current model predicts incorrect aesthetic labels, as shown as follows:

$$\begin{aligned} MP_{ada} &= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \omega_\beta \cdot \log\Big(Pr(\tilde{y} = \hat{y} \,|\, p, \theta)\Big) \\ \omega_\beta &= \frac{Pr(\tilde{y} = \hat{y} \,|\, p, \theta)^{-\beta} - 1}{Pr(\tilde{y} = \hat{y} \,|\, p, \theta)^{-\beta}} = 1 - Pr(\tilde{y} = \hat{y} \,|\, p, \theta)^{\beta} \end{aligned} \tag{7}$$

$$\begin{aligned} \frac{\partial MP_{ada}}{\partial \theta} &= \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \lambda \cdot \frac{\partial Pr(\tilde{y} = \hat{y} \,|\, p, \theta)}{\partial \theta} \\ \lambda &= \frac{1 - (1 + \beta \cdot \log Pr(\tilde{y} = \hat{y} \,|\, p, \theta)) \cdot (1 - \omega_\beta)}{Pr(\tilde{y} = \hat{y} \,|\, p, \theta)} \end{aligned} \tag{8}$$

where β is a positive number to control the adaptiveness of weight assignment. Given that $Pr(\tilde{y} = \hat{y} \,|\, p, \theta)^{-\beta}$ is in the range of $(1, +\infty)$, we normalize its value by subtracting 1 and dividing the result by $Pr(\tilde{y} = \hat{y} \,|\, p, \theta)^{-\beta}$. Figure 3 shows how the curves of ω_β change as a function of $Pr(\tilde{y} = \hat{y} \,|\, p, \theta)$ with different values of β. Intuitively, as the optimization progresses, the ratio of instances of high decision confidence increases. Consequently, the overall loss decreases, and meaningful training signals become weaker. To maintain meaningful training signals, we should allocate more computational resources to data points of relatively lower prediction confidence.

Different from the hinge loss which ignores data points that have been classified correctly, the MP_{ada} scheme constantly assigns

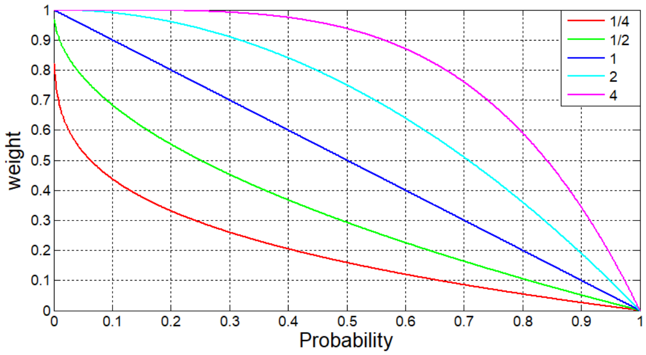

Figure 3: Curves of adaptive weights $\omega_\beta = 1 - Pr^\beta$ with different values of the hyperparameter β.

certain positive weights to those patches to help maintain correct predictions, similar to focal loss [17]. Moreover, this scheme prevents the training process from becoming unstable due to potential noisy data points or outliers.

The task of aesthetic assessment can be considered as a two-class classification problem, and the classifier may suffer from accuracy paradox if the dataset is unbalanced. In the training partition of AVA benchmark [25], more than 75% photos are labeled as aesthetically positive, whereas the rest are negative ones. To achieve a high accuracy, an assessment model may tend to predict a photo to be aesthetically positive. The MP_{ada} scheme can resolve this issue effectively because it assigns large weights ω_β to misclassified data points, which will essentially bias the model to pay attention to the minority class.

3.3 Network Architecture

In addition to the objective function, the network architecture is another indispensable part of an effective machine learning system. Among the popular architectures of CNNs, ResNet [9] is a good choice for our task because of its computational efficiency for training and inference processes.

Table 1 shows the details of our network architecture for all the experiments. For a fair comparison, we adopt the 18-layer ResNet architecture (plus the last classification layer), with the same depth of layers as models based on VGG16 nets [29], which are commonly used in previous methods for aesthetic assessment [19, 22, 23].

4 EXPERIMENTAL RESULTS

In this section, we describe our implementation details and present the experiment results. We compare our approach with state-of-the-art methods on the AVA benchmark dataset and validate the proposed solution via ablation study.

4.1 Training Data

We conduct experiments based on the AVA benchmark [25], which is the largest publicly available dataset for image aesthetic assessment. The AVA benchmark contains about 250,000 photos in total. Each photo has an aesthetic score, which is obtained by averaging the ratings from about 200 people. The scores range from 1 to 10, where 10 indicates the highest aesthetic quality.

Layers	Output names	Output shape
conv, 7x7, 64, stride 2	-	[112, 112, 64]
max pool, 3x3, stride 2	-	[56, 56, 64]
$\begin{bmatrix} conv, 3x3, 64 \\ conv, 3x3, 64 \end{bmatrix} x2$	0-0-BNReLU1 0-0-BNReLU2 0-0-ReLU 0-1-BNReLU1 0-1-BNReLU2 0-1-ReLU	[56, 56, 64]
$\begin{bmatrix} conv, 3x3, 128 \\ conv, 3x3, 128 \end{bmatrix} x2$	1-0-BNReLU1 1-0-BNReLU2 1-0-ReLU 1-1-BNReLU1 1-1-BNReLU2 1-1-ReLU	[28, 28, 128]
$\begin{bmatrix} conv, 3x3, 256 \\ conv, 3x3, 256 \end{bmatrix} x2$	2-0-BNReLU1 2-0-BNReLU2 2-0-ReLU 2-1-BNReLU1 2-1-BNReLU2 2-1-ReLU	[14, 14, 256]
$\begin{bmatrix} conv, 3x3, 512 \\ conv, 3x3, 512 \end{bmatrix} x2$	3-0-BNReLU1 3-0-BNReLU2 3-0-ReLU 3-1-BNReLU1 3-1-BNReLU2 3-1-ReLU	[7, 7, 512]
global average pooling	-	[512]
2d fc, softmax	-	[2]

Table 1: The architecture of deep CNN used in our experiments. We use the 18-layer ResNet for a fair comparison with alternative approaches. Each x-x-ReLU is the output of a residual block.

For a fair comparison, we use the same partition of training and test data similar to prior methods [19, 22, 23, 25] (i.e., 235,599 images for training and the rest for testing). We also follow the same procedure to assign a binary aesthetic label to each image in AVA. Specifically, images with average ratings less than or equal to 5 are aesthetically negative, whereas others are aesthetically positive.

4.2 Implementation Details

Given an image of arbitrary resolution in our dataset, we initially resize its shorter edge to be 256 while keeping its aspect ratio. In this manner, we can keep the testing data pipeline compatible with the training data pipeline without changing the aspect ratios of original images, which we assume is important for photo aesthetic assessment. Figure 4 shows that the distributions of aesthetic scores of images of different aspect ratios share similar patterns.

We then randomly crop several patches of resolution 224 × 224 from each resized image. Random horizontal flipping (50%

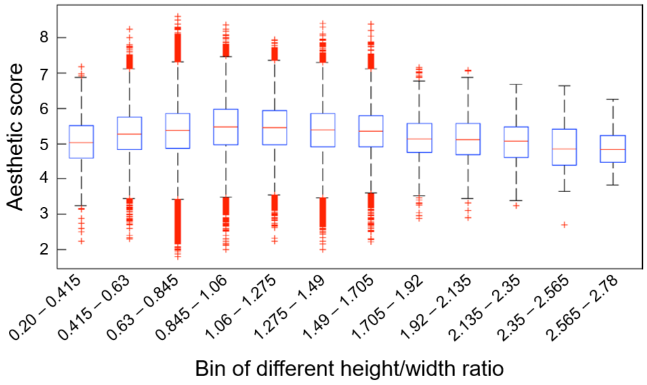

Figure 4: The distributions of aesthetic scores of images with various aspect ratios share similar patterns.

probability to flip) is conducted for data augmentation without color jittering or multi-scale cropping. Next, the collected patches are fed into the CNN to learn the feature representation for aesthetic assessment. An attention-based MP objective function (Section 3) is used to organize separated patches effectively and ensure that error back-propagation process eventually leads to desirable convergence with satisfactory assessment performance.

We build our system based on publicly available implementations of ResNet [9] using *Tensorflow*. Instead of training from scratch, our network is fine-tuned from a model pretrained on the ImageNet ILSVRC2012 dataset [27]. Each training process contains 32 epochs and it takes about 10 hours to complete on an NVIDIA TITAN X graphics card. At test time, it takes less than 0.1 ms to predict the aesthetic label of an image.

We apply a five-fold cross-validation technique on the training set of AVA to select the model hyperparameters. The hyperparameters include the learning rate (begins with 0.001 and is reduced by half every 20 epochs), fixed weight decay with 0.0001, and $\beta = 0.5$ in ω_β. The optimization process is performed using Nesterov SGD with a mini-batch size of 32.

4.3 Performance Evaluations

We use the *total classification accuracy* on the canonical AVA testing partition to validate the effectiveness of our proposed objectives for aesthetic assessment. In Table 2, we compare our results with several existing techniques, including handcrafted features [25], three VGG-Net based methods [19], a single-patch network [22] based on spatial pyramid pooling (SPP) [8], three types of aggregation based approaches as reviewed in Section 2, and a recent method [31] which casts aesthetic assessment as a regression task.

4.3.1 Our methods versus state-of-the-art approaches. Our proposed objective functions based on attention mechanism generally work better compared with existing techniques for image aesthetic assessment. The MP_{ada} scheme even outperforms the models with well-designed features that utilize hybrid information (e.g., [11, 19, 22, 23]). In Figure 5, we show four groups of aesthetic assessment results predicted by our MP_{ada} scheme based on the ResNet architecture, including aesthetically positive and negative predictions with high and low decision confidence.

Method	Core Features	Results
AVA [25]	handcrafted features	68.0
VGG-Scale [19]	non-uniform scaling	73.8
VGG-Pad [19]	uniform scaling + padding	72.9
SPP [22]	spatial pooling	76.0
VGG-Crop [19]	MP aggregation	71.2
DMA-Net [22]		75.41
MNA-CNN [19]		77.1
New-MP-Net [23]		81.7
DCNN [21]	multi-column aggregation	73.25
RAPID [21]		75.42
A&C CNN [12]		74.51
MTCNN [11]		78.56
MTRLCNN [11]		79.08
BDN [34]		78.08
Two-column DAN [6]		78.72
AA-Net [35]		76.9
DMA-Net-IF [22]	representation aggregation with explicit information	75.4
MNA-CNN-Scene [19]		77.4
A-Lamp [23]		82.5
NIMA [31]	distributions of human opinion scores	81.51
MP_{avg}	average weights	81.76
MP_{min}	minimum select	80.50
MP_{ada}	adaptive weights	**83.03**

Table 2: Comparisons between several state-of-the-art approaches and our proposed schemes. We list the core features of each method and the corresponding total classification accuracy on the AVA test set.

4.3.2 Comparisons between different attention-based schemes. Among all the three attention-based objectives proposed in Section 3, the MP_{ada} scheme achieves the highest aesthetic assessment accuracy. Figure 6 shows that the MP_{ada} scheme tends to assign larger weights to aesthetically negative examples with respect to positive ones. This adaptive scheme can help resolve the intrinsic class imbalance problem of our dataset and will lead to better assessment performance. We have also found that training based on the MP_{ada} scheme converges faster and reaches a lower minimum compared with the other two schemes.

4.3.3 β value for adaptive weight assignment. For a better understanding of the MP_{ada} scheme, we conduct additional experiments to train the model with different values of β (i.e., the hyperparameter for adaptive weight ω_β in Eqn 7). Figure 3 shows that when $\beta \in (0, 1)$, patches correspond to smaller probability and thus lower prediction confidence will be assigned a considerably larger weight, that is, a model with $\beta \in (0, 1)$ is more adaptive compared with those with $\beta \in (1, \infty)$. Our experimental results show that a model trained with $\beta \in (0, 1)$ generally outperforms the ones with $\beta \in (1, \infty)$ by a margin of $\sim$ 1% in terms of total classification accuracy.

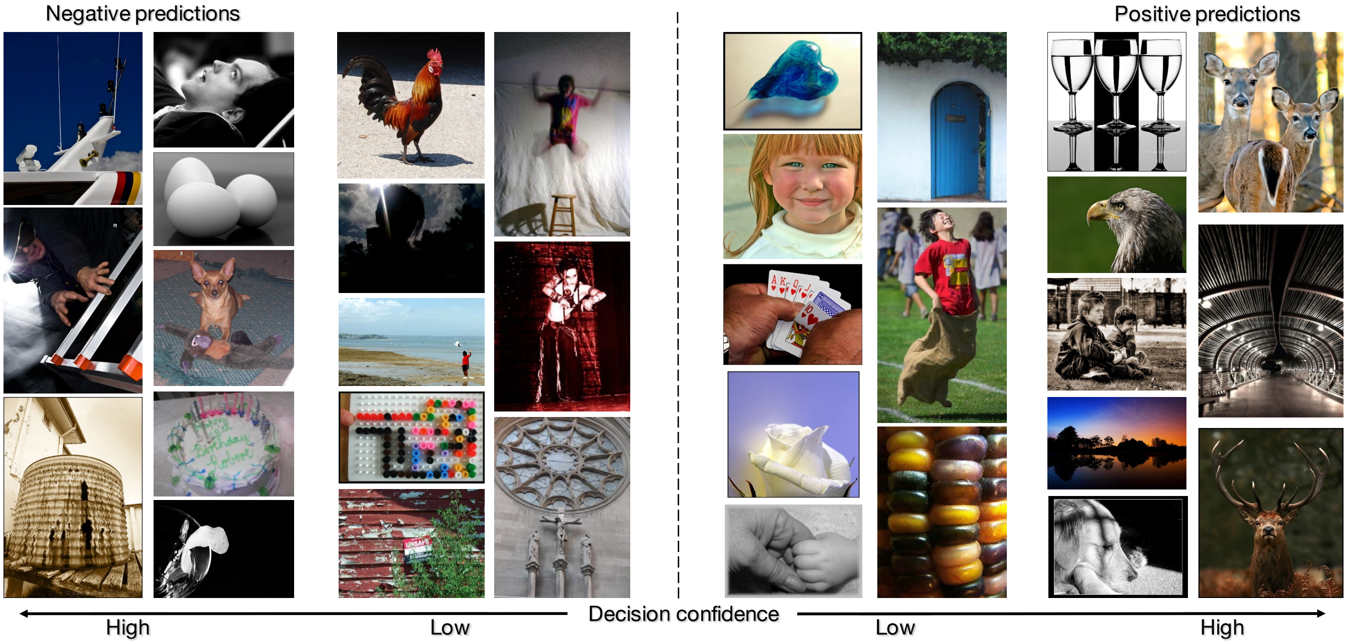

Figure 5: Our aesthetic assessment results on the AVA test set predicted by the MP_{ada} scheme.

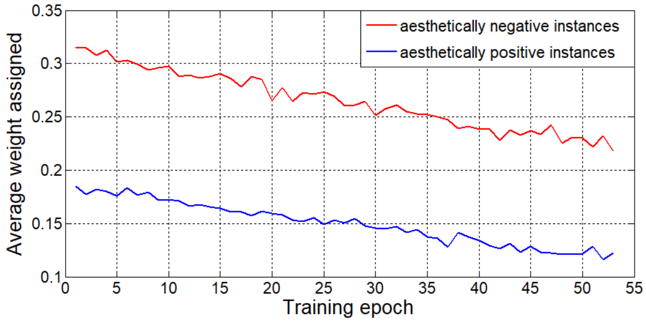

Figure 6: Average weights for patches from aesthetically positive and negative instances during the training process.

4.4 Further Investigation

In this section, we conduct ablation study to verify the effectiveness of our proposed method. We also investigate the learned representations to understand aesthetic-aware models further.

4.4.1 Ablation study: ResNet versus VGG. In addition to using the 18-layer ResNet [9], we also test the aesthetic assessment performance of the MP_{ada} scheme based on VGG16 nets, which are commonly used in previous studies [19, 22, 23]. In comparison with the number in Table 2, the accuracy of implementation using VGG16 nets is only approximately 0.5% lower. Note that a VGG16 net has a considerably larger network size (around 500MB) as compared to ResNet (approximately 40MB) and requires considerably more computational resources and time budget.

4.4.2 Representation correlation. For further understanding of the trained aesthetic-aware neural network, we investigate the representation vectors extracted from different layers of the ResNet architecture. We resort to singular vector canonical correlation analysis (SVCCA) [1] [24], a recently proposed powerful tool to understand the relationships among representations of various layers. We use $ResNet_{ImageNet}$, $ResNet_{AVA}$ and $ResNet_{Rand}$ to denote the original model pretrained on ImageNet, the fine-tuned model from $ResNet_{ImageNet}$ and the model trained from scratch, respectively. The SVCCA maps of $ResNet_{AVA}$ and $ResNet_{Rand}$ are shown in Figures 7 and 8, respectively. The notations of output nodes in the SVCCA maps are listed in Table 1. Figure 7 shows a positive correlation among various layers in the trained model, especially for high-level layers (e.g., the last six layers).

Our experiments show that $ResNet_{Rand}$ can still achieve good aesthetic classification results and is inferior to $ResNet_{AVA}$ by a moderate gap (around 1%). In comparison with the SVCCA map of $ResNet_{AVA}$ in Figure 7, the correlations of representations in $ResNet_{Rand}$ are relatively weaker, as shown in Figure 8. This result indicates that a weak representation correlation can cause degradation of aesthetic assessment performance.

4.4.3 Aesthetic-aware model versus object-oriented model. Since we start from a pre-trained object-oriented model $ResNet_{ImageNet}$ and then fine-tune it to obtain a model $ResNet_{AVA}$ for aesthetic assessment, it is interesting to see which parts of the network change the most during the fine-tuning process. Figure 9 shows the heatmap visualization where the SVCCA map of $ResNet_{AVA}$ has larger components than the SVCCA map of $ResNet_{ImageNet}$. This figure indicates that the high-level layers of a model for aesthetic assessment present stronger correlation compared with an object-oriented model.

[1] The code of SVCCA we use can be found at https://github.com/google/svcca.

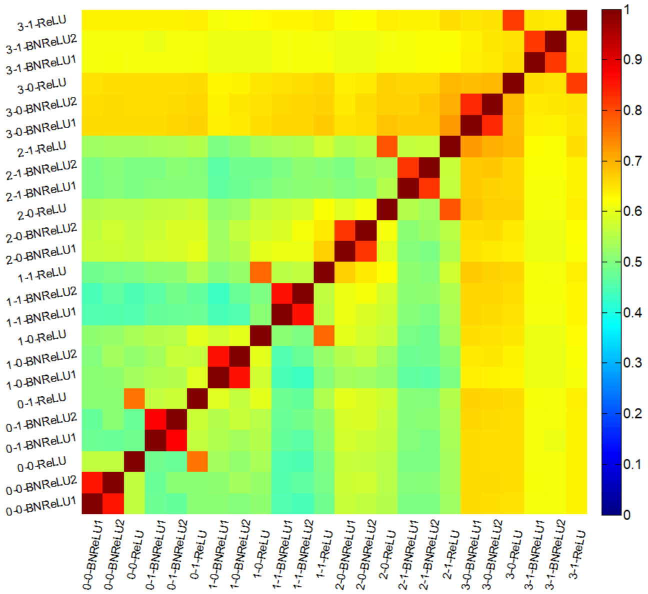

Figure 7: SVCCA map visualization of correlations between different layer representations in $ResNet_{AVA}$.

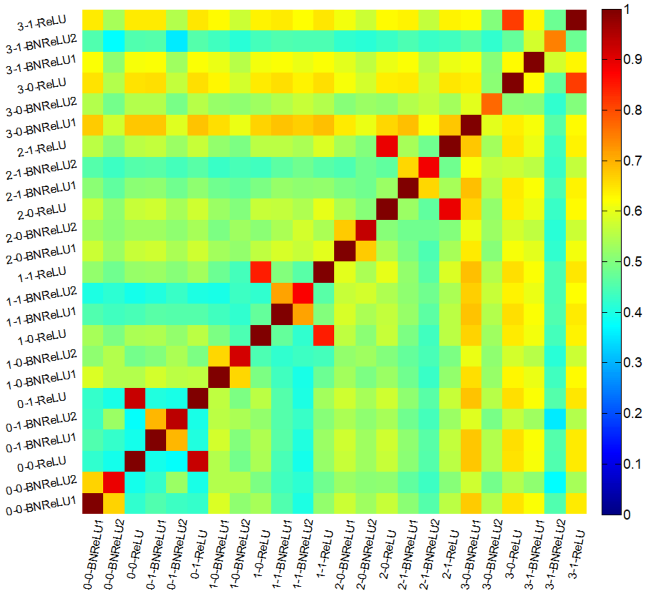

Figure 8: SVCCA map visualization of correlations between different layer representations in $ResNet_{Rand}$.

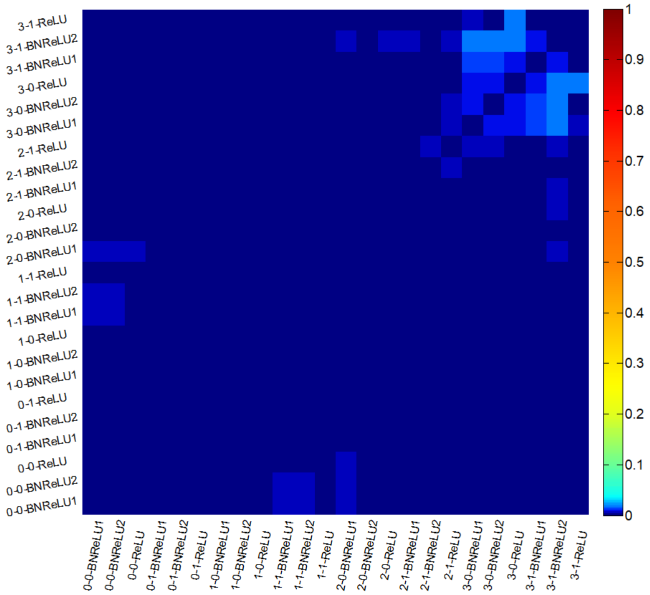

Figure 9: Heatmap visualization showing where the SVCCA map of $ResNet_{AVA}$ has larger components than the SVCCA map of $ResNet_{ImageNet}$.

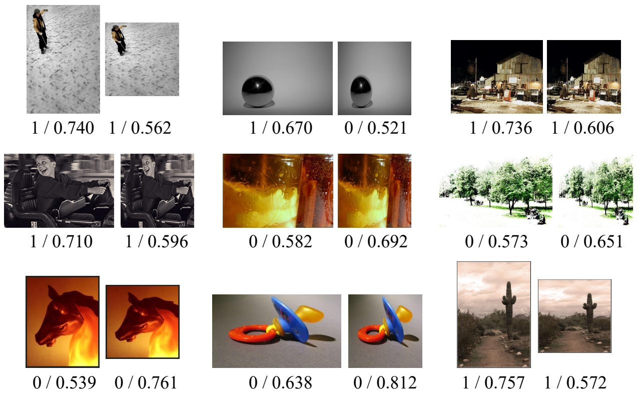

Figure 10: Our aesthetic assessment results and the corresponding prediction confidence with two scaling approaches: resizing shorter edge to be 256 while keeping the aspect ratio (left) and resizing to a resolution 256×256 (right).

4.4.4 Effect of image resizing methods. In our current system, we initially resize the shorter edge to be 256 for all the images to make the training and test data compatible with each other. This preprocessing operation is valid if aesthetic assessment result only relies on scale-invariant features and will not change after uniform resizing. In certain cases, such as the pictures with distinctive local contrast, resizing the images may change the aesthetic assessment results. We designate the study of image aesthetic assessment at the original resolution as future work.

In Figure 10, we compare our uniform scaling strategy with non-uniform scaling to a resolution of 256 × 256. For each image, we show our aesthetic assessment result and the corresponding prediction confidence. This figure shows that resizing without keeping the original aspect ratio will reduce the confidence for positive predictions and will increase the confidence for negative ones. This result is consistent with human visual perception because changing the aspect ratio of an ordinary image is likely to downgrade the image aesthetically.

5 CONCLUSIONS

In this study, we revisit the problem of image aesthetic assessment and propose a simple yet effective solution inspired by the attention mechanism. To learn a neural network based model for aesthetic assessment from training data with aesthetic labels only, we investigate three different weight assignment schemes for MP aggregation, namely, MP_{avg}, MP_{min}, and MP_{ada}. Our experimental results on the AVA test dataset show that our approach outperforms state-of-the-art approaches for image aesthetic assessment by a large margin. Among the three schemes presented, adaptive weight assignment MP_{ada} achieves the best aesthetic assessment performance due to larger weights assigned to meaningful instances during the optimization process, which help strengthen training signals and resolve the class imbalance issue of the dataset. We further validate our design choices via ablation study and evaluate the learned models by comparing different training strategies.

In the future we plan to investigate learning from unlabeled data to improve assessment performance further. Although our major goal is to improve the accuracy of image aesthetic assessment, another possible future avenue is to explore more compact aesthetic assessment models for various mobile applications, such as image enhancement and album thumbnail generation. Finally, combining image aesthetic assessment with other visual analysis tasks within a unified learning framework is also interesting.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable comments and suggestions. This work was supported by National Natural Science Foundation of China under nos. 61672520, 61573348, 61702488 and 61720106006, by Beijing Natural Science Foundation under No. 4162056, by the independent research project of National Laboratory of Pattern Recognition, and by CASIA-Tencent Youtu joint research project.

REFERENCES

[1] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2011. A holistic approach to aesthetic enhancement of photographs. *ACM Transactions on Multimedia Computing, Communications, and Applications* 7, 1 (2011), 21.

[2] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. 2015. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2956–2964.

[3] Kuang-Yu Chang, Kung-Hung Lu, and Chu-Song Chen. 2017. Aesthetic Critiques Generation for Photos. In *IEEE International Conference on Computer Vision*. IEEE, 3534–3543.

[4] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. 2017. Learning to Compose with Professional Photographs on the Web. In *Proceedings of ACM on Multimedia Conference*. ACM, 37–45.

[5] 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.

[6] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2017. Image Aesthetic Assessment: An experimental survey. *IEEE Signal Processing Magazine* 34, 4 (2017), 80–106.

[7] Michael Freeman. 2006. *The Complete Guide to Light and Lighting in Digital Photography (A Lark Photography Book)*. Lark Books.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*. Springer, 346–361.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *IEEE Computer Vision and Pattern Recognition*. IEEE, 770–778.

[10] Laurent Itti and Christof Koch. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience* 2, 3 (2001), 194.

[11] Yueying Kao, Ran He, and Kaiqi Huang. 2017. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing* 26, 3 (2017), 1482–1495.

[12] Yueying Kao, Kaiqi Huang, and Steve Maybank. 2016. Hierarchical aesthetic quality assessment using deep convolutional neural networks. *Signal Processing: Image Communication* 47, C (2016), 500–510.

[13] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The Design of High-Level Features for Photo Quality Assessment. In *IEEE Computer Vision and Pattern Recognition*. IEEE, 419–426.

[14] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint:1609.04836* (2016).

[15] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless C Fowlkes. 2016. Photo Aesthetics Ranking Network with Attributes and Content Adaptation. *European Conference on Computer Vision* (2016), 662–679.

[16] Yan Kong, Weiming Dong, Xing Mei, Chongyang Ma, Tong-Yee Lee, Siwei Lyu, Feiyue Huang, and Xiaopeng Zhang. 2016. Measuring and Predicting Visual Importance of Similar Objects. *IEEE Transactions on Visualization and Computer Graphics* 22, 12 (2016), 2564–2578.

[17] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision*. IEEE, 2999–3007.

[18] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *IEEE Computer Vision and Pattern Recognition*. IEEE, 3431–3440.

[19] Mai Long, Jin Hailin, and Liu Feng. 2016. Composition-preserving deep photo aesthetics assessment. In *IEEE Computer Vision and Pattern Recognition*. IEEE, 497–506.

[20] Ilya Loshchilov and Frank Hutter. 2015. Online Batch Selection for Faster Training of Neural Networks. *Mathematics* (2015).

[21] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. 2015. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia* 17, 11 (2015), 2021–2034.

[22] 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 *IEEE International Conference on Computer Vision*. IEEE, 990–998.

[23] Shuang Ma, Jing Liu, and Wen Chen Chang. 2017. A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *IEEE Computer Vision and Pattern Recognition*. IEEE, 722–731.

[24] Raghu Maithra, Gilmer Justin, Yosinski Jason, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*. 6076–6085.

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

[26] Clark V. Poling. 1975. *Johannes Itten, Design and Form: The Basic Course at the Bauhaus and Later*. Thames and Hudson. 368–370 pages.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S Bernstein, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[28] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training Region-Based Object Detectors with Online Hard Example Mining. In *IEEE Computer Vision and Pattern Recognition*. IEEE, 761–769.

[29] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint:1409.1556* (2014).

[30] Marijn F Stollenga, Jonathan Masci, Faustino Gomez, and Jürgen Schmidhuber. 2014. Deep networks with internal selective attention through feedback connections. In *Advances in neural information processing systems*. 3545–3553.

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

[32] Hanghang Tong, Mingjing Li, Hong-Jiang Zhang, Jingrui He, and Changshui Zhang. 2004. Classification of digital photos taken by photographers or home users. In *Pacific-Rim Conference on Multimedia*. Springer, 198–205.

[33] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. *IEEE Computer Vision and Pattern Recognition*, 3156–3164.

[34] Zhangyang Wang, Shiyu Chang, Florin Dolcos, Diane Beck, Ding Liu, and Thomas S Huang. 2016. Brain-inspired deep networks for image aesthetics assessment. *arXiv preprint:1601.04155* (2016).

[35] Wang Wenguan and Shen Jianbing. 2017. Deep cropping via attention box prediction and aesthetic assessment. In *IEEE International Conference on Computer Vision*. IEEE, 2186–2194.

[36] Luming Zhang. 2016. Describing Human Aesthetic Perception by Deeply-learned Attributes from Flickr. *arXiv preprint:1605.07699* (2016).