Learning to assess visual aesthetics of food images

Kekai Sheng^{1,2}, Weiming Dong² (✉), Haibin Huang³, Menglei Chai⁴, Yong Zhang⁵, Chongyang Ma³, and Bao-Gang Hu²

© The Author(s) 2020.

Abstract Distinguishing aesthetically pleasing food photos from others is an important visual analysis task for social media and ranking systems related to food. Nevertheless, aesthetic assessment of food images remains a challenging and relatively unexplored task, largely due to the lack of related food image datasets and practical knowledge. Thus, we present the Gourmet Photography Dataset (GPD), the first large-scale dataset for aesthetic assessment of food photos. It contains 24,000 images with corresponding binary aesthetic labels, covering a large variety of foods and scenes. We also provide a non-stationary regularization method to combat over-fitting and enhance the ability of tuned models to generalize. Quantitative results from extensive experiments, including a generalization ability test, verify that neural networks trained on the GPD achieve comparable performance to human experts on the task of aesthetic assessment. We reveal several valuable findings to support further research and applications related to visual aesthetic analysis of food images. To encourage further research, we have made the GPD publicly available at <https://github.com/Openning07/GPA>.

Keywords image aesthetic assessment; food image analysis; dataset; regularization

1 Introduction

Food is one of the most fundamental entities in our daily life. A great food photograph can convey feelings of warmth, awaken fond memories, conjure up fantasies, or just simply make you hungry [1]. It can also draw crowds flocking to a new restaurant or boost the sales of a food magazine. Thus, the ability to assess the aesthetic quality of food images plays an important role in various applications, such as food photo recommendation (see Fig. 1(a)), food photography assistance, and enhancement (Fig. 1(b)).

Human beings can easily gauge the visual aesthetics of food photos. However, it remains challenging for artificial intelligent agents to do so. During the past two decades, many researchers have considered various related fields, such as image aesthetic assessment [2–4] and food image analysis [5–7]. Some have already explored aesthetic assessment of food

Fig. 1 (a) When you browse photos of *strawberry cake*, it would be nice if photos are sorted by visual aesthetics, e.g., aesthetically negative (*top*) and positive ones (*bottom*). (b) Responses from 247 users on two questions: (*Left*) How often do you take food photos? (*Right*) Do you want better food photos?

1 Youtu Lab, Tencent, Shanghai 200233, China. E-mail: saulsheng@tencent.com.

2 NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: W. Dong, weiming.dong@ia.ac.cn (✉); B.-G. Hu, hubg@nlpr.ia.ac.cn.

3 Kuaishou Technology, Beijing 100085, China. E-mail: H. Huang, huanghaibin03@kuaishou.com; C. Ma, chongyangma@kuaishou.com.

4 Snap Inc., Santa Monica, 90405, USA. E-mail: mchai@snap.com.

5 AI Lab, Tencent Inc., Shenzhen 518000, China. E-mail: norriszhang@tencent.com.

Manuscript received: 2020-06-09; accepted: 2020-08-25

images [8], but they resorted to hand-crafted visual features and did not perform quantitative studies on a large-scale dataset.

To endow intelligent agents with the ability to assess food image aesthetics, there are two major challenges to solve. Firstly, while there are some aesthetic image datasets [2, 9] and food categorization benchmarks [5, 6, 10], no dataset is available for learning or evaluation of food image aesthetic assessment. Without a reliable dataset, we cannot investigate the topic quantitatively and provide scientific observations or insights. Secondly, prior knowledge is lacking in how to perform the task effectively. Any model needs to be regularized so that it generalizes well to unseen food images, from other sources than any training data.

To address the above two challenges, in this paper, we present the Gourmet Photography Dataset (GPD), containing 24,000 food images with corresponding binary aesthetic annotations. We have conducted a series of experiments with popular learning mechanisms for visual analysis tasks to verify the annotation quality of the GPD. We have also devised a non-statistical, effective regularization method, adaptive smoothing regularization (ASR), to combat over-fitting, to provide better generalization and better performance. We have quantitatively assessed the generalization abilities of optimized models on unseen food images. Extensive experiments in Section 5 demonstrate that the GPD provides practical help in tuning CNN models to predict important visual patterns allowing assessment of food aesthetics and to realize effective food photo aesthetic assessment. The proposed regularization strategy outperforms several common counterparts in the task of image aesthetic assessment. All these findings encourage further development in related applications of food image aesthetic assessment.

In summary, our contributions are as follows:

- the GPD, the first large-scale dataset to support research into aesthetic visual assessment of food images;
- a simple yet effective approach to properly regularize neural networks for enhanced generalization ability and better performance;
- a system with promising performance for the task of food image aesthetic assessment, which demonstrates good generalization ability.

We also provide practical knowledge for further research.

A preliminary extended abstract of this work appeared at SIGGRAPH Asia 2018 [11]. The code and dataset can be downloaded from <https://github.com/Openning07/GPA>.

2 Related work

2.1 Image aesthetic assessment

The goal of image aesthetic assessment is to gauge the aesthetics of input images; it has been extensively studied over the past decade. Early works on image aesthetic assessment resort to hand-crafted features [2, 12, 13]. Recently, thanks to large-scale datasets [2, 9, 14], convolutional neural networks (CNN) with effective learning mechanisms have been able to outperform their hand-crafted counterparts. Advanced methods have been developed, such as order-less multi-patch aggregation [15], aesthetic attribute graphs with adaptive patch selection [3], an Earth mover's distance based loss function [16], an attention-based learning scheme [17], visual feature aggregation [18], a semi-supervised deep active learning-based model [19], and multi-level pooling [4]. Progress in this topic has encouraged many aesthetic-aware applications (e.g., see Table 1). In this paper, we investigate the aesthetic assessment of food photos, which is an under-developed specific image domain with huge practical commercial value.

2.2 Food image analysis

There is an increasing amount of research into food images, because of its high value in commercial visual marketing. Many advanced methods and benchmarks

Table 1 Brief information about several datasets related to aesthetic visual analysis

Dataset	Amount	Domain
CUHK-PQ [9]	~17k	general image
AVA [2]	~250k	general image
AesCHN [20]	1k	Chinese handwriting
AutoTriage [21]	~16k	general image
AADB [14]	10k	general image
PCCD [22]	~4k	photo captioning
BlendPhotos [23]	1305	image blending
AesClothing [24]	—	clothing recommendation
GPD (ours)	24k	food aesthetics

have been proposed, e.g., for food categorization [6, 10, 25]), recipe retrieval [7], and calorie estimation [26]. In this paper, unlike previous literature, we attempt to investigate the possibility of *aesthetic visual analysis* of food images, with great potential value in visual commercial marketing.

2.3 Regularization method

In machine learning, regularization is intended to diminish generalization error, instead of training error. Developing effective regularization methods has always been a major research topic. Effective approaches include: softmax with temperature [27], label smoothing regularization (LSR) [28], dropout [29], data augmentation [30], etc. In this work, we propose an effective regularization strategy for image aesthetic assessment.

3 Our dataset

There were three steps to establishing the Gourmet Photography Dataset: food image collection, aesthetic label annotation, and inter-human agreement.

3.1 Food image collection

To learn how to assess food images as aesthetically positive or negative, we should aim for high variety during image collection, with respect to categories, viewpoint, lighting conditions, and layout. We collected food photos from the Internet and existing food categorization benchmarks. Firstly, we downloaded food images from four popular online communities: Flickr, Pinterest, 500px, and Pexels, using various food keywords (e.g., cakes, drinks, seafood) and regional cuisine indicators (e.g., Chinese, French, Mexican). We also retrieved images from various food categorization datasets [5, 10] in a class-balanced manner to enrich data complexity. In this way, we collected a rich variety of images with varying complexity. After collection, we removed irrelevant instances, such as duplicated images, collages, and photos with observable artificial additions. We also conducted additional pre-processing operations to provide a meaningful training signal, such as removing unnecessary image borders and rotation calibration.

3.2 Aesthetic label annotation

Following existing literature [2–4, 15], we treat visual aesthetic assessment of food photos as a binary image

classification problem. Given N pairs $\{\mathbf{I}_i, \hat{y}_i\}_{i=1}^N$, $\hat{y}_i$ is the aesthetic label for image $\mathbf{I}_i$, where $\hat{y}_i \in \{0, 1\}$ denotes negative or positive.

Figure 3 illustrates the annotation procedure used to provide binary aesthetic labels for images; Amazon's Mechanical Turk (AMT) was used. Workers were asked to judge whether a displayed image looked aesthetically pleasing. Some food images are aesthetically ambiguous, leading annotators to spend much time to provide an answer with low confidence. To mitigate this issue and ease their anxiety over such images, workers were allowed to *skip* images for which they could not confidently provide answers. Ensuring high confidence answers is crucial to limiting time consumption and guaranteeing that labels contain meaningful cues, such as personal or cultural preferences, with high recall ratio. Images that were skipped three times or labeled validly are not reissued to further workers. Moreover, each worker was allowed to annotate 3000 images at most, to avoid allowing a few annotators to dominate the aesthetic perception of the dataset. Overall, 57 workers participated in the annotation procedure. We obtained 29,042 valid image-aesthetic label pairs, with 2647 photos skipped.

3.3 Inter-human agreement

To ensure high-quality aesthetic labels, we removed controversial labels where possible. Eight additional expert photographers with good aesthetic taste were invited to re-check the collected labels. For each image-label pair, they could agree or disagree with the label, based on the tips (e.g., lighting, colour, quality) in Fig. 2. If more than four experts agreed, the annotation was kept; otherwise, the label was regarded as ambiguous and discarded. During this process, 5042 instances were eliminated due to potential controversy. Most of those annotations come from a few AMT workers, who were perhaps unqualified for the task.

The results form the GPD: 24,000 food images with corresponding aesthetic labels, 13,088 positive and 10,912 negative. We show some instances in Fig. 2. For simplicity, in the following experiments, we randomly divided the GPD into two partitions: 21,600 (11,779 positive/9821 negative) images for training and the remainder for testing.

Fig. 2 (Left) Instances in the proposed GPD. (Right) Some tips of good-looking food photos, indicating the important factors in assessing aesthetic quality of food images.

Fig. 3 Our annotation process to collect aesthetic annotations, which allows workers to skip assessing photos of ambiguous visual aesthetics for high confidence answers.

4 Methodology

4.1 Preliminary

Given a training set $\{I_i, \hat{y}_i\}_{i=1}^N$ of N image-aesthetic label pairs, we cast the aesthetic assessment as a binary classification problem and apply the cross-entropy loss:

$$\begin{aligned} \theta^* &= \arg \min_{\theta} - \sum_{i=1}^N \log \Pr(\tilde{y} = \hat{y} \mid I_i, \theta) \\ &= \arg \min_{\theta_E, \theta_C} - \sum_{i=1}^N \log g(f(I_i, \theta_E), \hat{y}, \theta_C) \end{aligned} \quad (1)$$

where $\tilde{y}_i$ is the label predicted by the model given an input image patch I_i , and θ is the trainable parameter of the model $\Pr(\cdot \mid \cdot, \theta)$, which contains the feature extractor $f(\cdot, \theta_E)$ and the classifier $g(\cdot, \theta_C)$. We apply the softmax function $\sigma(z_i) : \mathbb{R}^C \rightarrow (0, 1)^C$, to calculate the confidence for the aesthetically positive

class and the negative one, where

$$\sigma(z)_i = \frac{\exp(z_i/\tau)}{\sum_{j=1}^C \exp(z_j/\tau)} \quad (2)$$

z is a C -dimensional input vector ($C = 2$ in our case), and τ is the *temperature* parameter [27] to control the shape of the output probability distribution over different classes from the softmax (usually set to 1). When $\tau > 1$, the margin between the maximum logit and the others for each z_i will be diminished. Thus, the maximum confidence of aesthetic assessment reduce, relatively.

4.2 Motivation

As noted, we treat image aesthetic assessment as a binary classification problem. Arguably, the image-level aesthetic label cannot indicate differences in visual aesthetics for different local patches cropped from a single image. Without a proper strategy, we might overly penalize the negative class on patches from aesthetically positive images (see Fig. 4(right)). In other words, if we train the model naively, the model may assign full probability to the target aesthetic class for each input instance, leading to over-confidence. Given the distribution of aesthetic scores of images in the AVA benchmark [2] (Fig. 4(left)), it is improper to require the model to output a prediction with 100% confidence for every input image. It is also revealed in Ref. [31] that CNNs with ReLU activation function always yield high-confidence predictions far away from the training data. To ensure the generalization ability of optimized models, we need to handle the over-confidence issue properly.

Inspired by Laplace smoothing [32], which favors highlighting more certain examples while avoiding overly penalizing the others, we mitigate the over-

Fig. 4 (Left) The histogram of aesthetic score. (Right) Image level label cannot indicate the aesthetic difference between each local patches. Images come from the AVA dataset.

confidence issue by adaptively smoothing the shape of output probability. The intuitive motivation is that we diminish the output values of target classes by introducing K smoothing vectors in the last fully connected layer.

4.3 Adaptive smoothing regularization

We introduce K vectors in the last fully connected layer to implement ASR, i.e., θ_A . Figure 5 exhibits the core idea of ASR. Because $\exp(\mathbf{d}_j)$ is positive ($j \in \{1, \dots, K\}$), $\delta_{ASR}(\mathbf{z})_i$ ($i \in \{1, 2\}$), the confidence of binary visual aesthetics will reduce. In this way, we mitigate the over-confidence and promote generalization ability. Put these ideas together, the ASR method may be stated as

$$\begin{cases} \sigma_{ASR}(\mathbf{z})_i = \frac{\exp(\mathbf{z}_i)}{\sum_{j=1}^C \exp(\mathbf{z}_j) + \sum_{j=1}^K \exp(\mathbf{d}_j)} \\ \exp(\mathbf{z}_i) = \exp(f^T(\mathbf{I}, \theta_E) \cdot \theta_C^i) \\ \exp(\mathbf{d}_j) = \exp(f^T(\mathbf{I}, \theta_E) \cdot \theta_A^j) \end{cases} \quad (3)$$

The corresponding derivatives are

$$\frac{\partial \sigma_{ASR}(\mathbf{z}_i)}{\partial \mathbf{z}_i} = \sigma_{ASR}(\mathbf{z})_i - \sigma_{ASR}^2(\mathbf{z})_i \quad (4)$$

$$\frac{\partial \sigma_{ASR}(\mathbf{z}_i)}{\partial \mathbf{d}_j} = -\frac{\sigma_{ASR}^2(\mathbf{z})_i}{\exp(\mathbf{z}_i)} \quad (5)$$

We further note that the K introduced vectors can also help maintain the pace of optimization, as lower target confidence always means higher error and strengthens the training signal for back-propagation, according to Eq. (4).

We find that introducing K vectors is not enough, because their softmax output values decrease quickly as optimization progresses. To maintain the smoothing effect, we propose three procedures:

- Randomly select K vectors from the model pre-trained on ImageNet [33];

Fig. 5 The schematic illustration of the proposed ASR method.

- Every $\text{Freq}_{\text{comp}}$ training epochs (5 in our experiments), we pick the representation vectors of patches whose target class prediction confidence is relatively low (< 0.5), and use their mean vectors to update the K classes.
- We update θ_A incrementally ($\lambda = 0.3$ in this paper).

$$\theta_A = \lambda \theta_A + (1 - \lambda) \theta'_A$$

We give the main algorithm in Algorithm 1.

Algorithm 1 Adaptive smoothing regularization

Input:

Ep_{max} : maximum number of training epochs;

Fq_{comp} : frequency to maintain the K vectors;

τ_{th} : threshold to select features;

- 1: Initialize network θ_A and θ_C ; Let $\text{Ep}_{\text{idx}} = 1$;
 - 2: **while** $\text{Ep}_{\text{idx}} \leq \text{Ep}_{\text{max}}$ **do**
 - 3: Optimize θ_C and θ_A using Eq.(1);
 - 4: **if** $\text{mod}(\text{Ep}_{\text{idx}}, \text{Fq}_{\text{comp}}) = 0$ **then**
 - 5: Evaluate the target class confidences and extract visual features on the validation partition;
 - 6: Select the feature vector whose target class confidences are around τ_{th} ; // useful features
 - 7: Conduct DBSCAN clustering on the features to remove outliers and get K centroids, θ'_A ;
 - 8: Use the centroids to update θ_A to strengthen the regularization effect;
 - 9: **end if**
 - 10: $\text{Ep}_{\text{idx}} + 1$;
 - 11: **end while**
 - 12: Return θ_C .
-

Discussion. Compared to existing regularization methods, such as LSR [28], confidence penalty, or data augmentation [30, 34], the proposed method does not make the transformation-invariant assumption or assign some pre-defined values to curtail output confidences. Instead, we apply K introduced vectors to smooth the output space in a non-statistical manner. Moreover, the introduced vectors change flexibly during the training process. Thus, the proposed ASR is more flexible and reasonable.

4.4 Implementation details

We apply the SGD algorithm using a batch size of 32, with Nesterov momentum of 0.9 and weight decay of $5e^{-4}$. We begin with a learning rate of $1e^{-3}$, drop it by a factor of 0.1 after every 10 epochs, and keep it at $1e^{-5}$ after 20 epochs. We set $K = 2$ in Eq. (3) during the experiments, as we believe that

one introduced vector per aesthetic class suffices to make the regularization effect work, as we will see in Section 5. Using a single NVIDIA Titan X GPU, the learning process takes about 17 hours to finish all 40 epochs. We implemented the method with Tensorflow.

Without loss of generality, we adopted the 18-layer ResNet [35] (ResNet-18) model as the backbone network, using which we compared the results of different regularization strategies (see the last several rows of Tables 2–4).

5 Experiments

5.1 Baseline comparison

To ensure reliable results and a systematic evaluation, we applied several typical vision learning algorithms to the GPD. Their performance, in terms of percentage assessment accuracy, helps us explore the possibility of visual aesthetic assessment on the food images and assess the quality of aesthetic labels in the GPD.

5.1.1 SVM with color

Color information proves to be important in image aesthetic assessment [12]. We encoded color information for images as color histogram features, with 128 bins for RGB color channels. Zero-mean-unit-variance normalization was conducted before optimization.

5.1.2 SVM with GIST

GIST features [36] are another typical approach used to capture the global content of images. We extracted 512-dimensional gray-scale GIST features with an image size of 256×256 . Zero-mean-unit-variance normalization was also performed as a preprocessing step to facilitate the following optimization process.

5.1.3 SVM with VGG features

We extracted 4096-dimensional features from the penultimate layer of a 16-layer VGG (VGG-16) model [37]. For a full comparison, we experimented with VGG models with three typical semantics: VGG-object, VGG-scene, and VGG-food, trained on ImageNet [33], Places [38], and a food categorization dataset combining [5, 10], respectively.

5.1.4 GPD-supervised CNNs

We experiment with several typical CNNs for image classification task: AlexNet [30], VGG-16 [37],

InceptionV2 [34], and 18-layer ResNet (ResNet-18) [35]. To facilitate optimization, all the neural networks are initialized on ImageNet dataset [33]. As usual, we adopt the *cross-entropy* loss to optimize these neural networks.

5.1.5 Implementation details

For training, photos were re-scaled with respect to the shortest edge (259 for AlexNet and 256 for the others), and then patches (227×227 for AlexNet and 224×224 for the others) were randomly cropped. Random horizontal mirroring (0.5) was conducted for data augmentation. To maximize the performance of each model, we applied different training hyperparameters, such as batch-size or learning rate for CNNs and cost coefficient for SVM, via cross-validation experiments. For inference, we report and compare the average on 10 patches randomly cropped from input photos.

5.2 Evaluation on GPD

The results of the aforementioned methods on the GPD are listed in Table 2. Several typical aesthetically negative and positive instances are shown in Fig. 6. Figure 7 shows the histogram of confidences predicted by ResNet-18 on the training portion of the GPD. We also experimented with MP_{ada} [17], a state-of-the-art approach for image

Table 2 Training and test accuracy (%) of different machine learning algorithms and visual features on GPD

Solution	Training	Test
SVM classifier		
SVM + Color	72.4	63.3
SVM + GIST	78.1	64.4
SVM + VGG-object	90.8	74.7
SVM + VGG-scenes	86.8	72.4
SVM + VGG-foods	90.4	74.1
Vanilla CNNs		
AlexNet	89.1	88.6
VGG-16	90.6	87.2
InceptionV2	94.0	90.1
ResNet-18	93.3	89.7
CNNs for aesthetic assessment		
MP_{ada} [17]	94.6	90.4
ResNet-18 with regularization		
ResNet-18 + aug	93.6	89.9
ResNet-18 + LSR [28]	95.6	90.2
ResNet-18 + σ_T [27]	94.1	89.4
ResNet-18 + ASR (ours)	95.0	90.7

Fig. 6 Typical negative (left) and positive (right) photos found by our method. These images cover a wide variety of patterns and can be used in some food photography teaching class or online food photography assistance.

aesthetic assessment, using the authors' code from <https://github.com/Openning07/MPADA>. From the results, we obtained the following findings:

- *The scale of the GPD seems to be sufficient to support training learning algorithms.* The gaps between training performance and testing results are generally less than 10% in each row of Table 2. It should be noted that we do not adopt any data augmentation or complex training tricks. We also conducted the same experiments with different partitions of data, i.e., 10-fold cross-validation on the whole GPD, and obtained results close to those in Table 2. We may conclude that the

Fig. 7 The distribution of aesthetic scores on unseen food images from the ResNet-18 model trained on the proposed GDP.

results demonstrate the effectiveness of the GPD for assessing visual aesthetics of food photos.

- *Our regularization module outperforms other approaches.* Comparing the results for ResNet models with different regularization strategies, our proposed ASR works better than other regularization strategies. We further observe that during the inference stage, the confidence values for positive/negative visual aesthetics are always larger than those for the introduced K vectors. This shows that that our regularization method works for the task of binary aesthetic visual assessment.
- *GPD-supervised CNNs achieve the best testing results amongst the tested learning mechanisms.* Further, SVM with VGG features generally outperforms SVM with hand-crafted features. These findings of the effectiveness of CNNs are not new and are consistent with mainstream conclusions from the computer vision community [39]; they make effective image aesthetic assessment of food photos possible. Another interesting observation is that the visual features from VGG-scenes do not work as well as those from VGG-foods or VGG-objects. This demonstrates the importance of object semantics and food semantics in assessing visual aesthetics of food photos.

5.3 Generalization ability test

To test the generalization ability of tuned models, we collected 825 unseen food photos from *WeChat*, one of the largest online communities. We then invited 50 qualified candidates to give their opinions as to whether the photos looked aesthetically pleasing or not. Based on the 41,250 responses, we measured consistency of the aesthetic assessments from the models with human perception via the following equation:

$$V(S_c) = \frac{100}{|S_c|} \sum_{I \in S_c} \frac{\text{vote}_I^c}{U}, \quad c \in \{\text{pos}, \text{neg}\}$$

where S_{pos} , S_{neg} denote which images are positive or negative as predicted by the model, $\text{vote}_I^{\text{pos}}$, $\text{vote}_I^{\text{neg}}$ indicate the number of votes from reviewers who believe the image I to be positive or negative, 100 is the normalization factor for percentage, and U indicates the number of candidates ($U = 50$ in our experiments). The results are listed in Table 3; we calculate the *best* / *worse* results in a greedy manner. To better visualize the comparisons, we show some images together with aesthetic assessment results from different methods in Fig. 8. Based on these results, we draw some empirical conclusions

Table 3 The results of generalization ability test on several approaches via food photos collected from WeChat, which is different from the sources where the images of GPD come from.

Solution	$V(S_{\text{pos}})$	$V(S_{\text{neg}})$
Bounds		
Best	75.5	83.9
Worst	16.1	24.5
Random	37.3	62.5
Human Expert	72.1	81.0
SVM classifier		
SVM + Color	38.8	65.9
SVM + GIST	42.0	67.7
SVM + VGG-object	66.5	68.9
SVM + VGG-scene	65.5	69.2
SVM + VGG-food	65.2	71.0
Vanilla CNNs		
AlexNet	63.5	71.6
VGG-16	65.4	71.9
InceptionV2	67.7	72.3
ResNet-18	61.1	72.5
ResNet-18 with regularization		
ResNet-18 + AVA	38.5	65.7
ResNet-18 + aug	71.0	71.5
ResNet-18 + LSR [28]	61.3	72.4
ResNet-18 + σ_T [27]	61.1	72.5
ResNet-18 + ASR (ours)	73.5	72.1

as follows:

- *Human experts achieve results close to the theoretical best (75.5, 83.9), and significantly better than random.* These observations indicate that a good model for image aesthetic assessment on food photos should be able to generalize well, like human experts. They also indicate that the 825 food images with collected responses from reviewers can be used to test model generalization ability.
- *GPD-supervised CNNs possess good generalization ability in assessing visual aesthetics of food photos.* The aesthetic assessments of food photos from GPD-supervised CNNs are consistent with those of human experts. For positive aesthetics, ResNet-18 with ASR even outperforms human experts in the experiment. This shows that neural networks tuned on the proposed GPD dataset possess good generalization ability in assessing visual aesthetics of food photos. Consequently, these results demonstrate the validity and utility of the GPD and the proposed regularization method in the task of aesthetic visual assessment on food photos (e.g., food image triage or recommendation).
- *Negative food visual aesthetics seem to be easier to assess than positive ones.* We have more food images with positive aesthetics than ones with negative aesthetics in the GPD, whereas the tuned models consistently achieve better results for negative cases than positive ones. Supportive cues arise from the observations that $V(S_{\text{neg}})$ is generally higher than $V(S_{\text{pos}})$ across each row of Table 3. These results indicate that, on assessing the visual aesthetics of food photos, people achieve consensus on negative visual aesthetics more often than on positive ones, if they are forced to make a judgment. This insight provide some useful guidelines for further developments. For example, we need more training instances of positive visual aesthetics, and we should take personal preferences into consideration when offering certain services related to positive visual aesthetics.

5.4 Further investigation into K

To further investigate how the hyper-parameter K in Eq. (3) influences regularization and the final aesthetic assessment accuracy, we conducted

Fig. 8 Predictions from humans, SVM with hand-crafted / VGG-based features, and deep convolutional neural networks trained on the AVA benchmark [2] (AVA-Networks) and on the proposed GPD dataset (GPD-Networks).

additional experiments with different values of K on the GPD. To compute accurate statistics, we conducted 10-fold cross-validation; the results are exhibited in Fig. 9. It can be seen that $K = 2$ is generally a good choice, in comparison to other options. On the other hand, we note that improperly introducing K smoothing vectors can hurt the assessment accuracy.

It is also interesting to note that, when we shift the backbone from ResNet-18 to InceptionV2 or AlexNet, sometimes the K smoothing vectors output confidences larger than the two main classes. Further work is needed to investigate the underlying mechanism to exploit ASR better.

5.5 Additional experiments on AVA

To make the proposed ASR more convincing, we conducted additional comparative experiments on the large-scale AVA benchmark [2], which is widely

used [3, 15, 16, 40]. Without loss of generality, we experimented with ResNet-18 [35] models using common regularization strategies, followed by the common pipeline on the AVA dataset [3, 15, 16, 40]: e.g., we used 5.0 as the threshold value for binary aesthetic assessment labels, 230k images for training and the remaining 20k for testing.

The accuracy comparison is shown in Table 4. Our regularization strategy achieves comparable results to the *state-of-the-art* method [17] and outperforms other regularization methods. We do not claim superior results, but simply verify the effectiveness of our proposed ASR method, which works differently to existing methods [3, 4, 15–17]. Consequently, we further verify the effectiveness of the proposed regularized softmax in the task of image aesthetic assessment.

Table 4 The classification accuracy (%) of ResNet-18 models with various regularization strategies and other typical approaches on the AVA dataset

Solution	Accuracy
Handcrafted features [2]	68.0
DMA-Net-IF [15]	74.5
MNA-CNN-Scene [40]	77.1
Reg+Rank+Att+Cont [14]	77.3
NIMA [16]	81.5
Pool-3FC [4]	81.7
GPF-CNN [41]	81.81
CombFeatures [18]	81.95
A-Lamp [3]	82.5
MP_{ada} [17]	83.0
ResNet-18	81.8
ResNet-18 + aug	80.9
ResNet-18 + LSR [28]	82.5
ResNet-18 + σ_T [27]	82.3
ResNet-18 + ASR (ours)	83.3

Fig. 9 The image aesthetic assessment accuracy goes with K in Eq. (3) on the test partition of the GPD dataset.

5.6 Assistance in food image generation

5.6.1 Motivation

Another practical use for the ability to assess visual aesthetics of food image is to diminish bad instances (e.g., with observable artifacts) generated by CNN models. Currently, many researchers are working on image generation or enhancement [23, 42, 43]. However, the lack of effective methods to distinguish low-quality outputs impedes practical application of such methods.

5.6.2 Approach

With an aesthetic assessment model with good generalization ability, we gauged the aesthetic scores of original images and output ones, and then selected outputs with relatively high aesthetic scores or with moderate degraded score. This process is inspired by a refinement-based-on-evaluation procedure instead of manual annotation, and is akin to web-supervised learning [44] and evaluation without ground truth [45]. Intuitively, the generator model and the assessment model benefit each other in the long-term trend.

5.6.3 Results

We conducted an experiment to assist food image generation with pizzaGAN [46], a generative adversarial network (GAN) based model to generate pizza images conditioned by a pizza photo and a cooking instruction (e.g., add corn, or remove ham). All the original images and the manipulated results can be found at <http://pizzagan.csail.mit.edu/#>.

The results of aesthetic assessment on the original food images and the manipulated ones are shown in Fig. 10. With the procedure described above, we can distinguish good results from low-quality ones without the need for a time-consuming user study. In this way, we can discard improper output from generative models, making related applications on food images more practical.

6 Conclusions

To support research into food image aesthetic assessment, this work presents the GPD, the first related complex, large-scale dataset with corresponding binary aesthetic labels. To combat over-confidence, we have given a simple yet effective regularization strategy, ASR, which can improve the generalization ability of optimized CNN models. Extensive experiments with several typical machine learning approaches demonstrate that the proposed GPD can provide valuable help, enabling computer vision models to predict visual aesthetic of food photos. Furthermore, the proposed regularization strategy is better than alternatives in helping CNN models to achieve generalization, on the GPD and the AVA. Even on unseen food photos, CNN models trained on the GPD and armed with the proposed ASR perform comparably with human experts in assessing visual aesthetics of food photos. All these empirical findings should encourage further research and practical applications related to aesthetic visual

Fig. 10 With the help of GPD-ResNet, we pick good cases (*top column*) generated from pizzaGAN model and avoid the outputs with observable improper artifacts (*bottom column*). *Original* means the input image to pizzaGAN, and *Add/Remove xxx* mean conduct one certain cooking operation with specific ingredient xxx (e.g., pepperoni, arugula) on the original image.

analysis of food images.

For future work, we hope to expand the scale of the GPD and enrich its attributes such as viewing angle, layout, and scenes. We also hope to exploit the proposed dataset to further facilitate related applications in the specific domain of food images.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61832016 and 61672520, and by a CASIA–Tencent YouTu joint research project.

References

- [1] Manna, L. Digital food photography. Cengage Learning PTR, 2015.
- [2] Murray, N.; Marchesotti, L.; Perronnin, F. Ava: A large-scale database for aesthetic visual analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2408–2415, 2012.
- [3] Ma, S.; Liu, J.; Chen, C. W. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 722–731, 2017.
- [4] Hosu, V.; Goldlücke, B.; Saupe, D. Effective aesthetics prediction with multi-level spatially pooled features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9367–9375, 2019.
- [5] Bossard, L.; Guillaumin, M.; van Gool, L. Food-101—mining discriminative components with random forests. In: *Computer Vision—ECCV 2014. Lecture Notes in Computer Science, Vol. 8694*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 446–461, 2014.
- [6] Zhang, X. J.; Lu, Y. F.; Zhang, S. H. Multi-task learning for food identification and analysis with deep convolutional neural networks. *Journal of Computer Science and Technology* Vol. 31, No. 3, 489–500, 2016.
- [7] Salvador, A.; Hynes, N.; Aytar, Y.; Marin, J.; Oi, F.; Weber, I.; Torralba, A. Learning cross-modal embeddings for cooking recipes and food images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3068–3076, 2017.
- [8] Li, Y.; Sheopuri, A. Applying image analysis to assess food aesthetics and uniqueness. In: Proceedings of the IEEE International Conference on Image Processing, 311–314, 2015.
- [9] Luo, W.; Wang, X.; Tang, X. Content-based photo quality assessment. In: Proceedings of the IEEE International Conference on Computer Vision, 2206–2213, 2011.
- [10] Chen, X.; Zhu, Y.; Zhou, H.; Diao, L.; Wang, D. ChineseFoodNet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743*, 2017.
- [11] Sheng, K. K.; Dong, W. M.; Huang, H. B.; Ma, C. Y.; Hu, B. G. Gourmet photography dataset for aesthetic assessment of food images. In: Proceedings of the SIGGRAPH Asia 2018 Technical Briefs, Article No. 20, 2018.
- [12] Datta, R.; Joshi, D.; Li, J.; Wang, J. Z. Studying aesthetics in photographic images using a computational approach. In: *Computer Vision—ECCV 2006. Lecture Notes in Computer Science, Vol. 3953*. Leonardis, A.; Bischof, H.; Pinz, A. Eds. Springer Berlin Heidelberg, 288–301, 2006.
- [13] Zhang, F. L.; Wang, M.; Hu, S. M. Aesthetic image enhancement by dependence-aware object recomposition. *IEEE Transactions on Multimedia* Vol. 15, No. 7, 1480–1490, 2013.
- [14] Kong, S.; Shen, X. H.; Lin, Z.; Mech, R.; Fowlkes, C. Photo aesthetics ranking network with attributes and content adaptation. In: *Computer Vision—ECCV 2016. Lecture Notes in Computer Science, Vol. 9905*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 662–679, 2016.
- [15] Lu, X.; Lin, Z.; Shen, X.; Mech, R.; Wang, J. Z. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In: Proceedings of the IEEE International Conference on Computer Vision, 990–998, 2015.
- [16] Talebi, H.; Milanfar, P. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* Vol. 27, No. 8, 3998–4011, 2018.
- [17] Sheng, K. K.; Dong, W. M.; Ma, C. Y.; Mei, X.; Huang, F. Y.; Hu, B. G. Attention-based multi-patch aggregation for image aesthetic assessment. In: Proceedings of the 26th ACM International Conference on Multimedia, 879–886, 2018.
- [18] Kucer, M.; Loui, A. C.; Messinger, D. W. Leveraging expert feature knowledge for predicting image aesthetics. *IEEE Transactions on Image Processing* Vol. 27, No. 10, 5100–5112, 2018.
- [19] Liu, Z. G.; Wang, Z. P.; Yao, Y. Y.; Zhang, L. M.; Shao, L. Deep active learning with contaminated tags for image aesthetics assessment. *IEEE Transactions on Image Processing* doi: 10.1109/TIP.2018.2828326, 2018.

- [20] Sun, R.; Lian, Z.; Tang, Y.; Xiao, J. Aesthetic visual quality evaluation of Chinese handwritings. In: Proceedings of the International Joint Conferences on Artificial Intelligence, 2510–2516, 2015.
- [21] Chang, H. W.; Yu, F.; Wang, J.; Ashley, D.; Finkelstein, A. Automatic triage for a photo series. *ACM Transactions on Graphics* Vol. 35, No. 4, Article No. 148, 2016.
- [22] Chang, K.-Y.; Lu, K.-H.; Chen, C.-S. Aesthetic critiques generation for photos. In: Proceedings of the IEEE International Conference on Computer Vision, 3514–3523, 2017.
- [23] Hung, W.-C.; Zhang, J.; Shen, X.; Lin, Z.; Lee, J.-Y.; Yang, M.-H. Learning to blend photos. In: Proceedings of the European Conference on Computer Vision, 70–86, 2018.
- [24] Yu, W. H.; Zhang, H. D.; He, X. N.; Chen, X.; Xiong, L.; Qin, Z. Aesthetic-based clothing recommendation. In: Proceedings of the World Wide Web Conference, 649–658, 2018.
- [25] Hassannejad, H.; Matrella, G.; Ciampolini, P.; de Munari, I.; Mordonini, M.; Cagnoni, S. Food image recognition using very deep convolutional networks. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, 41–49, 2016.
- [26] Meyers, A.; Johnston, N.; Rathod, V.; Korattikara, A.; Gorban, A.; Silberman, N.; Guadarrama, S.; Papandreou, G.; Huang, J.; Murphy, K. P. Im2Calories: Towards an automated mobile vision food diary. In: Proceedings of the IEEE International Conference on Computer Vision, 1233–1241, 2015.
- [27] Hinton, G. E.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv preprint* arXiv:1503.02531, 2014.
- [28] Szegedy, C.; Vanhoucke, V.; Iofie, S.; Shlens, J.; Z. Wojna. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826, 2016.
- [29] Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* Vol. 15, No. 1, 1929–1958, 2014.
- [30] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* Vol. 60, No. 6, 84–90, 2017.
- [31] Hein, M.; Andriushchenko, M.; Bitterwolf, J. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 41–50, 2019.
- [32] Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [33] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 248–255, 2009.
- [34] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9, 2015.
- [35] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016.
- [36] Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* Vol. 42, No. 3, 145–175, 2001.
- [37] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556v6, 2015.
- [38] Zhou, B. L.; Lapedriza, A.; Khosla, A.; Oliva, A.; Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 6, 1452–1464, 2018.
- [39] Zhang, R.; Efros, A. A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 586–595, 2018.
- [40] Mai, L.; Jin, H.; Liu, F. Composition-preserving deep photo aesthetics assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 497–506, 2016.
- [41] Zhang, X. D.; Gao, X. B.; Lu, W.; He, L. H. A gated peripheral-foveal convolutional neural network for unified image aesthetic prediction. *IEEE Transactions on Multimedia* Vol. 21, No. 11, 2815–2826, 2019.

- [42] Deng, Y.; Loy, C. C.; Tang, X. Aesthetic-driven image enhancement by adversarial learning. In: Proceedings of the 26th ACM International Conference on Multimedia, 870–878, 2018.
- [43] Hu, Y.; He, H.; Xu, C.; Wang, B.; Lin, S. Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics* Vol. 37, No. 2, Article No. 26, 2018.
- [44] Xu, Z.; Huang, S. L.; Zhang, Y.; Tao, D. C. Webly-supervised fine-grained visual categorization via deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 5, 1100–1113, 2018.
- [45] Sheng, K. K.; Dong, W. M.; Kong, Y.; Mei, X.; Li, J. L.; Wang, C. J.; Huang, F.; Hu, B. Evaluating the quality of face alignment without ground truth. *Computer Graphics Forum* Vol. 34, No. 7, 213–223, 2015.
- [46] Papadopoulos, D. P.; Tamaazousti, Y.; Oi, F.; Weber, I.; Torralba, A. How to make a pizza: Learning a compositional layer-based GAN model. In: proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8002–8011, 2019.

researcher engineer at Youtu Lab, Tencent Inc. His research interests include image quality evaluation, domain adaptation, and AutoML.

Kekai Sheng received his Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences in 2019. He received his B.Eng. degree in telecommunication engineering from the University of Science and Technology, Beijing in 2014. He is currently a

science in 2001 and 2004 from Tsinghua University. He received his Ph.D. degree in information technology from the University of Lorraine, France, in 2007. His research interests include visual media synthesis and evaluation. He is a member of the ACM and IEEE.

Weiming Dong is a professor in the Chinese–French Joint Laboratory for Computer Sciences, Control, and Applied Mathematics and the National Laboratory of Pattern Recognition at the Institute of Automation, Chinese Academy of Sciences. He received his B.Eng. and M.S. degrees in computer

Haibin Huang is a senior research scientist at Kuaishou Technology. He obtained his Ph.D. degree in computer science from UMass Amherst. He obtained his B.S. and an M.S. degrees in the Department of Mathematics, Zhejiang University. His research focuses on visual content analysis and creation.

Menglei Chai is a senior research scientist at Snap Inc. He received his Ph.D. and B.Eng. degrees in computer science from Zhejiang University in 2017 and 2011 respectively. His research interests are in computer vision and graphics, especially in photo manipulation and physics-based simulation.

Yong Zhang is a senior researcher in the Tencent AI Lab. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2018. He was supervised by Prof. Bao-Gang Hu and Prof. Weiming Dong at the National Laboratory of Pattern Recognition. He obtained his B.Eng degree in automation from Hunan University in 2012. His research is on computer vision and machine learning, particularly human facial behavior analysis, face recognition, and face synthesis.

Chongyang Ma received his B.S. degree in fundamental science (mathematics and physics) from Tsinghua University in 2007 and his Ph.D. degree in computer science from the Institute for Advanced Study of Tsinghua University in 2012. He is currently a research leader at Kuaishou Technology. His research interests include computer graphics and computer vision.

Bao-Gang Hu is a full professor at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. He received his M.S. degree from the University of Science and Technology, Beijing, China in 1983, and his Ph.D. degree from McMaster University, Canada in 1993, both in mechanical engineering. He worked as a lecturer in the University of Science and Technology,

Beijing, from 1983 to 1987. From 1994 to 1997, he was a research engineer and senior research engineer at C-CORE, the Memorial University of Newfoundland, Canada. From 2000 to 2005, he was the Chinese Director of the Chinese–French Joint Laboratory for Computer Science, Control and Applied Mathematics. He is Senior Member of the IEEE.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made.

The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Other papers from this open access journal are available free of charge from <http://www.springer.com/journal/41095>. To submit a manuscript, please go to <https://www.editorialmanager.com/cvmj>.