Signal Processing: Image Communication 67 (2018) 12–21

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Visual aesthetic understanding: Sample-specific aesthetic classification and deep activation map visualization

Chao Zhang [a,b], Ce Zhu [a,*], Xun Xu [c], Yipeng Liu [a], Jimin Xiao [d], Tammam Tillo [e]

[a] University of Electronic Science and Technology of China, Chengdu, China
[b] Sichuan Police College, Luzhou, China
[c] National University of Singapore, Singapore
[d] Xi'an Jiaotong-Liverpool University, Suzhou, China
[e] Libera Universit di Bolzano-Bozen, Bozen-Bolzano, Italy

ARTICLE INFO

Keywords:
Visual aesthetic quality assessment
Aesthetic understanding
Sample-specific weighting

ABSTRACT

Currently image aesthetic estimation using deep learning has achieved great success compared with the traditional methods by hand-crafted features. Similar to recognition problem, aesthetic estimation categorizes images into visually appealing or not. Nevertheless, it is desirable to understand why certain images are visually more appealing, in specific, which part of the image is contributing to the aesthetic preference. In fact, most traditional approaches adopting hand-crafted feature are, to some extent, able to understand part of image's aesthetic and content information while few studies have been conducted in the context of deep learning. Moreover, we discover that aesthetic rating is ambiguous so that many examples are uncertain in aesthetic level. This has caused a highly imbalanced distribution of aesthetic ratings. To tackle all these issues, we propose an end-to-end convolutional neural network (CNN) model which simultaneously implements aesthetic classification and understanding. To overcome the imbalanced aesthetic ratings, a sample-specific classification method that re-weights samples' importance is proposed. We find that dropping out ambiguous image, as common adopted by recent deep learning models, is a special case of the sample-specific method, and also figure out that as the weights of the non-ambiguous images increase, the performance is positively affected. In order to understand what is learned in the deep model, global average pooling (GAP) following the last feature map is employed to generate aesthetic activation map (AesAM) and attribute activation map (AttAM). AesAM and AttAM respectively represent the likelihood of aesthetic level for spatial location, and the likelihood of different attribute information. In particular, AesAM mainly accounts for what is learned in deep model. Experiments are carried out on public aesthetic datasets and state-of-the-art performance is achieved. Thanks to the introduction of AttAM, the aesthetic preference is explainable by visualization. Finally, a simple application on image cropping based on the AesAM is presented. The code and trained model will be publicly available on https://github.com/galoiszhang/AWCU.

1. Introduction

Image aesthetic analysis is becoming an increasingly important topic in computer vision and multimedia research community due to its application in image retrieval and image editing. Many attempts have been made to formulate the aesthetic analysis as a classification problem [1–3], i.e. categorizing images into discrete levels of aesthetic quality. Traditionally, support vector machine (SVM) combined with hand-crafted visual features such as color [4] and SIFT [5] was adopted to predict aesthetic quality [6]. Though outperforming the specific aesthetic features [7], the generic visual feature is not designed for aesthetic categorization.

As a step further, deep learning has emerged as the prevailing approach towards image classification problems [8], so does the study into deep aesthetic understanding [9]. Lu et al. [9] proposed to learn deep convolutional neural networks directly and developed a two-branch CNN to account for global and local cues. The deep models have been proved to be effective on public benchmarks, e.g. A Large-Scale Database for Aesthetic Visual Analysis (AVA) dataset, and outperform the conventional hand-crafted generic features.

* Corresponding author.
E-mail addresses: galoiszhang@gmail.com (C. Zhang), eczhu@uestc.edu.cn (C. Zhu), elexuxu@nus.edu.sg (X. Xu), yipengliu@uestc.edu.cn (Y. Liu), jimin.xiao@xjtlu.edu.cn (J. Xiao), ttillo@unibz.it (T. Tillo).

https://doi.org/10.1016/j.image.2018.05.006
Received 18 July 2017; Received in revised form 8 May 2018; Accepted 8 May 2018
Available online 29 May 2018

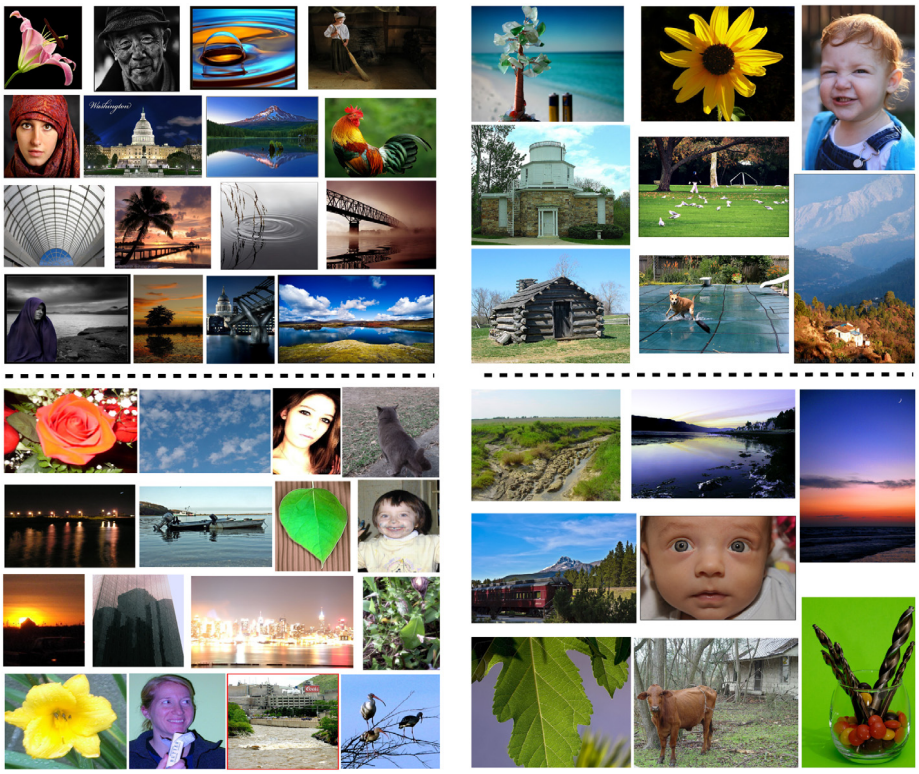

Fig. 1. Comparison of images from AVA dataset with high and low aesthetic qualities. A detailed analysis is found in the text.

Though high classification accuracy is achieved, the modern deep model approaches simply learn weights between different layers in the network and fail to reveal why certain images are more aesthetically appealing [9,10]. Such reasons can often be explained by an class activation map (CAM) [11] using the global average pooling (GAP) which reveal aesthetic supportive areas with high activation values. Nevertheless, GAP has rarely been studied for the purpose of aesthetic understanding, in particular visualizing the spatial aesthetic levels. Discovering aesthetic activation map can help us understand how the aesthetic preference is determined. Aesthetic activation map bears similarity with saliency detection [12], all aiming at localizing interesting areas and areas that are supportive to the action or aesthetic preference. By simply examining high aesthetic level area, it is very straightforward to judge if the approach is learning meaningful contents. Furthermore, localizing the aesthetic area can in turn improve the evaluation of aesthetic level by excluding the influence from irrelevant pixels. Finally, localized aesthetic area can be naturally helpful to image cropping where visually appealing areas are cropped. Therefore, it is highly desirable to discover from the image which area is the most supportive to the aesthetic preference.

Apart from the aesthetic activation map, the inherent ambiguity in aesthetic labeling poses a big challenge to data-driven approaches. Humans are often good at judging absolutely beautiful or ugly images, while those images of mediocre quality are often hard to be judged. We illustrate the case by looking into the examples in Fig. 1. Images above and below the dashed line are with high and low aesthetic values respectively. It might be easier for humans to assess the aesthetic values (either high or low) for the images on the left than the right. This is because the images on the right are of average scores (close to 5), thus more ambiguous while the left are of extreme scores (close to either 1 or 10), thus less ambiguous for judging aesthetic quality. Unfortunately, those clear examples are often rare. As a proof, the state-of-the-art public aesthetic benchmark, A Large-Scale Database for Aesthetic Visual Analysis (AVA), exhibits a highly non-uniform distribution over the aesthetic score as illustrated in Fig. 2. The majority images concentrate around the intermediate/average values (more than 80%). Regardless of such imbalanced distribution, the traditional models learn parameters by treating all available samples equally, therefore, the model could be overwhelmed by those ambiguous examples. This violates the commonsense that the more definite data should decide how the model is trained. Recently, people have been aware of the data imbalance issue in aesthetic prediction tasks. A common solution is dropping out average quality images [9,3,13,10]. Conclusions have been made that simply dropping out average quality images would deteriorate the overall performance. Without exception, the data imbalance issue cannot be remedied by the state-of-the-art deep neural networks. This can be easily understood as the majority of average quality image would dominate the parameter learning if all training samples are equally treated. As a result, recent studies have stressed the importance of pre-processing for imbalanced data in the framework of deep neural networks [14,15].

Finally, the attribute information is often visually ambiguous. Many of them are not visually detectable, such as `Emotive`, `Abstract`, `History`, `Humorous`, `Political`, `Science and Technology`, etc. According to some previous works [10,16], unless rearranging the attribute categories, the attribute information does little help to AVA's aesthetic assessment. Therefore, the attribute categorization module should not share parameters with aesthetic prediction but instead be learned separately.

Motivated by these reasons, in this paper, we focus on image aesthetic binary classification and understanding. We design two submodules that simultaneously classify the visual aesthetic/attribute and discover spatial locations in the image for supporting aesthetic preference and attribute classification (as depicted in Fig. 3). We, in particular, name the level of spatial activation as aesthetic predictive saliency. Such map potentially coincides with the interpretation of high/low aesthetic regions of the image and is supportive to the aesthetic prediction of images given standard labeling criteria. Finally, to cope with non-uniform distribution of aesthetic score, we propose a sample-specific loss function that re-weights each sample's importance. Specifically, the samples with average aesthetic scores (score close to 5) would contribute

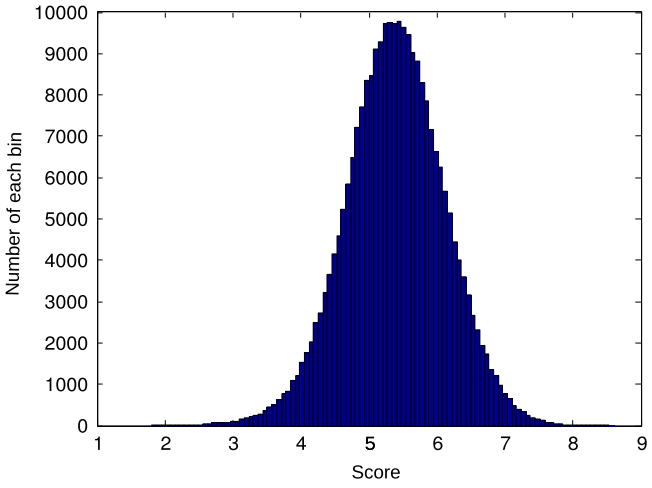

Fig. 2. The motivation of sample-specific classification. More than 80% samples in AVA are located at the intermediate interval of the score label. Sample specified weight is assigned to each sample to balance its importance in the training phase.

less to parameter learning while the more clear samples (score close to 1 or 10) should contribute more. Such a re-weighting scheme can be achieved by a customized weighting function.

Contributions. The objective of this paper is to classify the aesthetic level of a given image and understand what is learned in a deep model. The overview of our method is illustrated in Fig. 3, and the main contributions are as follows:

- Motivated by the score distribution of the images in AVA dataset, we propose a sample-specific classification model to re-weight the importance of different samples. Ambiguous samples are given lower weights while clear samples are weighted higher.
- Our model simultaneously conducts classification and visualization of aesthetic information in an end-to-end fashion. Compared with previous works, our model is simple and effective, and can also greatly reduce the number of parameters and memory.
- In order to explain what is learned in deep aesthetic classification, we jointly train two branches where the second branch on attribute information (AttAM) is added to compare with AesAM. On one hand, we discover that AesAM potentially corresponds to aesthetic predictive salient areas. On the other hand, we find AesAM partly coincides with AttAM.
- We study an application on aesthetic based image cropping. Our method is different from the previous work in aesthetic cropping using bottom-up salient object or scanning window.

The rest of the paper is organized as follows. Section 2 discusses two related topics: image and aesthetic estimation, and image aesthetic cropping. Our proposed model is explained in Section 3. The experimental results are presented in Section 4.

2. Related work

Image estimation and aesthetic estimation. Image estimation refers to score an image according to certain assessing criterion. For instance, when a photo with person is given, how can we estimate his/her age? In general, image estimation can be formulated as two tasks, i.e., classification or regression. In the problem with regression task, huge number of images with accurate labels can hardly be obtained. Image collection with the annotation of label is really time-consuming and subjective. Many applications mainly focus on image classification instead of regression by converting label into different categories. In fact, psychological evidence [17] showed that humans prefer to conduct evaluations on qualitative analysis instead of quantitative analysis, i.e., preferring to different levels. In our experience, human do not like to describe image aesthetic with exact value in practice. Instead, qualitative adjectives are usually used, such as excellent, good, and bad. Therefore, asking subjects to qualitatively evaluate image quality is a natural way to conduct subjective experiments, and can dramatically reduce the randomness of the scores and the burden placed on subjects [18].

Before 2012, many previous works on image estimation [19,20,7,21,22,4] mainly use support vector machine (SVM), random forest or support vector regression (SVR) on hand-crafted features. In recent years, CNNs [23,24,8,25,26] models have shown great success in image classification, which becomes the primary choice for researchers. Since the popularity of deep learning, image estimation also applies CNNs models [9,27,28], and it has received great improvement. The hand-crafted feature is generally complex compared with deep model learning which benefits from the use of the powerful hierarchical feature extractors. However, deep CNN models are always just regarded as a black-box, i.e., inputting an image, and outputting the probability vector for each class. In fact, the hand-crafted feature in traditional method sometimes can give us richer information though it cannot get as good performance as deep learning based methods.

Visual aesthetics estimation is an application on image estimation (or image classification), and it is also a subjective task. Assessing aesthetic quality of images automatically is still challenging in the field of computer vision. For image aesthetic estimation, many data-driven approaches [29,30,9,19,7,31,6,4] have been proposed to address this problem. Most of these methods aim to discover a meaningful and better aesthetic representation, and often formulate the representation learning as a single and standard classification task such as SVM, random forest and so on. Since the popularity of deep learning, many approaches [16,9,3,10,13,32,1,33] are based on CNN, which define image aesthetic estimation as a binary classification problem : high- or low-level aesthetic categories.

[9] was the first work using deep learning on aesthetic estimation, and then all subsequent works are based on this work. [1] shows that using patches from the original high-resolution images largely improves the performance. Input images need to be transformed via cropping, padding and so on, in some CNNs, which often damages image composition, reduces image resolution, or causes image distortion, thus compromising the aesthetics of the original images. In [3], to solve the problem of aesthetic distortion on resized image, they present a composition-preserving deep model method that directly learns aesthetics features from the original input images without any image transformations. Specifically, they add an adaptive spatial pooling layer upon the regular convolution and pooling layers to directly handle input images with various sizes. To allow for multi-scale feature extraction, they develop the Multi-Net Adaptive Spatial Pooling `ConvNet` architecture which consists of multiple sub-networks with different adaptive spatial pooling sizes and leverage a scene-based aggregation layer to effectively combine the predictions from multiple sub-networks. [16] proposes to learn a deep convolutional neural network to rank photo aesthetics in which the relative ranking of photo aesthetics are directly modeled in the loss function. This model incorporates joint learning of meaningful photographic attributes and image content information which can help regularize the complicated photo aesthetics rating problem. [10] addresses the correlation issue between automatic aesthetic quality assessment and semantic recognition. A correlation item between these two tasks is further introduced to the framework by incorporating the inter-task relationship learning.

Deep learning has received promising results in aesthetic classification. However, there are still some open problems. All previous works treat samples of AVA equally in the loss function. Considering the uncertain sample's side effect, we propose a model that can highlight certain samples while assign a low weight to uncertain samples. In

addition, binary classification for aesthetic analysis can distinguish high- or low-quality images, but fails to provide the reason why images are judged in a certain way. This observation motivates us to consider aesthetic visualization rather than simply assigning binary labels. Finally aesthetic visualization helps image cropping study.

Image cropping. Traditionally, automatic image cropping techniques follow two mainstreams, i.e. attention-based [34,35] and aesthetics-based methods [36,37]. Before the popularity of the deep learning, these works use the hand-crafted feature to get salient areas, and then crop these areas. To the best of our knowledge, there are few works that based on deep learning, and can be trained end-to-end. Most recently, [38] conducts an extensive study on traditional approaches as well as ranking-based croppers trained on various image features, which also release a good dataset for image cropping. They employ the deep feature that was extracted in advance, which does not train an unified end-to-end network. [3] also gives a result of the image cropping. They slide a cropping window through the whole image with the step size of 20 pixels, and then input the cropped window into the trained model to get its aesthetic score. They choose some high aesthetic windows as the final cropping result very intuitively. They get a good performance, while this method may be time-consuming. [39] proposed an automatic image cropping technique based on aesthetic map and gradient energy map. Based on the above maps, they give a preservation model to evaluate the quality of the composition for crops. [32] builds a connection between aesthetic assessment and aesthetic manipulation, with a focus on aesthetic-based image cropping. To the best of our knowledge, this is the first work that uses deep learning to implement image cropping in an end-to-end fashion.

3. Aesthetic classification and understanding

In this section, we propose the aesthetic level classification and understanding models, and analyze them from deep network and mathematical form. Firstly, we overview the whole structure of the model. Then, we introduce the sample-weighted aesthetic classifier. Thirdly, we demonstrate the pixelwise activation of our model to reflect the spatial aesthetic level predicted by the model. Finally, a simple application to image cropping is proposed.

3.1. Overview of the proposed model

The proposed network is composed of four parts, feature encoder (f_{enc}), global average pooling (f_{gap}), aesthetic level classifier (f_{hl}) and aesthetic attribute classifier (f_{att}), as shown in Fig. 3. Image classification in deep learning generally includes two parts: feature extraction layers f_{enc} and classifier learning layers f_{hl} or f_{att}. In the training phase, our model includes two branches of sub-network: the aesthetic level classifier f_{hl} with high/low level supervision, and the attribute classifier f_{att} with attribute category supervision. Given an input image $\mathbf{I}$, the network firstly extracts the feature $\mathbf{X} = f_{enc}(\mathbf{I}, \theta_{enc}), \mathbf{X} \in \mathbb{R}^{M\times S\times S}$, where θ_{enc} is the encoder parameter, M and S denote the channel number and the spatial size of each channel respectively. Then the global average pooling f_{gap} is applied to vectorize the last convolutional feature map by summing up each channel's map. Finally, the vectorized result is followed by two classifiers f_{hl} and f_{att}. All the two parallel operations can be formulated as

$$\begin{aligned}\hat{\mathbf{y}}_{hl} &= f_{hl}(f_{gap}(f_{enc}(\mathbf{I};\theta_{enc});\theta_{gap});\theta_{hl})\\ &= f_{hl}(f_{gap}(\mathbf{X};\theta_{gap});\theta_{hl}), \hat{\mathbf{y}}_{hl} \in \mathbb{R}^{C\times 1},\end{aligned} \tag{1}$$

and

$$\begin{aligned}\hat{\mathbf{y}}_{att} &= f_{att}(f_{gap}(f_{enc}(\mathbf{I};\theta_{enc});\theta_{gap});\theta_{att})\\ &= f_{att}(f_{gap}(\mathbf{X};\theta_{gap});\theta_{att}), \hat{\mathbf{y}}_{att} \in \mathbb{R}^{C'\times 1},\end{aligned} \tag{2}$$

where C and C' are the number of aesthetic level classes and aesthetic attribute classes respectively, $\hat{\mathbf{y}}_{hl}$ is the aesthetic level confidence predicted by f_{enc}, f_{gap} and f_{hl}, and $\hat{\mathbf{y}}_{att}$ is the aesthetic attribute confidence predicted by f_{enc}, f_{gap} and f_{att}.

The training process optimizes two tasks with softmax loss and multi-class sigmoid cross entropy loss using stochastic gradient descent in an unified framework. We jointly train the model as $\mathcal{L} = \mathcal{L}_{hl} + \mathcal{L}_{att}$, where $\mathcal{L}$, $\mathcal{L}_{hl}$ and $\mathcal{L}_{att}$ are the total loss, the loss for aesthetic level classification model and the loss for aesthetic attribute classification model respectively. The loss function and its gradient are as follows,

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{hl} + \mathcal{L}_{att}\\ &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{1}(\mathbf{y}_{hl}^{ic}=1)\log p(\hat{\mathbf{y}}_{hl}^{ic}=1|\mathbf{X}_i)\\ &\quad -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C'}\{\mathbf{y}_{att}^{ic}\log\sigma(\hat{\mathbf{y}}_{att}^{ic})+(1-\mathbf{y}_{att}^{ic})\log(1-\sigma(\hat{\mathbf{y}}_{att}^{ic}))\},\end{aligned} \tag{3}$$

and

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\Theta} &= -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{1}(\mathbf{y}_{hl}^{ic}=1)\frac{\log p(\hat{\mathbf{y}}_{hl}^{ic}=1|\mathbf{X}_i)}{\partial\Theta}\\ &\quad -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C'}\{\mathbf{y}_{att}^{ic}\frac{\partial\log\sigma(\hat{\mathbf{y}}_{att}^{ic})}{\partial\Theta}+(1-\mathbf{y}_{att}^{ic})\frac{\partial\log(1-\sigma(\hat{\mathbf{y}}_{att}^{ic}))}{\partial\Theta}\},\end{aligned} \tag{4}$$

where $\mathbf{X}_i$ represents the ith input image, Θ represents general parameter of whole model, $\hat{\mathbf{y}}_{hl}^{ic}$ is the ith image's aesthetic level confidence predicted on the cth class, $p(\hat{\mathbf{y}}_{hl}^{ic}=1|\mathbf{X}_i)$ is a softmax function of the aesthetic level confidence predicted on the cth class, $\hat{\mathbf{y}}_{att}^{ic}$ is the ith image's aesthetic attribute confidence predicted on the cth class, $\sigma(\hat{\mathbf{y}}_{att}^{ic})$ is sigmoid function of $\hat{\mathbf{y}}_{att}^{ic}$.

In the testing phase, f_{hl} branch generates aesthetic level prediction and aesthetic activation map, and f_{att} branch generates aesthetic attribute prediction and attribute activation map. We wish to obtain not only the aesthetic level/attribute prediction but also are interested in why certain images are predicted in the way they are. For this purpose, we reverse the order of GAP f_{gap} and the classifiers f_{hl} or f_{att} in the testing process. These operations of prediction can be concluded as follows,

$$\mathbf{X}^{hl} = f_{hl}(\mathbf{X};\theta_{hl}), \mathbf{X}^{hl} \in \mathbb{R}^{C\times S\times S}, \tag{5}$$

$$\hat{\mathbf{y}}_{hl} = f_{gap}(\mathbf{X}^{hl};\theta_{gap}) = f_{gap}(f_{hl}(\mathbf{X};\theta_{hl});\theta_{gap}), \hat{\mathbf{y}}_{hl} \in \mathbb{R}^{C\times 1\times 1}, \tag{6}$$

and

$$\mathbf{X}^{att} = f_{att}(\mathbf{X},\theta_{att}), \mathbf{X}^{att} \in \mathbb{R}^{C'\times S\times S}, \tag{7}$$

$$\hat{\mathbf{y}}_{att} = f_{gap}(\mathbf{X}^{att};\theta_{gap}) = f_{gap}(f_{att}(\mathbf{X};\theta_{att});\theta_{gap}), \hat{\mathbf{y}}_{att} \in \mathbb{R}^{C'\times 1\times 1}, \tag{8}$$

where $\mathbf{X}^{hl}$ and $\mathbf{X}^{att}$ represent aesthetic level activation map (AesAM) and aesthetic attribute activation map (AttAM) respectively. Comparing Eq. (1) and Eq. (5), and Eq. (2) and Eq. (7), we just reverse the order of GAP f_{gap} and the classifier f_{hl}(or f_{att}). Both activation maps characterize the spatial intensity of aesthetic level and attribute predictions, i.e. elements of the cth channel of $\mathbf{X}^{hl}$ (or $\mathbf{X}^{att}$) with high value indicate the corresponding area is contributing to the prediction of the cth aesthetic level or attribute. For convenience, we denote both activation maps as the deep activation map. The example of AesAM and AttAM is shown in Fig. 3, and we also give their cropping example.

Justifications for GAP. One may argue the global average pooling layer f_{gap}, as adopted here, lose too much spatial information compared with alternative mappings, e.g. fully connected layer. As seen in Figs. 3 and 4, f_{gap} follows the last feature layer with $S\times S$ convolutional layer. The output of f_{gap} is of the size $M\times 1\times 1$. We reckon the global average pooling is superior for two reasons. First, we choose f_{gap} instead of fully connected layers for generating class activation map. Intuitively,

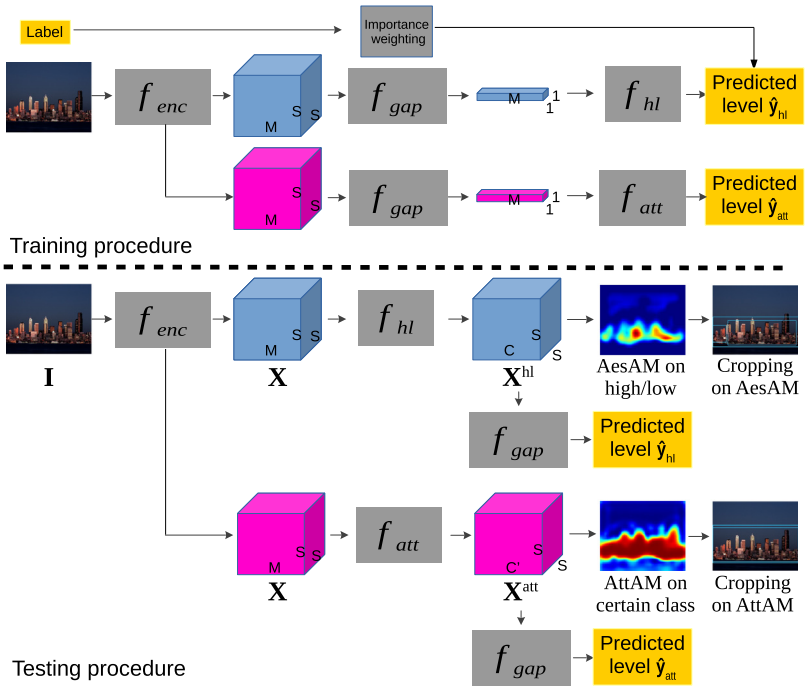

Fig. 3. The pipeline for aesthetic classification and understanding model, including the training and testing procedure.

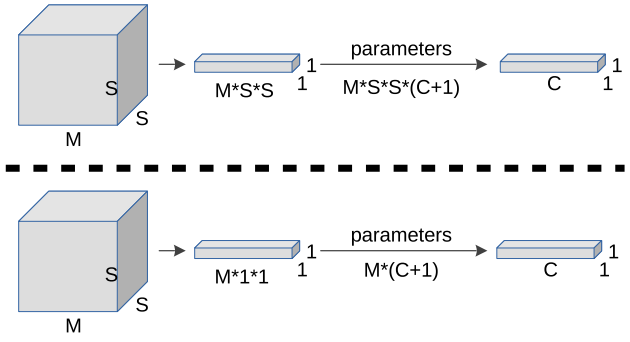

Fig. 4. The parameter analysis in the vectoring layer: the top and bottom are corresponding to the parameters of common fully connected layer and f_{gap} in the vectoring layer.

f_{gap} is very similar to fully convolutional layers in which the feature's spatial size is equal to the convolutional kernel size $S \times S$. Some helpful information is lost in f_{gap} because we use the average value to replace each channel of $\mathbf{X}$. But for the purpose of deep activation map, this procedure is necessary (see the detailed analysis in Section 3.3). Second, as seen in Fig. 4, f_{gap} relies on much fewer parameters than fully connected layer. For example, in VGG−16, the vectoring layer's parameters in fully connected layer is $256\times14\times14\times(4096+1) \approx 206\times10^6$, while the vectoring layer's parameters in GAP is $256\times(4096+1) \approx 1\times10^6$. In general, our model lower the volume of parameters, which is a key aspect for deep learning.

3.2. Sample-weighted aesthetic classification

In this section, we propose a sample-weighted aesthetic classification approach to counter-balance the biased distribution of aesthetic labels. As seen from Fig. 2. Almost 212092/255520 = 83% of samples are allocated in the interval [4.5, 6.5]. That is to say, there are too many ambiguous samples located at such narrow interval in AVA. To tackle this issue, we propose to provide each sample with a weighting importance. Given an image, if the aesthetic score is located at the both ends a high weight is assigned, otherwise intermediate score will be assigned with low weight. According to the distribution of the data, we choose a binary weight function. We note that the distribution of aesthetic scores is near a Gaussian distribution. An alternative weighting function could be an inverse-Gaussian function. Nevertheless, such a weighting function requires either sophisticated calibration of two parameters by hand, i.e., the mean and variance, or learning from data. Both are not trivial and may not yield better performance than binary weighting. In addition, Gaussian distribution does not mean implementing the Gaussian weighting. Therefore, a weighted softmax loss $\mathcal{L}_{whl}$ based on the binary per-sample weight is used to train the model, as given by

$$\mathcal{L}_{whl} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_i \mathbb{1}(\mathbf{y}_{hl}^{ic}=1)\log p(\hat{\mathbf{y}}_{hl}^{ic}=1|\mathbf{X}_i),$$
$$w_i = \begin{cases} a, & 4 < A_{score}^{i} < 6 \\ b, & others \end{cases} \tag{9}$$

where w_i is the per-sample weight, and A_{score}^{i} denotes the ith image's aesthetic score. Here we develop a binary weight w_i. In fact, we do not care about the exact values of a and b but instead the ratio b/a that controls the weight assignment.

At the training phase, the proposed total loss $\mathcal{L}$ for two branches can be given as

$$\mathcal{L} = \mathcal{L}_{whl} + \mathcal{L}_{att}, \tag{10}$$

and the gradient of weighted aesthetic level loss function $\mathcal{L}_{whl}$ on general parameter Θ can be given as follows,

$$\frac{\mathcal{L}_{whl}}{\partial\Theta} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} w_i \mathbb{1}(\mathbf{y}_{hl}^{ic}=1)\frac{\partial \log p(\hat{\mathbf{y}}_{hl}^{ic}=1|\mathbf{X}_i)}{\partial\Theta}. \tag{11}$$

where w_i is calculated outside the gradient.

3.3. Deep activation map for aesthetic understanding

In this section, we give an analysis on deep activation map and aesthetic understanding. As introduced in Eq. (5), Eq. (6), Eq. (7) and Eq. (8), in the testing phase, we swap f_{gap} and classifier f_{hl} (or f_{att}) to get deep activation map that provides the neuron activation on spatial

location, i.e., a cue for aesthetic understanding. Our model generates AesAM ($\mathbf{X}^{hl}$) and AttAM ($\mathbf{X}^{att}$) which are used to interpret the importance of spatial locations from aesthetic prediction and attribute prediction respectively. The former reveals the area which potentially appears aesthetic predictive saliency (see Fig. 3), and the latter interprets spatial locations from attribute information. They are termed as deep activation map rather than general saliency map [12]. From the point of neuron activation, they represent salient areas for aesthetic prediction, while general saliency map [12] mainly focuses on salient object. In fact not every aesthetic image include an object, such as scene or texture images. So we call it deep activation map instead of saliency map. On the other hand, as seen in Fig. 3, the average of AesAM (or AttAM) also represents the likelihood of aesthetic level (or attribute category).

For a given image $\mathbf{I}$, the feature map $\mathbf{X} \in \mathbb{R}^{M\times S\times S}$ can be represented by $\mathbf{X} = \{\mathbf{X}(m,:,:)\}_{m=1,2,\ldots,M}$, where $\mathbf{X}(m,:,:)$ denotes each channel of $\mathbf{X}$. Because the process of generating AesAM and AttAM are similar, we only need to discuss AesAM. In fact, f_{hl} intends to learn a $M\times C$ weight matrix $\mathbf{W} = \{w_m^c\}_{M\times C}$, where w_m^c is the weight parameter corresponding to class c for unit m. Obviously $\mathbf{X}(m,x,y)$ represents the activation of unit m in the last convolutional layer at spatial location (x, y). Then for unit m, in the training phase, the output of performing global average pooling is $f_{gap}(\mathbf{X}(m,:,:)) = \sum_{(x,y)}\mathbf{X}(m,x,y)$. In general, the predicted confidence probability on the cth class $\hat{\mathbf{y}}_{hl}^c$ is the weighted sum of each unit $\sum_m w_m^c f_{gap}(\mathbf{X}(m,:,:))$. Essentially, w_m^c indicates the neuron activation of $\mathbf{X}(m,:,:)$ for class c. In the testing phase, we denote AesAM as $\{A_c(x,y)\}_{c=1,2,\ldots,C}$. $\hat{\mathbf{y}}_{hl}^c$ also can be represented as $\sum_{(x,y)}A_c(x,y)$. The detailed relationship is given as follow,

$$\begin{aligned}\hat{\mathbf{y}}_{hl}^c &= \sum_m w_m^c f_{gap}(\mathbf{X}(m,:,:)) = \sum_m w_m^c \sum_{(x,y)}\mathbf{X}(m,x,y)\\ &= \sum_m\sum_{(x,y)} w_m^c\mathbf{X}(m,x,y)\\ &= \sum_{(x,y)}\sum_m w_m^c\mathbf{X}(m,x,y) = \sum_{(x,y)} f_{hl}(\mathbf{X}(:,x,y)) = \sum_{(x,y)} A_c(x,y) = \hat{\mathbf{y}}_{hl}^c.\end{aligned} \tag{12}$$

Hence $A_c(x,y)$ indicates the importance of the activation at spatial grid (x, y) leading to the classification of an image to class c. Eq. (12) corresponds to Fig. 3, which analyzes the relationship between training and testing phase.

Intuitively, each unit $\mathbf{X}(m,:,:)$ is activated by the weight matrix $\mathbf{W} = \{w_m^c\}_{M\times C}$. The element of deep activation map $A_c(x,y) = \sum_m w_m^c\mathbf{X}(m,x,y)$ is simply a weighted linear summation of each spatial location. $A_c(x,y)$ points out aesthetic predictive salient areas which potentially coincide with pleasing/unpleasant areas. Deep activation map includes two aspects of information: the aesthetic intensity in spatial location and aesthetic level for the whole map. For the former, $A_c(x,y)$ represents the activation of the high and low level aesthetic image. For the later, the average $\sum_{(x,y)}A_c(x,y)$ of all spatial locations represents the image's aesthetic level. In the same way, we can get attribute activation map corresponding to the attribute information.

3.4. Image cropping as an application

After obtaining AesAM and AttAM, we use them to guide image cropping as a potential application. For each map, we follow the method [11] to get the bounding boxes, in which they first segment the activation map using a threshold value that is 30% of the maximum value, and then take the bounding box that covers the largest connected component in the segmentation map. For the image cropping, here we give two alternatives: only using AesAM, and using both AesAM and AttAM. In the former, we use AesAM to get the bounding boxes (see the column 2, 3, 4 in Fig. 7). In the latter one, AesAM and AttAM are used to get two groups of bounding boxes, and which are merged by calculating the overlapping region of bounding boxes. By adjusting the threshold large overlapping bounding boxes can be retained. Finally we can obtain the cropped patches that include both aesthetic level and attribute information (see the last column in Fig. 7). With the absence of AttAM, we can still get the cropping from AesAM (see the fourth column in Fig. 7). However this cropping is hard to explain from aesthetic content information.

We do not quantitatively evaluate the quality of image cropping in this work due to: (i) The definition of image cropping ground-truth is hard itself due to the ambiguity and subjectivity of definition. (ii) The aesthetic activation map picks up the region which favors aesthetic prediction. Nevertheless, we believe a quantitative evaluation aesthetic-based image cropping is an interesting topic and deserve further investigation in the future.

4. Experiments

We carry on experiments on challenging public aesthetic datasets AVA to verify the effectiveness of our proposed framework and then analyze the visualization of aesthetic activation and its application on image cropping.

4.1. Dataset

The AVA dataset [6] is the largest publicly available aesthetics dataset providing over 250,000 images in total. Each image in the dataset was rated by roughly 200 people with the rating score ranging from 1 to 10, with 10 indicating the highest aesthetics quality. For a fair comparison, we follow the experimental settings in [9,16], and use the same collection of training data and testing data: 230,000 images for training and 20,000 images for testing. For the ease of evaluation, all images are divided into two categories, i.e., low-aesthetic images with aesthetic score from 1–5 and high-aesthetic images with score from 6–10, following the same criteria as in [6,9,16,10].

In addition to the aesthetic ratings, each image is associated with no more than 2 tags (attributes) annotated out of 66 attribute categories. We use these attributes to explain what is learned in the deep model. The frequencies of attributes are highly imbalanced. Some attributes appear significantly less than others, thus leading to very complicated correlations between attributes. With few exceptions, attribute classification is still rarely touched up until now. In particular, [16] proposed a mini-dataset Based on AVA for attribute classification by rearranging the attribute information. In this work, we focus on using attribute for explaining what is learned in aesthetic level classification model.

4.2. Experimental settings

We implement experiment with three tasks: aesthetic level classification, deep activation map generation and image cropping. In aesthetic classification, we employ `Alex-net` [8] and `VGG-16-net` [25] pre-trained on ImageNet [8,11] as our encoder model f_{enc} respectively. In the pre-processing phase the original image is resized to a fixed size, and then a patch is randomly cropped from the resized image. The former mainly meets the requirement of CNN input, and the latter can reduce the risk of over-fitting in the training phase. In both training and testing stages, we resize each images to 256×256 pixels and crop five regions each with 227×227 pixels. Four of the croppings are made by aligning the cropping window along the four corners of the raw image and the last cropping is made by positioning the window in the mid-center. For `Alex-net`, the initial learning rate is set at 0.001, and periodically annealed by 0.1. For `VGG-16`, the parameters differ from `Alex-net` by setting initial learning rate at 0.0005. Weight decay is set as 0.0005 and momentum is set as 0.9.

After training the network, we use the trained classifier f_{hl} and f_{att} to weight the last convolutional feature map. Then two deep activation maps: AesAM and AttAM, are obtained. Based on the deep activation map, we use f_{gap} to get the level confidence. The evaluation criteria is classification accuracy.

Table 1
Comparison of different combination of modules for aesthetic classification.

Model	AesCNN	AesCNN-W	AesAttCNN	AesAttCNN-W
Alex-Net	76.82	**77.39**	76.77	**77.18**
Vgg-Net	78.60	**78.87**	76.89	**78.62**

4.3. Aesthetic classification

In this section, we elaborate how does each module contribute to the aesthetic level classification performance. Extensive combinations of two-branch training and weighted training are evaluated for comparison. In specific, we first evaluate the basic model that relies on the f_{hl} branch alone to predict aesthetic level. This is a single branch model which only takes aesthetic supervision. This model is equivalent to cutting off the attribute prediction branch in Fig. 3. We term this model as **AesCNN**. Furthermore, we are wondering if attribute prediction can benefit learning a better encoder network. For this reason, we study the full two-branch model as illustrated by Fig. 3. The two-branch model updates the encoder network by jointly optimizing both the aesthetic and attribute prediction objectives. We term this two-branch model as **AesAttCNN**. Finally, as we notice in Section 3.2, assigning uniform weights to all training samples is not the optimal solution. To counterbalance the impact of imbalance training samples, we develop a sample-weighted aesthetic classification strategy by weighting the training sample according to its aesthetic score. We term the weighted models as **AesCNN-W** and **AesAttCNN-W** respectively. We report the performance of above four combinations on AVA dataset in Table 1.

For both `alex` and `vgg`, models with weighting strategy are consistently better than without weighting. This suggests the effectiveness of the proposed weighting method. Although the two-branch model (AesAttCNN) performs slightly worse than the single-branch model (AesCNN), we believe the attribute information is a good indication of the spatial area which supports the aesthetic prediction. In addition, the attribute activation map provides a baseline for validating the aesthetic level activation map. In general, all the above comparisons have demonstrated the effectiveness of weighting method on aesthetic level prediction.

4.4. Re-weighting the samples

In this section, we analyze how does the choice of re-weighting parameter affect the performance of aesthetic classification. In specific, we denote the scheme of re-weighting as $w(a:b)$ where a and b are the parameters introduced in Section 3.2. We firstly study a special case of re-weighting, i.e., dropping out average quality images which has been studied by [10,3]. Dropping out strategy is equivalent to setting $a = 0$ and $b = 1$ ($w(0:1)$). Furthermore, we evaluate gradually by reducing the ratio b/a until assigning uniform weights to all samples ($w(1:1)$). The results for both evaluations are presented in Table 2 and Fig. 5. It is clear that the dropping out strategy is by no means better than re-weighting scheme regardless of the parameters (b/a) we choose. In addition, by increasing the ratio b/a from 1 to 7, i.e., the range {1, 2, 3, 4, 5, 6, 7}, we observe that recognition accuracy has been improved gradually (see Fig. 5). Both observations suggest the effectiveness of the re-weighting scheme.

4.5. Comparison with the state-of-the-art

In this section, we further compare our model with the state-of-the-art models on AVA dataset. Specifically, we consider the following models, RDCNN [9], DMA-Net [1], MNA-CNN [3], Reg + Rank + Att + Cont [16] and Triplet-loss [40]. [9] is the first work using deep learning on aesthetic estimation, they achieved 74.46% accuracy comparing to the hand-crafted feature with shallow classifier which achieves 68.00%. [1] uses patches cropped from the original high-resolution images improves performance from 74.46% to 75.41%. [3] presents a composition-preserving deep model method that directly learns aesthetics features from the original input images without any image transformations by an adaptive spatial pooling layer, their performance is 77.40%. [16] proposes to rank aesthetic level in which the relative ranking of photo aesthetics are directly embedded in the loss function. This model incorporates joint learning of meaningful photographic attributes and image content information which can help regularize the complicated photo aesthetics rating problem. They achieved 75.48% and 77.33% accuracy on `alex` and `vgg` respectively. [10] addresses the correlation between automatic aesthetic quality assessment and semantic recognition, in which multiple losses are combined in the decision layer. They achieve 77.35% and 78.46% accuracy on `alex` and `vgg`.

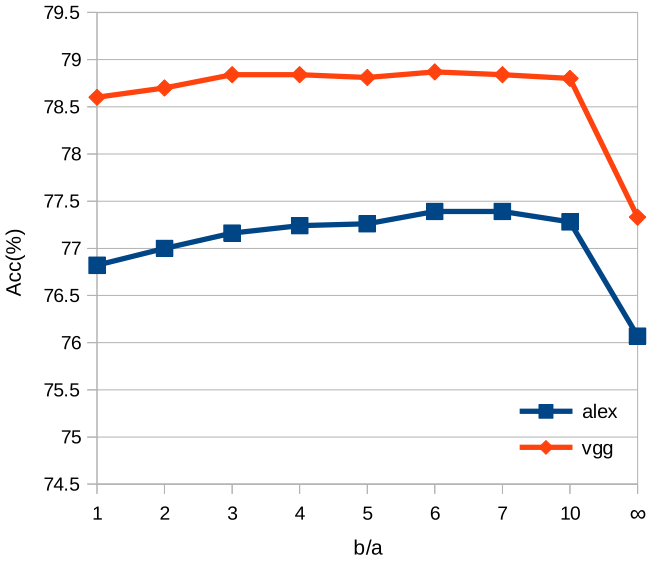

Fig. 5. Results of different weights: As the ratio b/a increases, its performance will also increases. After a certain point the performance will decreases.

We report the comparative results in Table 3. As we observe, our final models, AesCNN-W and AesAttCNN-W, outperform most existing models. In specific AesCNN-W based on `alex` and `vgg` encoder have achieved 77.39% and 78.87% accuracy, which is the state-of-the-art result. We note that MTRLCNN achieved rather close performance. MTRLCNN's good performance is mainly derived from the exploration of relationship between the aesthetic classification and aesthetic attribute. Both `Reg+Rank+Att` and `Reg+Rank+Att+Cont` combine score regression, ranking information, attribute information and additional content information to train the model, and also perform well. The superiority of `Reg+Rank+Att+Cont` on `Reg+Rank+Att` comes from the additional content preprocessing that is very important for AVA dataset. In fact, AVA dataset content information exploration is a very interesting topic. MNA-CNN-Scene gives a novel adaptive spatial pooling net, while its overall performance is not the best. In general we observe that our model, without any bells and whistles, achieves the state-of-the-art result. It is worth to note that only two losses are used in the proposed method. We believe that our proposed method can be easily extended to other models and obtain further improvement.

4.6. Deep activation maps

In the testing procedure, we adjust the order of GAP and classifier to get the deep activation map. Given a test image, we first obtain the AesAM and AttAM. Here we mainly analyze AesAM's representation (see Fig. 6). AesAM incorporates two aspects: aesthetic intensity for spatial location and global aesthetic level. In Fig. 6, each image's low level

Table 2
Comparison of different weight w for aesthetic classification.

Model	w(1:1)	w(1:2)	w(1:3)	w(1:4)	w(1:5)	w(1:6)	w(1:7)	w(1:10)	w(0:1)
b/a	1	2	3	4	5	6	7	10	∞
Alex-Net	76.82	77.00	77.16	77.24	77.26	77.39	77.39	77.28	76.07
Vgg-Net	78.60	78.70	78.84	78.84	78.81	78.87	78.84	78.80	77.33

Table 3
Comparison of performance between different models on the AVA dataset. Results are arranged by publishing time.

Model	Year	AVA Accuracy
Murrary [6]	2012	68.00%
RDCNN [9]	2014	74.46%
DMA-Net-ImgFustat [1]	2015	75.41%
Reg-Rank + Att(alex) [16]	2016	75.48%
Reg-Rank + Att + Cont(alex) [16]	2016	77.33%
MNA-CNN-Scene(vgg) [3]	2016	77.40%
Triplet-Loss [40]	2016	75.83%
MTCNN(alex) [10]	2017	76.15%
MTCNN(vgg) [10]	2017	77.73%
MTRLCNN(alex) [10]	2017	77.35%
MTRLCNN(vgg) [10]	2017	78.46%
AesCNN-W(alex)		**77.39%**
AesCNN-W(vgg)		**78.87%**
AesAttCNN-W(alex)		77.18%
AesAttCNN-W(vgg)		78.62%

AesAM and high level AesAM are shown. The high level AesAM points out images' high level aesthetic predictive salient areas while the low level AesAM points out the image's low level ones. For example, the face in both left and right in the fifth row looks pleasing from human perception. Their faces in high level AesAM, instead of low level AesAM, both have very strong activation though some off-the-target mini areas still exist. That is to say, pleasing area from human perception generally coincides with the strong activation in high level AesAM. This idea corresponds to Eq. (12) and Section 3.3. The aesthetic activation map is an indicator why certain images are predicted with the aesthetic high or low level. For instance, the tip of the leave on the last row is the area with high contrast and contribute the most to high aesthetic categorization.

The definition of aesthetic, in this work, comes from AVA dataset's groundtruth label, i.e., high or low. From the point of psychology, human perception has a special definition on aesthetic, and different people may have different opinions on the aesthetic. For instance, different people may have different aesthetic feelings/judgments on the photos in Fig. 1. In general, this definition is different from that of high/low level aesthetic, thus there still exist some slight differences on the visual understanding. As shown in Fig. 6, for different people, few AesAM may not coincide with human's perception very well. Nevertheless, the correlation between aesthetic activation map and human perceived pleasing area is yet established. Further psycho-visual experiments are required to fill this gap.

In addition, we can also consider the global state of the AesAM. On the left side of Fig. 6, the sum or average of the low level AesAM is smaller than the high level AesAM, then we predict that this image is with aesthetic high level. On the right side of Fig. 6, the sum or average of the low level AesAM is larger than the high level AesAM, then we predict that this image is with aesthetic low level. Considering the face in the fifth row again, on the left side of Fig. 6, the average activation of all pixels in the third column is stronger than the second column, i.e., $\sum_{(x,y)} A_{low}(x,y) < \sum_{(x,y)} A_{high}(x,y)$. While for the face on the right side, the average activation of all pixels in the fourth column is weaker than the second column though the face area active very strong, i.e., $\sum_{(x,y)} A_{low}(x,y) > \sum_{(x,y)} A_{high}(x,y)$. It means the surrounding area of the face is non-pleasing area which have stronger activation than the face area. In general, what we have learned in AesAM is explainable on visualization.

Finally, in order to explain AesAM on content information, we give a comparison between AesAM and AttAM. From Fig. 7, we can see that the area of the strong activation on the second column (high level AesAM) will also have strong activation on the fifth column (certain attribute AttAM). AesAM implicitly coincides with AttAM which includes some attribute information. That is to say, what we have learned in AesAM is also explainable on the attribute. As illustrated in Fig. 7, there is some overlap between the AesAM and AttAM indicating both supervisions are supported by the similar visual cues.

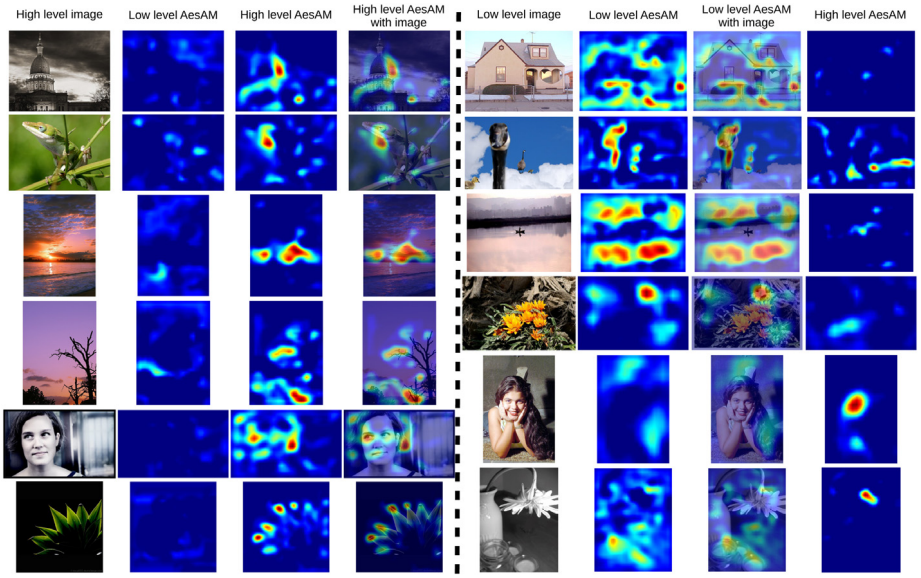

Fig. 6. Deep activation map aiming at spatial location: (i) Four columns in the left/right correspond to high/low level image, low level activation map(non-appealing), high level activation map(appealing) and the merged image. (ii) Each image is associated with high/low level AesAM. On the left, the sum of high level AesAM is larger than the sum of low level AesAM, and otherwise on the right. (Best viewed in color and magnifier.).

Fig. 7. Deep activation map and its application on image cropping.

In general, high level AesAM coincides with the pleasing area from both the deep model analysis (Eq. (12)) and human perception (Fig. 6) though there are still some off-the-target instances. On the other hand, AesAM partly coincides with AttAM (Fig. 7).

4.7. Application: image cropping

Based on the AesAM and AttAM, we can get the bounding box by the method [11]. We first segment the AesAM and AttAM using the threshold value that is 30% of the maximum value in each map, and then take the bounding box that covers the largest connected component in the segmentation map. About the image cropping, there are two choices. On one hand, we can get the cropping result by AesAM alone (see the fourth column in Fig. 7). On the other hand, we combine two groups of bounding boxes that includes both the aesthetic and attribute information. In the experiment, we set the threshold of the intersection over union (IOU) as 0.3. The detailed description is given at Section 3.4. If the IOU of two bounding boxes is larger than the threshold, we select two bounding boxes' intermediate point, i.e., the average of two bounding boxes' coordinates, as the new bounding box's coordinates. The concrete examples can be seen in Fig. 7.

The 66 attributes have very complex relation. The scope of attribute information is very general, in which many attributes come from life experience or from different point of view such as `science and technique`, `macro`, `Emotive`, `Abstract`, `History`, `Humorous`, `Political`. Each image may belong to multiple attribute classes. For each image, we choose the top 3 attribute map because of the complicated relationship between attributes. If the attribute label is in the top 3, then we set it as the correctly generated AttAM. Some other attributes are hardly visualized such as `humorous`, `black and white`, `still`, `History` and so on. So we consider 49 attributes of them in the procedure of AttAM generation.

The cropping results are shown in Fig. 7. From the first column to the eighth column, they respectively represent: the input image, high level AesAM, AesAM with its bounding boxes, input image with its bounding boxes, AttAM, AttAM with its bounding boxes, input image with its bounding boxes, final cropping result.

5. Conclusion and future work

In this paper, we present a sample-specific aesthetic classification model that simultaneously implements classification and the deep activation map visually. We point out which area or part support the aesthetic prediction. The aesthetic predictive salient areas potentially coincides with human perception of pleasing areas. The final aesthetic prediction is achieved by averaging out the aesthetic predictive areas. We also compare the AesAM and AttAM to describe that AesAM is also reasonable on attribute information. This builds the connection between image aesthetic and image attribute. Overall we not only point out which area is beautiful, but also know why this part is beautiful.

Acknowledgments

This research is supported by National Natural Science Foundation of China (NSFC, No. 61571102, No. 61602091, No. 61372187), Applied Basic Research Programs of Science and Technology in Sichuan (No. 2018JY0035), Sichuan Police College Research Program (No. 13SCJYKY42, No. 13SCJYKY43).

References

[1] X. Lu, Z. Lin, X. Shen, R. Mech, J.Z. Wang, Deep multi-patch aggregation network for image style, aesthetics, and quality estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 990–998.
[2] J. Guo, S. Gould, Deep CNN ensemble with data augmentation for object detection, 2015, arXiv preprint. arXiv:1506.07224.
[3] L. Mai, H. Jin, F. Liu, Composition-preserving deep photo aesthetics assessment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 497–506.
[4] X. Tang, W. Luo, X. Wang, Content-based photo quality assessment, IEEE Trans. Multimed. 15 (8) (2013) 1930–1943.
[5] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[6] N. Murray, L. Marchesotti, F. Perronnin, AVA: A large-scale database for aesthetic visual analysis, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2408–2415.
[7] L. Marchesotti, F. Perronnin, D. Larlus, G. Csurka, Assessing the aesthetic quality of photographs using generic image descriptors, in: IEEE International Conference on Computer Vision, 2011, pp. 1784–1791.
[8] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[9] X. Lu, Z. Lin, H. Jin, J. Yang, J.Z. Wang, Rapid: Rating pictorial aesthetics using deep learning, in: ACM International Conference on Multimedia, 2014, pp. 457–466.
[10] Y. Kao, R. He, K. Huang, Deep aesthetic quality assessment with semantic information, IEEE Trans. Image Process. 26 (3) (2017) 1482–1495.
[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[12] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[13] Y. Kao, K. Huang, S. Maybank, Hierarchical aesthetic quality assessment using deep convolutional neural networks, Signal Process., Image Commun. 47 (2016) 500–510.
[14] C. Huang, Y. Li, C. Change Loy, X. Tang, Learning deep representation for imbalanced classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5375–5384.
[15] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, P.J. Kennedy, Training deep neural networks on imbalanced data sets, in: Neural Networks (IJCNN), 2016 International Joint Conference on, IEEE, 2016, pp. 4368–4374.
[16] S. Kong, X. Shen, Z. Lin, R. Mech, C. Fowlkes, Photo aesthetics ranking network with attributes and content adaptation, in: European Conference on Computer Vision, 2016, pp. 662–679.
[17] E. Hutchins, Cognition in the Wild. 1995, vol. 14, MIT Press, Cambridge, USA, 1995, pp. 399–406.
[18] W. Hou, X. Gao, D. Tao, X. Li, Blind image quality assessment via deep learning, IEEE Trans. Neural Netw. Learn. Syst. 26 (6) (2015) 1275–1286.
[19] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image retrieval: Ideas, influences, and trends of the new age, ACM Comput. Surv. 40 (2) (2008) 5.
[20] J.-Y. Zhu, A. Agarwala, A.A. Efros, E. Shechtman, J. Wang, Mirror mirror: Crowdsourcing better portraits, ACM Trans. Graph. 33 (6) (2014) 234.
[21] Y. Fu, G. Guo, T.S. Huang, Age synthesis and estimation via faces: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1955–1976.
[22] H. Han, C. Otto, A.K. Jain, Age estimation from face images: Human vs. machine performance, in: IEEE International Conference on Biometrics, 2013, pp. 1–8.
[23] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[24] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, 2012, arXiv preprint. arXiv:1207.0580.
[25] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint. arXiv:1409.1556.
[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, 2014, arXiv preprint. arXiv:1409.4842.
[27] G. Levi, T. Hassner, Age and gender classification using convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2015, pp. 34–42.
[28] D. Yi, Z. Lei, S.Z. Li, Age estimation by multi-scale convolutional network, in: Asian Conference on Computer Vision, Springer, 2015, pp. 144–158.
[29] T. Malisiewicz, A. Gupta, A. Efros, et al., Ensemble of exemplar-SVMs for object detection and beyond, in: IEEE International Conference on Computer Vision, 2011, pp. 89–96.
[30] R. Datta, D. Joshi, J. Li, J.Z. Wang, Studying aesthetics in photographic images using a computational approach, in: European Conference on Computer Vision, 2006, pp. 288–301.
[31] L. Marchesotti, N. Murray, F. Perronnin, Discovering beautiful attributes for aesthetic image analysis, Int. J. Comput. Vis. 113 (3) (2015) 246–266.
[32] Y. Deng, C.C. Loy, X. Tang, Image aesthetic assessment: An experimental survey, 2016, arXiv preprint. arXiv:1610.00838.
[33] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, X. Xu, A multi-scene deep learning model for image aesthetic evaluation, Signal Process., Image Commun. 47 (2016) 511–518.
[34] F. Stentiford, Attention based auto image cropping, in: Workshop on Computational Attention and Applications, ICVS, 2007.
[35] J. Chen, G. Bai, S. Liang, Z. Li, Automatic image cropping: A computational complexity study, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 507–515.
[36] M. Nishiyama, T. Okabe, Y. Sato, I. Sato, Sensation-based photo cropping, in: Proceedings of the 17th ACM International Conference on Multimedia, 2009, pp. 669–672.
[37] C. Fang, Z. Lin, R. Měch, X. Shen, Automatic image cropping using visual composition, boundary simplicity and content preservation models, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 1105–1108.
[38] Y.-L. Chen, T.-W. Huang, K.-H. Chang, Y.-C. Tsai, H.-T. Chen, B.-Y. Chen, Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study, in: IEEE Winter Conference on Applications of Computer Vision, 2017, pp. 226–234.
[39] K. Yueying, R. He, H. Kaiqi, Automatic image cropping with aesthetic map and gradient energy map, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2017, pp. 1982–1986.
[40] K. Schwarz, P. Wieschollek, H. Lensch, Will people like your image? 2016, arXiv preprint. arXiv:1611.05203.