This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2911428, IEEE Transactions on Multimedia

A Gated Peripheral-Foveal Convolutional Neural Network for Unified Image Aesthetic Prediction

Xiaodan Zhang, Xinbo Gao, *Senior Member, IEEE,* Wen Lu, *Member, IEEE,* and Lihuo He, *Member, IEEE*

***Abstract*—Learning fine-grained details is a key issue in image aesthetic assessment. Most of the previous methods extract the fine-grained details via random cropping strategy, which may undermine the integrity of semantic information. Extensive studies show that humans perceive fine-grained details with a mixture of foveal vision and peripheral vision. Fovea has the highest possible visual acuity and is responsible for seeing the details. The peripheral vision is used for perceiving the broad spatial scene and selecting the attended regions for the fovea. Inspired by these observations, we propose a Gated Peripheral-Foveal Convolutional Neural Network (GPF-CNN). It is a dedicated double-subnet neural network, *i.e.* a peripheral subnet and a foveal subnet. The former aims to mimic the functions of peripheral vision to encode the holistic information and provide the attended regions. The latter aims to extract fine-grained features on these key regions. Considering that the peripheral vision and foveal vision play different roles in processing different visual stimuli, we further employ a Gated Information Fusion (GIF) network to weight their contributions. The weights are determined through the fully connected layers followed by a sigmoid function. We conduct comprehensive experiments on the standard AVA and Photo.net datasets for unified aesthetic prediction tasks: (i) aesthetic quality classification; (ii) aesthetic score regression; and (iii) aesthetic score distribution prediction. The experimental results demonstrate the effectiveness of the proposed method.**

***Index Terms*—Visual aesthetic quality assessment, image aesthetics, deep learning**

I. Introduction

Automatic image aesthetic assessment aims to endow computers with the ability of perceiving aesthetics as human beings. It plays an important role in many real-world applications, such as image recommendation, photo organization and image enhancement [1]–[4]. Early attempts in this area focus on handcrafted features which are based on the known aesthetic principles such as the rule-of-thirds, simplicity or diagonal rules [5]–[9]. But most photographic rules are too abstract to be modeled mathematically.

Deep learning methods have shown great success in various computer vision tasks [10]–[13]. More and more researchers try to apply deep learning methods to image aesthetic assessment [14]–[16]. But most of the networks ignore the fine-grained high resolution details, which is critical in aesthetic prediction. Previous study [17] tried to address this limitation by adding a local-subnet to the traditional CNN. The input of the local-subnet is represented by a randomly cropped patch which may undermine the integrity of semantic information. Recently, Lu *et al.* [18] proposed a Deep Multi-Patch Aggregation Network (DMA-Net) to extract local fine-grained features from multiple randomly cropped patches. This method achieves some promising results, but it ignores the global spatial layout information. Considering this, Ma *et al.* [19] proposed a layout-aware framework in which an attribute graph is added to DMA-Net. Whereas, the nodes of the attribute graph need to be predefined, which is not applicable in practical applications.

It is universally acknowledged that humans perceive scenes with a mixture of high-acuity foveal vision and coarser peripheral vision [20], [21]. The former has the highest density of cones, and is responsible for encoding fine-grained details. The latter contains a significantly lower density of cones, and is mainly used for encoding the broad spatial scene and seeing large objects [21], [22]. More importantly, peripheral vision also actively participates in attentional selection of visual space to be processed by fovea [23].

Considering the above observations, we mimic this process and develop a Gated Peripheral-Foveal Convolutional Neural Networks (GPF-CNN). As its name would suggest, it consists of two subnets, *i.e.* a peripheral subnet and a foveal subnet. The peripheral subnet mimics the functions of peripheral vision. It composes of a bottom-up feed-forward network to encode the global composition and a top-down neural attention feedback process to create an image-sized attention map. The input of the peripheral subnet is a downsampled low resolution image and is referred as peripheral view. Similarly with foveal vision, the proposed foveal subnet aims to extract fine grained high resolution details on the attended regions of the peripheral view. The input of the foveal subnet is a high-resolution image, and is denoted as the foveal view or simply fovea. Fig. 1 shows an example of the attention map, the peripheral and foveal view. The model selects a foveal window from the peripheral view with the guidance of top-down neural attention. The corresponding region from the high-resolution images is then cropped for extracting fine-grained details. Finally, features extracted in the fovea subnet are fused with features extracted in the peripheral subnet.

Recent studies show that foveal vision and peripheral vision play different roles in processing different visual stimuli [21],

Manuscript received December 1, 2018; revised March 8, 2019; accepted April 7, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61432014, 61772402, U1605252 and 61671339, in part by the National Key Research and Development Program of China under Grant 2016QY01W0200, and in part by National High-Level Talents Special Support Program of China under Grant CS31117200001.

X. Zhang, W. Lu and L. He are with the Video and Image Processing System Laboratory, School of Electronic Engineering, Xidian University, Xian 710071, China (email:xdanzhang@stu.xidian.edu.cn; luwen@mail.xidian.edu.cn; lihuo.he@gmail.com).

X. Gao is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xian 710071, China (e-mail: xbgao@mail.xidian.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2911428, IEEE Transactions on Multimedia

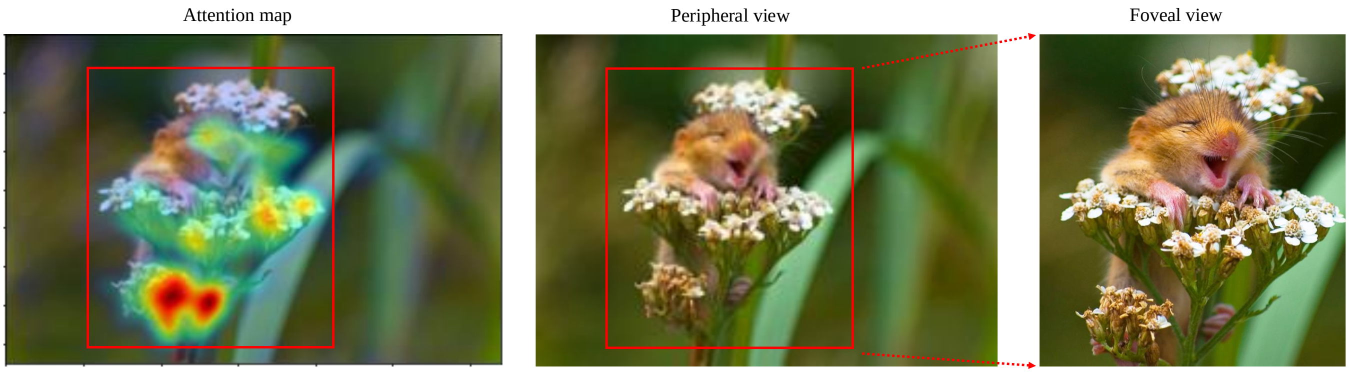

Fig. 1. Illustration of the top-down neural attention map (left), peripheral view (middle) and foveal views (right) of an image.

[24]. Categories such as portrait and animal rely more on fine-grained details information to make aesthetic decision. Thus they are associated more with fovea representations. Other categories, as in the case of landscape and architecture, rely more on global shape and large-scale integration. Therefore, they are associated more with peripheral representations. Motivated from these findings, we propose a gated information fusion network to weight the foveal and peripheral branch adaptively: if one branch is better at processing a given image, the gating layer will direct more information to that branch by increasing the value of that gating node.

Overall, this paper makes the following contributions.

- We propose a deep architecture that not only encodes the global composition but also automatically focuses on aesthetic-relevant regions to extract the fine-grained details.
- We also develop a gated information fusion module which can adaptively weight the contributions of the global layout and local fine-grained features according to the input. By combining the weighted global and local features, the proposed module can greatly boost the performance.
- We conduct comprehensive experiments for unified aesthetic prediction tasks: aesthetic classification, aesthetic regression and aesthetic label distribution. For all these tasks, the proposed model achieves superior performance over the state-of-the-art approaches on public datasets.

The remainder of this paper is organized as follows. In section II, we briefly summarize the related work. In section III, we introduce the architecture of the GPF-CNN model. In section IV, we quantitatively evaluate the effectiveness of the proposed model and compare it with state-of-the art methods. Finally, we wrap up with conclusions and ideas for future work in section V.

II. Related Work

Contemporary image aesthetic assessment research can be roughly outlined by the following two important components: extraction of more advanced features and utilization of more sophisticated learning algorithms. Thus, we summary the previous research from these two perspectives: the visual representations and the learning algorithms.

A. *Visual Representations*

There is a vast literature on the problem of designing effective features for aesthetic assessment, starting with the seminal work of [25] and leading to recent works of [6], [7], [9]. These features are based on the person's aesthetic perception and photographic rules. For example, Datta *et al.* [25] extracted 56 features to model the photographic technique such as rule of thirds, colorfulness, or saturation. Tang *et al.* [7] modeled the photographic rules (composition, lighting, and color arrangement) by extracting the visual features according to the variety of photo content. Nishiyama *et al.* [9] proposed to use the *bags-of-color-patterns* to model the color harmony in aesthetics. Later work proposed by Zhang *et al.* [26] focused on constructing the small-sized connected graphs to encode the image composition information. However, the above methods with hand-designed features can achieve only limited success because 1) such handcrafted features cannot be applied to all the image categories since the photographic rules vary considerably among different images. 2) these handcrafted features are heuristic and some photography rules are difficult to be quantified mathematically.

Recently, some researchers have tried to apply the deep learning networks to image aesthetic quality assessment. Tian *et al.* [14] proposed a query-dependent aesthetic model with deep learning for aesthetic quality assessment. Their method suffers deteriorate accuracy since they just use the networks as feature extractor. Kao *et al.* [27] explored the deep multi-task networks to leverage the semantic information to image aesthetic prediction. Different from the aforementioned methods, [17]–[19] focused on the fixed-size input constraint of deep networks when applied for aesthetic prediction. The inputs need to be transformed via scaling, cropping, or padding before feeding into the neural network. Images after these transformations often lose the holistic information and the high-resolution fine-grained details. Lu *et al.* [17] tried to tackle this problem by proposing a double column network called *RAPID*. In particular, they represented the global view via padded or warped image, the local view via the randomly cropped single patch. In order to capture more high resolution fine-grained details, Lu *et al.* [18] extended the *RAPID* to a deep multi-patch aggregation network (*DMA-Net*). In *DMA-*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2911428, IEEE Transactions on Multimedia

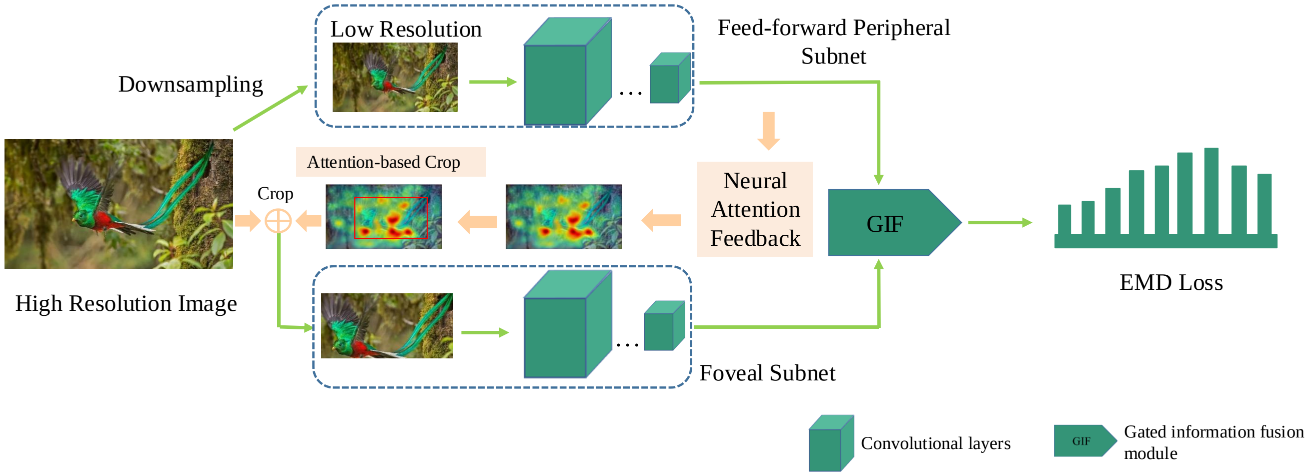

Fig. 2. Overall architecture of the GPF-CNN. The input of the peripheral subnet is low resolution image, and the input of foveal subnet is the selected attended region. The GIF module is used to balance the weights of the two subnets. More detailed illustrations of GIF module can be seen in Fig. 3.

Net, the input image was represented with a bag of random cropped patches. Two network layers (statistics and sorting) were used to aggregate the multiple patches. However, *DMA-Net* failed to encode the global layout of the image. Ma *et al.* [19] tried to address this limitation by adding an object-based attribute graph to *DMA-Net*. Their method relies on strong hypothesis. The number of attribute graph node need to be given in advance, which is unapplicable in most cases. Our work is also related to fusing the global and local features for aesthetic prediction. But we focus on extracting fine-grained details on the aesthetic-related regions and using a gated information fusion module to adaptively weight the global and local features according to the inputs.

B. Learning Algorithms

Early attempts in image aesthetic assessment cast this problem as a classification task [18], [19], [27]–[29]. They classified the images into high or low aesthetic quality based on the threshold of the weighted mean scores of human rating. Other research such as [15], [16] used the regression model to predict the aesthetic score. However, the image aesthetic quality assessment is highly subjective. The rated scores of different people may differ greatly due to the cultural background. Thus a scalar value is insufficient to provide the degree of consensus or diversity of opinion among annotators [30]. Considering this, recent research focuses on directly predicting the label distribution of the scores. In [30], Jin *et al.* proposed a new CJS loss to predict the aesthetic label distribution. Murray *et al.* [31] used the Huber loss to predict the aesthetic score distribution. But they predicted each discrete probability independently. Talebi *et al.* [32] treated the score distribution as ordered classes and used squared Earth Mover's Distance (EMD) loss to predict the score distributions. In this paper, similar with [32], we optimize our networks by minimizing EMD loss.

III. Gated Peripheral-Foveal Convolutional Neural Network

The proposed model includes two subnets: the peripheral subnet and the foveal subnet. Given a high resolution image, the image is first downsampled and then fed into the peripheral subnet. The peripheral subnet is responsible for encoding the global composition and providing the key region. Then, a top-down back-propagation pass is done to calculate the attention map which is informative about the model's decisions. Based on the neural attention map, the attended region is selected and fed into the foveal subnet. A GIF module is followed to effectively weight the extracted features from these two subnets. The overall architecture of the model is shown in Fig. 2.

The traditional methods often formulate the aesthetic assessment as binary classification as we have discussed earlier. The binary labels are typically derived from a distribution of scores (e.g. from $1-10$ in www.dpchallenge.com and from $1-7$ in www.photo.net). They compute and threshold the mean score of distributions. However, the single binary label removes the useful information of the ground-truth score distribution, such as the variance, the median, etc. These removed information is useful to investigate the consensus and diversity of opinions among annotators. Thus in this paper, we formulate the aesthetic assessment as a label distribution predicting problem. Each image in the dataset consists of its ground truth (user) ratings q. Let $q=[q_{s_1}, q_{s_2}, q_{s_N}]$ denote the score distributions of the images. s_i represents the i-th score bucket. N is the total number of score buckets. q_{s_i} denotes the number of voters that give a discrete score of s_i to the image. As for AVA dataset, $N=10$, $s_1=1$, $s_N=10$, but for Photo.net dataset, $N=7$, $s_1=1$, $s_N=7$ (The detailed introduction of AVA and Photo.net dataset can be found in section IV). The score distributions are l_1-normalized as a preprocessing step, and thus $\sum_{i=1}^{N} q_{s_i}=1$. The loss function

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2911428, IEEE Transactions on Multimedia

used in our paper is defined as follows:

$$EMD(q,\hat{q}) = (\frac{1}{N}\sum_{k=1}^{N}|CDF_q(k) - CDF_{\hat{q}}(k)|^r)^{\frac{1}{r}}, \quad (1)$$

where $CDF_q(k)$ is the cumulative distribution function, r is set as 2 to penalize the Euclidean distance between the CDFs. Our proposed GPF-CNN is applicable to a variety of CNN, such as AlexNet [33], VGGNet [34], ResNet [13] as demonstrated in the experiment part. For fair comparison with most of the aesthetic assessment methods, we select the VGG16 [34] as our baseline.

A. Top-down Neural Attention Feedback

The detail information locates in the original high resolution image. Training deep networks with large-size inputs requires a significantly larger dataset, and hardware memory. In this work, we use the top-down neural attention to discover the most important region of an image. The network then directs the high resolution "fovea" to extract fine-grained details. This offers a two-fold bonus. First, it helps to reduce the parameters. If we estimate the saliency map via a new saliency network, the number of learning parameters tends to be quite large. And it will increase the amount of computation and difficulty of training. Second, extracting local fine-grained features based on the neural network's attention is more valid, and can achieve better performance. Generally speaking, the neural attention map shows where the CNN is looking at when it makes the decisions. It is task-relevant. Traditional saliency map shows where the human gaze when seeing the images. And it is task-irrelevant. With the top-down neural attention map, we can focus on the aesthetic-relevant regions to extract fine-details.

Recently, lots of methods have been proposed to explore where the neural networks "look" in an image for evidence for their predictions [35], [36]. Our work is inspired by the excitation backprop method [36] which generates the top-down neural attention map based on the probabilistic Winner Take All (WTA) model. Given a selected output class, the probabilistic WTA scheme uses a stochastic sampling process to generate a soft attention map. The winning (sampling) probability $P(a_i)$ is defined as

$$P(a_i) = \sum_{a_j \in \varrho_i} P(a_i \mid a_j)P(a_j), \quad (2)$$

where $a_i \in N$ (N is the overall neuron set), ϱ_i is the parent node set of a_i (top-down order). As Eq. 2 indicates, $P(a_i)$ is a function of the winning probability of the parent nodes in the preceding layers [36]. Thus, the winner neurons are recursively sampled in a top down fashion based on a conditional winning probability $P(a_i \mid a_j)$. The conditional winning probability is defined as

$$P(a_i \mid a_j) = \begin{cases} Z_j \hat{a}_i w_{i,j} & if w_{i,j} \geq 0 \\ 0 & otherwise, \end{cases} \quad (3)$$

where Z_j is the normalization factor, $\hat{a}_i$ is the response of a_i, $w_{i,j}$ is the connection weight between a_i and a_j. Recursively propagating the top-down signal based on Eq. 2 and Eq. 3 layer by layer, we can compute the attention map of the predicted class. The computed attention map indicates which pixels are more relevant for the class. Next, we crop and zoom in the attended region to finer scale with higher resolution to extract fine-grained features.

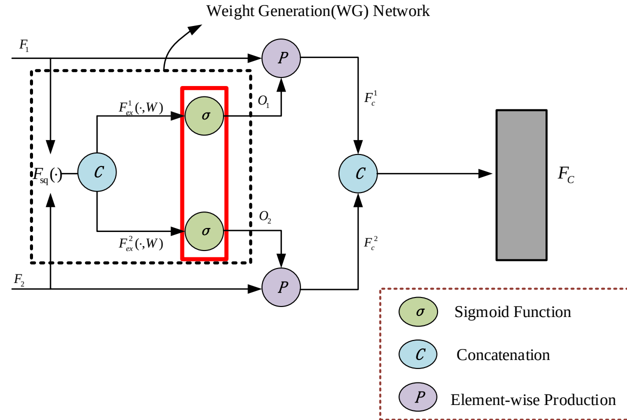

Fig. 3. The structure of the proposed GIF network. The GIF module produces weights O_1 and O_2 by applying the fully connected layers and the sigmoid function. Then, O_1 and O_2 are multiplied to the features to get the weighted information fusion results.

Attention based automatic image cropping tries to identify the most important region in the image. It aims to search for the smallest region inside which the summed attention is maximized. Suppose G is a non-negative valued top-down neural attention map. Larger attention values in G indicate higher visual importance. Without loosing generality, the attended regions can be found by optimising the following problem:

$$min \parallel R(\tau) \parallel s.t. \sum_{y \in R(\tau)} G(y) \geq \tau \sum_{y} G(y), \quad (4)$$

where τ is the minimum percentage of total attention to be preserved, $R(\tau)$ is the smallest rectangle that contain τ percentage of total attention, $\parallel R(\tau) \parallel$ is the rectangular area of $R(\tau)$. It should be emphasized that for a given τ[1], $R(\tau)$ may not be unique. In our algorithms, we always choose $R(\tau)$ with the largest summed attention value.

B. Gated Information Fusion (GIF) Network

The GIF module aims to balance the global and local feature according to the feature maps. The overall structure is shown in Fig. 3. Similar gated information fusion mechanism has been proposed for multi-modal learning [38]. In this paper, we generalise this design and focus on weighting the features by modeling the relationship between channels. The same idea has been adopted in SENet [39]. Let F_1 and F_2 denote the $M \times N \times K$ feature maps from the peripheral subnet and the foveal subnet. The GIF module consists of two parts: the weight generation part and the feature fusion part. In the weight generation part, a global pooling layer $F_{sq}(\cdot)$ is applied before concatenating the feature maps F_1

[1]We use the search strategy of [37], and follow the default parameter setting in the paper, *i.e.* $\tau = 0.7$

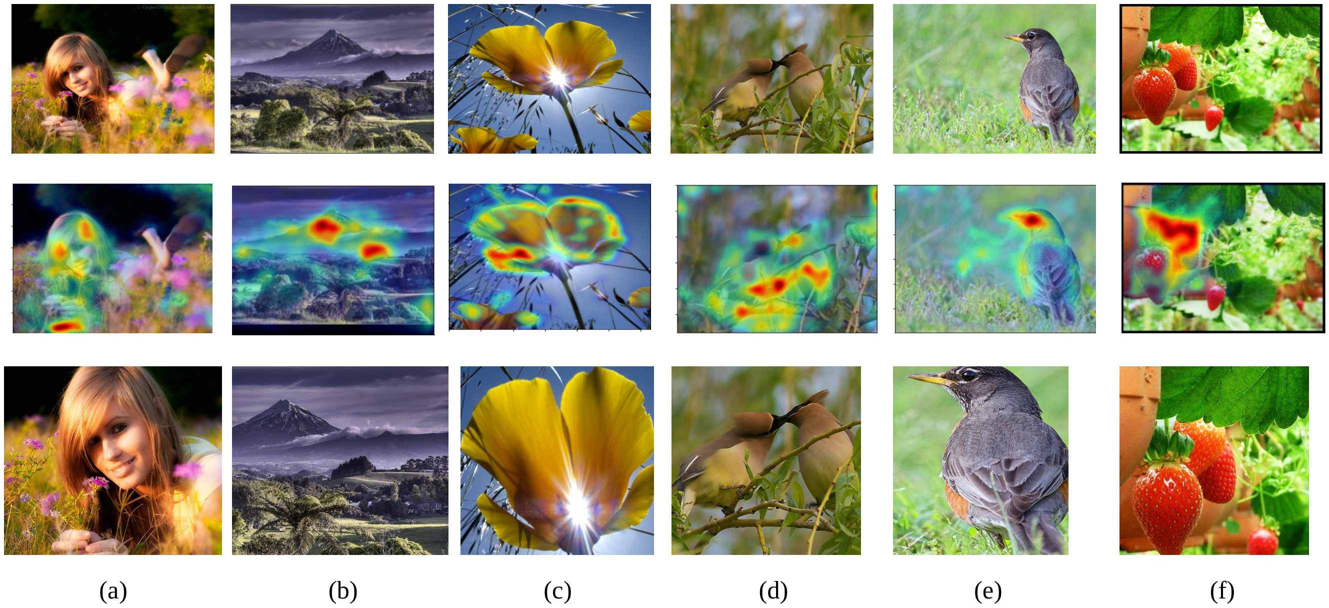

Fig. 4. Example patches cropped with neural attention. First row: original images; second row: top-down neural attention map; third row: patches cropped with neural attention map.

and F_2. $F_{sq}(\cdot)$ is used to squeeze global spatial features into channel descriptors [39]. Then, a bottleneck with two fully connected (FC) layers is applied in parallel to fully capture channel-wise dependencies. The sigmoid gating layer is employed to modulate the learned weights. Finally, the weighted feature maps are fed into the fully connected layers and the classification layer. Let U denote the concatenated features of F_1 and F_2. We summarize the operations of the GIF module as follows.

$$F_{sq}(u_c) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} u_c(i,j) \tag{5}$$

$$F_{ex}^i(\cdot, W) = W_2 \delta(W_1, z), i = 1, 2 \tag{6}$$

$$O_i = \sigma(F_{ex}^i(\cdot, W)), i = 1, 2 \tag{7}$$

$$F_c^i = \delta(W, x_i), i = 1, 2 \tag{8}$$

where δ denotes the *ReLU* function [40], W_1 is a dimensionality-reduction layer and W_2 is a dimensionality-increasing layer as defined in SENet [39], z refers the output features of $F_{sq}(u_c)$, and x_i denotes the input features of i-th branch.

IV. Experiments

In this section, we verify the effectiveness of the proposed photo aesthetic prediction approach on different datasets and CNN architectures. First, we perform the ablation studies on AVA dataset. The training networks include AlexNet [33], VGGNet [34], ResNet [13], and InceptionNet [41]. For all the architectures, our proposed scheme learns to perform better than the original networks. Next, we compare the performance of our scheme with state-of-the-art methods on AVA and Photo.net dataset.

A. Datasets and Implementation Details

AVA Dataset: The AVA aesthetic dataset [42] includes $250,000$ images, which is the largest public available aesthetics assessment dataset. The images are collected from www.dpchallenge.com. Each image has about 200 aesthetic ratings ranging from one to ten. We use the same partition of training data and testing data as the previous work [5], [18], [19], *i.e.* $230,000$ images for training and validation, the rest $19,000$ images for testing.

Photo.net Dataset: The Photo.net dataset [25] is collected from www.photo.net. It consists of $20,278$ images but only $17,200$ images have aesthetic label distribution. Distribution (counts) of aesthetics ratings are from one to seven. From the overall images, $15,000$ images are used to train, 1000 images are used for validation and the rest 1200 images are used for test.

Considering that the peripheral subnet is used for encoding the global composition features, we do not rescale the input into fixed size but use downsampling to keep its original aspect ratios. The longest dimension of the input image is kept to 224. The training process includes two steps. In the first step, we initialize convolutional layers in the peripheral subnet by the pre-trained VGG16 from ImageNet [34]. We first train the peripheral subnet with softmax loss to classify the images into two categories, i.e. high quality and low quality. After training the peripheral subnet, we can get the attended regions by feeding back the top-down neural attention. In the second step, we freeze the convolutional layers of peripheral subnet, and start to train the foveal subnet and the GIF module. Each input

TABLE I
EFFECTIVENESS OF NEURAL NETWORK ATTENTION MODULE.

Network architecture	Accuracy (%)↑	SRCC(mean)↑	LCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
VGG16	74.41	0.6007	0.5869	0.4611	0.5878	0.0539
Random-VGG16	78.54	0.6274	0.6382	0.4410	0.5660	0.0510
Saliency-VGG16	79.19	0.6601	0.6711	0.4228	0.5430	0.0475
PF-CNN	80.60	0.6604	0.6712	0.4176	0.5387	0.047
GPF-CNN	**80.70**	**0.6762**	**0.6868**	**0.4144**	**0.5347**	**0.046**

TABLE II
PARAMETER SETTINGS OF GIF IN DIFFERENT ARCHITECTURES

		AlexNet	VGG16	InceptionNet	ResNet-18
Operation	Layer type	Parameter	Parameter	Parameter	Parameter
$F_{ex}^1(\cdot, W)$	fully connected	[128,256]	[256,512]	[512,2048]	[128,512]
$F_{ex}^2(\cdot, W)$	fully connected	[128,256]	[256,512]	[512,2048]	[128,512]
F_c^1	fully connected	2048	2048	2048	2048
F_c^2	fully connected	2048	2048	2048	2048
F_C	fully connected	4096	4096	4096	4096

image is normalised through mean RGB-channel subtraction. Both the two steps adopt the SGD optimization algorithm. The minibatch samples 64 images randomly in each iteration. The momentum is 0.9. The initial learning rate is set to 0.001 and reduced by a factor of 10 every 10 epochs. The training continues until the validation loss reaches a plateau for 10 epochs. The hyper-parameters for the first and second step training are unified. All experiments are implemented based on the open source PyTorch framework with a NVIDIA Pascal TITAN X GPU.

Unlike most traditional methods that are designed to perform the binary classification, we evaluate our proposed method with respect to three aesthetic quality tasks: (i) aesthetic score regression, (ii) aesthetic quality classification, and (iii) aesthetic score distribution prediction. For the aesthetic score regression task, the mean score of the label distribution is computed via $\mu = \sum_{i=1}^{N} s_i \times p_{s_i}$. For the aesthetic quality classification, we threshold the mean score using the threshold 5 just as the work of [5], [6], [19], [27]. Images with predicted scores above 5 are categorized as high quality and vice versa. The evaluation metrics related to the three prediction tasks are as follows.

- *Image aesthetic score regression*: We report the Spearman rank-order correlation coefficient (SRCC), Pearson linear correlation coefficient (LCC), root mean square error (RMSE) and mean absolute error (MAE). These are the most significant for testing the performance of an IQA method. Of these criteria, SRCC measures the prediction monotonicity, and the LCC provides an evaluation of prediction accuracy. Both SRCC and LCC range from 0 to 1, and larger value indicates better result. While for MAE and RMSE, the smaller value indicates the better results.
- *Image aesthetic quality classification*: We report the overall accuracy, defined as $Accuracy = \frac{TP+TN}{P+N}$.
- *Image aesthetic score distribution*: We use EMD to measure the closeness of the predicted and ground truth rating distribution. The EMD is defined in Eq.1 with $r = 1$.

B. Ablation Studies

Traditional methods extract the local features based on random cropping [18]. The random cropping method is independent of the image content. It is unlikely to capture the semantic meaning. Another alternative is to extract the fine-grained details based on salient object detection or saliency detection. The salient object detection can perform well on condition that there is one salient object. When there are multiple objects, it is difficult to choose which is the most important one. Besides, for most landscape images, there is no salient object in the image. The human saliency detection shows where human gaze when seeing the images. It adds extra computation burden to the network as we described earlier. However, extracting fine-grained features based on neural attention can tackle the above challenges. Fig. 4 shows some examples of patches cropped with neural attention. Fig. 4(a)(c)(e) have only one subject in the image. The cropped patches can capture the important region and preserve the semantic integrity. Fig. 4(d)(f) have multiple objects in the image. But the cropped patches can capture both of them.

To validate the effectiveness neural attention module quantitatively, we conduct three baselines: VGG16, Random-VGG16, and Saliency-VGG16 methods. The VGG16 is pre-trained on ImageNet and fine-tuned to predict the aesthetic quality. The input of VGG16 is obtained by wrapping the original input image to the fixed size of 224×224. The Random-VGG16 is a double-column deep convolutional neural networks. The first column encodes the global views and the size of input image is 224×224. The second column uses random

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2911428, IEEE Transactions on Multimedia

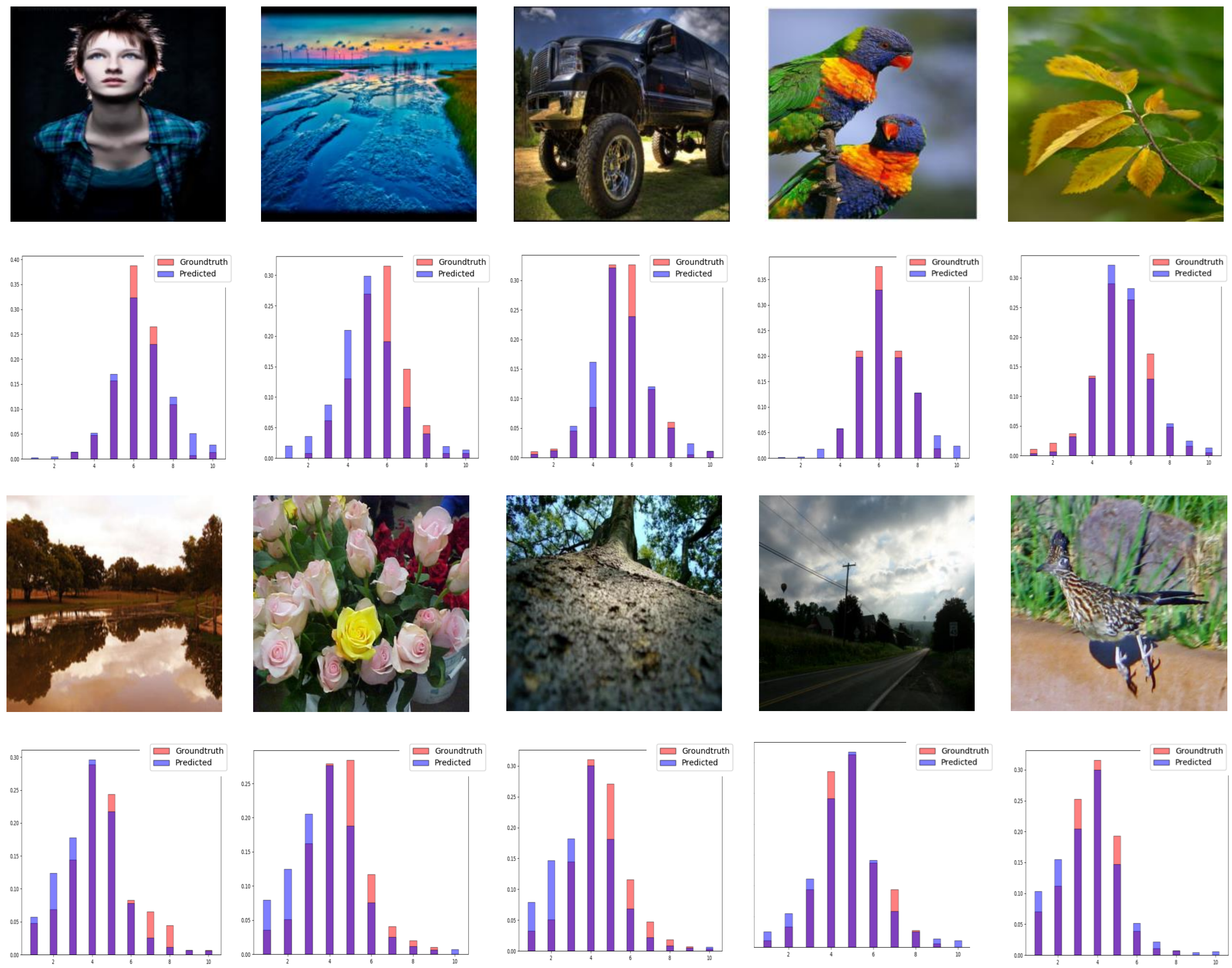

Fig. 5. Top 2 rows: high quality images, as predicted by our GPF-CNN (VGG16), coupled with plots of their ground-truth and predicted score distributions. Bottom 2 rows: the low quality images, as predicted by our GPF-CNN (VGG16), coupled with plots of their ground-truth and predicted score distributions.

cropping method to extract the local fine-grained information. The cropped patch size is fixed to be 224×224. The Saliency-VGG16 has a very similar architecture with Random-VGG16, i.e. including two column CNNs. But the Saliency-VGG16 extracts local fine-details with the guidance of saliency map predicted by DVA [43] model (DVA model achieves state-of-the-art performance on human saliency prediction task.). The PF-CNN is a simplification of proposed GPF-CNN by removing the GIF module. It uses the neural attention to extract the fine-grained details. The attended regions of foveal subnet are resized to 224×224 in training and testing. For fair comparison, we use the same network architecture and unify the hyper-parameters. The results are shown in Table I. It can be seen that the Random-VGG16, Saliency-VGG16, and PF-CNN achieve better performance compared with VGG16, which indicates that incorporating local fine-grained features can improve the prediction results. This is consistent with the results of [18], [19], who uses the random cropping strategy to capture fine-grained details. The Saliency-VGG16 and PF-CNN perform much better than Random-VGG16. This illustrates the importance of using attention mechanism to encode the fine-grained details. Furthermore, PF-CNN outperforms the Saliency-VGG16, which demonstrates the superiority of neural attention map over predicted saliency map.

In order to see whether the GIF module is effective, we compare the GPF-CNN with PF-CNN. Compared with PF-CNN, GPF-CNN has GIF module to weight the global and local features. The baseline network is still VGG16. The detailed parameters of GIF module in VGG16 are illustrated in Table II. The comparison results are shown in Table I. GPF-CNN performs better than PF-CNN on all three tasks, i.e. aesthetic classification, aesthetic regression and aesthetic label distribution prediction.

In conclusion, our experimental results confirm the importance of fusing global and local fine-grained details, emphasizing the critical importance of neural attention and GIF module in our framework.

TABLE III
EXTENSIONS TO OTHER NETWORK ARCHITECTURE.

Network architecture	Accuracy (%)↑	SRCC(mean)↑	LCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
AlexNet	76.37	0.5549	0.5665	0.4733	0.6063	0.0525
ResNet-16	77.91	0.6394	0.6505	0.4346	0.5583	0.0484
InceptionNet	79.43	0.6756	0.6865	0.4154	0.5359	0.0466
GPF-CNN(AlexNet)	**78**	**0.5996**	**0.6121**	**0.4539**	**0.5820**	**0.0507**
GPF-CNN(ResNet-18)	**80.30**	**0.6711**	**0.6817**	**0.4185**	**0.5390**	**0.0474**
GPF-CNN(InceptionNet)	**81.81**	**0.6900**	**0.7042**	**0.4072**	**0.5246**	**0.045**

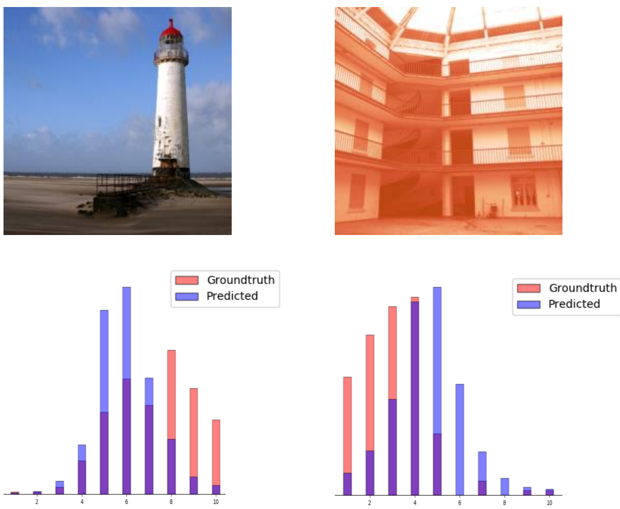

Fig. 6. Failure cases of our model. Our model performs poorly on bimodal distribution or on very skewed distributions.

C. Extension to Other Network Architectures

We next investigate the performance of GPF-CNN mechanism on several other architectures: AlexNet [33], ResNet-18 [13], and InceptionNet [41]. The parameters of GIF module that are integrated with AlexNet, ResNet-18 and InceptionNet are shown in Table II. The $F_{ex}^{i}(\cdot, W), i = 1, 2$ is a bottleneck with two fully connected (FC) layers: a dimensionality-reduction layer with parameters W_1, a *ReLU*, and then a dimensionality increasing layer with parameters W_2. The W_1 and W_2 are set 128 and 256 respectively for AlexNet, 256 and 512 for VGG16, 512 and 2048 for InceptionNet (We have tried other parameters. But we have not seen any improvements.). The comparison results are illustrated in Table III. As with the previous experiments, we observe significant performance improvements induced by the GPF-CNN mechanism.

D. Content-based Photo Aesthetic Analysis

In this section, we demonstrate the effectiveness of the proposed method on various types of images. We select eight category images from the test set of AVA dataset: *i.e.* animal, landscape, cityscape, floral, food-drink, architecture, portrait, still-life. The image collection is the same with previous works of [5], [19], [42], about $2.5K$ in each of the categories. In each category of images, we systematically compare the proposed GPF-CNN with VGG16, Random-VGG16, and PF-CNN. The experimental results are illustrated in Table IV. For all the seven categories, random-VGG16, PF-CNN, and GPF-CNN perform better than VGG16. These results indicate that fine-details information is quite important for image aesthetic prediction. We can also find that the performance of the proposed GPF-CNN significantly outperforms the baselines in most of the categories. The portrait shows substantial improvements, reaching a 6.79% improvement compared with VGG16. This is because the fine details in the face, such as light, contrast is quite important in portrait aesthetic assessment. The proposed GPF-CNN is sensitive to the faces since it uses the neural attention to extract the fine-grained details (see Fig. 4(a)).

E. Comparison with the State-of-the-Art on AVA Dataset

In this section, we quantitatively compare our approach with eight photo aesthetics evaluation methods. The compared methods include one handcrafted feature based approach, i.e. *Graphlet-based method* [26], as well as six deep learning based classification methods i.e. *MTRLCNN* [27], *A-Lamp* [19], *MNA-CNN* [5], *RAPID* [17], *DMA-Net* [18], *MS-DLM* [44], and one label distribution method, i.e. *NIMA* [32]. Table V shows the comparison results. The handcrafted feature based method [26] uses different data partition to test the model performance. Other methods follow a standard data partition. To date, the AVA dataset that follows the standard partition is considered to be the most challenging. That is why [26] achieves 83.24% but our method achieves 81.81%. Note that our GPF-CNN achieves the best performance among methods that follow standard data partition. Methods of *RAPID* [17], *DMA-Net* [18] and *MSDLM* [44] are based on shallow networks, achieving 74.2%, 75.42% and 76.94% respectively. But the proposed GPF-CNN (AlexNet) achieves 78%. This is a 3.8%, 2.58% and 1.06% performance improvement, respectively. For the larger VGG16 network, our method GPF-CNN (VGG16) performs slightly worse than *A-Lamp* [19] but outperforms *MTRLCNN* [27] and *MNA-CNN* [5] by 4.5% and 2.14% respectively. Note that *A-Lamp* [19] only performs binary classification. Our method provides richer and more precise information than binary classification. *NIMA* [32] is closely related to our work since they use the *EMD* loss to optimise their network. The *SRCC* and *LCC* of NIMA is 0.592 and 0.610 respectively on VGG16, while *GPF-CNN* achieves 0.6762 and 0.6868. This is a 8.42%

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2911428, IEEE Transactions on Multimedia

TABLE IV
ABLATION STUDY ON EIGHT CATEGORY IMAGES.

Category	Network architecture	Accuracy (%)↑	SRCC(mean)↑	LCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
animal	**GPF-CNN**	**80.80**	**0.7480**	**0.7478**	**0.3941**	**0.5051**	**0.045**
	PF-CNN	80.23	0.7274	0.7277	0.4091	0.5238	0.0462
	VGG-Random16	77.9	0.6587	0.6654	0.4475	0.5688	0.0516
	VGG16	76.17	0.6212	0.6267	0.4959	0.6325	0.0546
landscape	**GPF-CNN**	**85.42**	**0.7746**	**0.7822**	**0.3713**	**0.4705**	**0.0422**
	PF-CNN	84.78	0.7606	0.7685	0.3831	0.4863	0.0431
	VGG-Random16	82.97	0.7230	0.7318	0.4051	0.5153	0.0480
	VGG16	80.04	0.6780	0.6876	0.4968	0.6276	0.0541
cityscape	**GPF-CNN**	**81.68**	**0.7533**	**0.7539**	**0.3956**	**0.5103**	**0.0443**
	PF-CNN	80.59	0.7365	0.7362	0.4096	0.5284	0.0456
	VGG-Random16	77.02	0.6808	0.6827	0.4438	0.5676	0.0508
	VGG16	76.22	0.6460	0.6424	0.5074	0.6481	0.0552
floral	**GPF-CNN**	**79.95**	**0.7374**	**0.7348**	**0.3681**	**0.4785**	**0.0423**
	PF-CNN	79.15	0.7196	0.7171	0.3794	0.4921	0.0433
	VGG-Random16	77.19	0.6564	0.6576	0.4147	0.5312	0.0487
	VGG16	75.58	0.6184	0.6220	0.4455	0.5709	0.05
fooddrink	**GPF-CNN**	**80.22**	**0.7389**	**0.7476**	**0.3919**	**0.4948**	**0.0443**
	PF-CNN	79.46	0.7180	0.7288	0.4081	0.5125	0.0456
	VGG-Random16	77.58	0.6642	0.6740	0.4357	0.5498	0.0503
	VGG16	74.01	0.6163	0.6278	0.4876	0.6208	0.0536
architecture	**GPF-CNN**	**81.60**	**0.7431**	**0.7410**	**0.3704**	**0.4822**	**0.0421**
	PF-CNN	81.16	0.7221	0.7213	0.3840	0.4970	0.0433
	VGG-Random16	79.16	0.6708	0.6709	0.4191	0.5379	0.0476
	VGG16	76.83	0.6032	0.6069	0.4748	0.6091	0.0521
portrait	**GPF-CNN**	**83.52**	**0.6987**	**0.7047**	**0.4228**	**0.5389**	**0.0475**
	PF-CNN	82.72	0.6774	0.6866	0.4386	0.5549	0.0487
	VGG-Random16	81.71	0.6215	0.6331	0.4672	0.5884	0.0522
	VGG16	76.93	0.5671	0.5726	0.5347	0.6784	0.0583
still life	**GPF-CNN**	**76.35**	**0.7001**	**0.7127**	**0.4039**	**0.5153**	**0.0455**
	PF-CNN	75.23	0.6772	0.6909	0.4210	0.5338	0.0468
	VGG-Random16	71.49	0.6158	0.6329	0.4488	0.5683	0.0522
	VGG16	71.14	0.5652	0.5810	0.49	0.6187	0.0537

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON AVA DATASET.

Network architecture	Accuracy (%)↑	SRCC(mean)↑	LCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
Graphlet-based method [26]	83.24	-	-	-	-	-
RAPID(AlexNet) [17]	74.2	-	-	-	-	-
DMA-Net(AlexNet) [18]	75.42	-	-	-	-	-
MSDLM(AlexNet) [44]	76.94	-	-	-	-	-
MNA-CNN(VGG16) [5]	76.1	-	-	-	-	-
A-Lamp(VGG16) [19]	82.5	-	-	-	-	-
MTRLCNN(VGG16) [27]	78.46	-	-	-	-	-
NIMA(VGG16) [32]	80.6	0.592	0.610	-	-	0.052
NIMA(InceptionNet) [32]	81.51	0.612	0.636	-	-	0.05
GPF-CNN(AlexNet)	**78**	**0.5996**	**0.6121**	**0.4539**	**0.5820**	**0.0507**
GPF-CNN(VGG16)	**80.70**	**0.6762**	**0.6868**	**0.4144**	**0.5347**	**0.046**
GPF-CNN(InceptionNet)	**81.81**	**0.6900**	**0.7042**	**0.4072**	**0.5246**	**0.045**

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS ON PHOTO.NET DATASET.

Network architecture	Accuracy (%)↑	SRCC(mean)↑	LCC (mean)↑	MAE↓	RMSE ↓	EMD ↓
GIST_SVM [28]	59.90	-	-	-	-	-
FV_SIFT_SVM [28]	60.8	-	-	-	-	-
MTCNN(VGG16) [27]	65.2	-	-	-	-	-
GPF-CNN(AlexNet)	**69.03**	**0.4106**	**0.4258**	**0.4860**	**0.5876**	**0.0794**
GPF-CNN(VGG16)	**75.6**	**0.5217**	**0.5464**	**0.4242**	**0.5211**	**0.070**
GPF-CNN(InceptionNet)	**77.19**	**0.5018**	**0.5250**	**0.3996**	**0.5019**	**0.0706**

TABLE VII
OTHER VARIANTS OF NETWORK ARCHITECTURE.

Network architecture	Reduction ratio	Accuracy (%)↑	SRCC(mean)↑	LCC(mean)↑	MAE↓	RMSE ↓	EMD ↓
GPF-CNN(VGG16)	(4,4)	**80.70**	**0.6762**	**0.6868**	**0.4144**	**0.5347**	**0.046**
GPF-CNN(VGG16)	(4,8)	79.35	0.6687	0.6794	0.4187	0.5399	0.0468
PF-SENet(VGG16)	4	79.52	0.6680	0.6787	0.4203	0.5408	0.0467

and 7.6% improvement. This is, to the best of our knowledge, the state-of-the-art performance on AVA dataset.

Fig. 5 shows the top six and bottom six images randomly selected in the AVA test set. Plots of the ground-truth and predicted distributions are also shown. We can find that the model can achieve a high degree of accuracy, with almost perfect reconstruction in some cases. Fig. 6 shows some failure cases of our model. Our trained model performs poorly on images which have very non-Gaussian distributions. But the Gaussian functions perform adequately for 99.77% of all the images in the AVA dataset, as reported by Murray [42].

F. Evaluating Performance on Photo.net Dataset

The comparison results on Photo.net are shown in Table VI. GPF-CNN(VGG16) outperforms the baselines by a large margin, achieving 75.6% accuracy rate. This is around 10.4% better than *MTCNN* [27], and 14.8% better than traditional methods. We also observe that GPF-CNN(VGG16) achieves 4.57% improvement in classification accuracy rate, 1.02% improvement in SRCC, and 1.02% in LCC, compared with GPF-CNN(AlexNet). But GPF-CNN(InceptionNet) only achieves 1.59% improvement in accuracy rate compared with GPF-CNN(VGG16). The SRCC and LCC are slightly lower than GPF-CNN(VGG16). This is mainly because the number of training images on Photo.net is $1,5000$, which is too small to support training GPF-CNN(InceptionNet).

G. Other Variants

We have also investigated other variants of GPF-CNN on AVA dataset. For example, we replace the GIF module with a Squeeze-and-Excitation block (SE block) to fuse the global and local features, and denote the new baseline as PF-SENet. For fair comparison, the reduction ratios in PF-SENet and GPF-CNN are unified. Tabel VII illustrates the performance in terms of six metrics. From the results, the GPF-CNN outperforms PF-SENet by 1.38% in aesthetic classification task, which indicates the superiority of GIF module over single SE block. Another variant is to use the GIF module but the reduction ration in two branches are different. We conduct experiments based on GPF-CNN(VGG16). The inner brackets $(4,4)$ in Table VII indicates that the reduction ratios of the two branches in the GIF module are both 4. While inner brackets $(4,8)$ indicates that the reduction ratios in the two branches are 4 and 8 respectively. From the table, we can observe that using same reduction ration in two branches achieves better performance. Consequently, we unify the reduction ratios in two branches of GIF module for all experiments.

V. CONCLUSION

This paper presents a biological model for photo aesthetic assessment. In human vision system, the fovea has the highest possible visual acuity and is responsible for seeing the fine details. The peripheral vision has a significantly lower density of cones and is used for perceiving the broad spatial scene. Besides, foveal and peripheral vision play different roles in processing different visual stimuli. We are inspired by these observations and propose the GPF-CNN architecture. It can learn to focus on the important regions of top-down neural attention map to extract the fine details features. The GIF module can adaptively fuse the global and local features according to the input feature map. The experimental results on the large-scale AVA and Photo.net datasets show that our GPF-CNN can significantly improve the state-of-the-art for three tasks: aesthetic quality classification, aesthetic score regression and aesthetic score distribution prediction. In the future work, we will further explore the human vision system and design more powerful model for aesthetic prediction tasks.

REFERENCES

[1] F.-L. Zhang, M. Wang, and S.-M. Hu, “Aesthetic image enhancement by dependence-aware object recomposition,” *IEEE Trans. Multimedia*, vol. 15, no. 7, pp. 1480–1490, 2013.

[2] A. Samii, R. Měch, and Z. Lin, "Data-driven automatic cropping using semantic composition search," *Computer Graphics Forum*, vol. 34, no. 1, pp. 141–151, 2015.

[3] S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 271–280.

[4] H. Talebi and P. Milanfar, "Learned perceptual image enhancement." [Online]. Available: https://arxiv.org/abs/1712.02864

[5] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 27-30, 2016, pp. 497–506.

[6] L. Guo, Y. Xiong, Q. Huang, and X. Li, "Image esthetic assessment using both hand-crafting and semantic features," *Neurocomputing*, vol. 143, pp. 14–26, 2014.

[7] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[8] C. Chamaret and F. Urban, "No-reference harmony-guided quality assessment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 23-28, 2013, pp. 961–967.

[9] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato, "Aesthetic quality classification of photographs based on color harmony," in *Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition*, June 20-25, 2011, pp. 33–40.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015.

[11] Z. Jiao, X. Gao, Y. Wang, J. Li, and H. Xu, "Deep convolutional neural networks for mental load classification based on EEG data," *Pattern Recognition*, vol. 76, pp. 582–595, 2018.

[12] X. Zhang, X. Gao, W. Lu, L. He, and Q. Liu, "Dominant vanishing point detection in the wild with application in composition analysis," *Neurocomputing*, vol. 311, pp. 260–269, 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[14] X. Tian, Z. Dong, K. Yang, and T. Mei, "Query-dependent aesthetic model with deep learning for photo quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2035–2048, 2015.

[15] S. Kong, X. Shen, Z. L. Lin, R. Mech, and C. C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proceedings of 14th European Conference on Computer Vision*, 2016, pp. 662–679.

[16] B. Jin, M. V. O. Segovia, and S. Süsstrunk, "Image aesthetic predictors based on weighted cnns," in *Proceedings of International Conference on Image Processing*, September 25-28, 2016, pp. 2291–2295.

[17] X. Lu, Z. L. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, 2015.

[18] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proceedings of IEEE International Conference on Computer Vision*, December 7-13, 2015, pp. 990–998.

[19] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, July 21-26, 2017, pp. 722–731.

[20] H. Strasburger, I. Rentschler, and M. Jüttner, "Peripheral vision and pattern recognition: A review," *Journal of Vision*, vol. 11, no. 5, pp. 13–13, 2011.

[21] P. Wang and G. W. Cottrell, "Central and peripheral vision for scene recognition: a neurocomputational modeling exploration," *Journal of Vision*, vol. 17, no. 4, pp. 9–9, 2017.

[22] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, M. Messner, G. R. Bradski, P. Baumstarck, S. Chung, and A. Y. Ng, "Peripheral-foveal vision for real-time object recognition and tracking in video," in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, January 6-12, 2007, pp. 2115–2121.

[23] C. J. Ludwig, J. R. Davies, and M. P. Eckstein, "Foveal analysis and peripheral selection during active visual sampling," *Proceedings of the National Academy of Sciences*, vol. 111, no. 2, pp. E291–E299, 2014.

[24] A. M. Larson and L. C. Loschky, "The contributions of central versus peripheral vision to scene gist recognition," *Journal of Vision*, vol. 9, no. 10, pp. 6–6, 2009.

[25] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proceedings of European Conference on Computer Vision*, May 7-13, 2006, pp. 288–301.

[26] L. Zhang, Y. Gao, R. Zimmermann, Q. Tian, and X. Li, "Fusion of multichannel local and global structural cues for photo aesthetics evaluation," *IEEE Trans. Image Processing*, vol. 23, no. 3, pp. 1419–1429, 2014.

[27] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Trans. Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[28] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *Proceedings of IEEE International Conference on Computer Vision*, 2011, pp. 1784–1791.

[29] M. Kucer, A. C. Loui, and D. W. Messinger, "Leveraging expert feature knowledge for predicting image aesthetics," *IEEE Trans. Image Processing*, vol. 27, no. 10, pp. 5100–5112, 2018.

[30] X. Jin, L. Wu, X. Li, S. Chen, S. Peng, J. Chi, S. Ge, C. Song, and G. Zhao, "Predicting aesthetic score distribution through cumulative jensen-shannon divergence," in *Proceedings of the AAAI Conference on Artificial Intelligence*, February 2-7, 2018, pp. 77–84.

[31] N. Murray and A. Gordo, "A deep architecture for unified aesthetic prediction." [Online]. Available: https://arxiv.org/abs/1708.04890

[32] H. Talebi and P. Milanfar, "NIMA: neural image assessment," *IEEE Trans. Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, December 3-6, 2012, pp. 1106–1114.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[35] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 27-30, 2016, pp. 2921–2929.

[36] J. Zhang, Z. L. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *Proceedings of the 14th European Conference on Computer Vision*, October 11-14, 2016, pp. 543–559.

[37] J. Chen, G. Bai, S. Liang, and Z. Li, "Automatic image cropping: A computational complexity study," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 27-30, 2016, pp. 507–515.

[38] J. Kim, J. Koh, Y. Kim, J. Choi, Y. Hwang, and J. W. Choi, "Robust deep multi-modal learning based on gated information fusion network." [Online]. Available: https://arxiv.org/abs/1807.06233

[39] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 18-22, 2018, pp. 7132–7141.

[40] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, June 21-24, 2010, pp. 807–814.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[42] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 16-21, 2012, pp. 2408–2415.

[43] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.

[44] W. Wang, M. Zhao, L. Wang, J. Huang, C. Cai, and X. Xu, "A multi-scene deep learning model for image aesthetic evaluation," *Sig. Proc.: Image Comm.*, vol. 47, pp. 511–518, 2016.

Xiaodan Zhang received the B.Eng. degree in electronic information engineering from Xidian University, Xian, China, in 2014, where she is currently pursuing the Ph.D. degree in intelligence information processing. Her current research interests include visual quality assessment, visual attention, and computationally modeling of human visual system.

Xinbo Gao (M'02-SM'07) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a research fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a post-doctoral research fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education of P. R. China, a Professor of Pattern Recognition and Intelligent System, and the Dean of Graduate School of Xidian University. His current research interests include Image processing, computer vision, multimedia analysis, machine learning and pattern recognition. He has published six books and around 300 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology and a Fellow of the Chinese Institute of Electronics.

Wen Lu received the BSc, MSc and PhD degrees in signal and information processing from Xidian University, China, in 2002, 2006and 2009 respectively. He is currently the professor at Xidian University and postdoctoral research in the department of electrical engineering at Stanford University, USA. His research interests include image quality metric, human vision system, computational vision. He has published 2 books and around 30 technical articles in refereed journals and proceedings including IEEE TIP, TSMC, Neurocomputing, Signal processing etc.

Lihuo He is currently an associate professor at Xidian University. He received the B.Sc.degree in electronic and information engineering and Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xian, China, in 2008 and 2013. His research interests focus on image/video quality assessment, cognitive computing, and computational vision.