Integrity-Preserving Image Aesthetic Assessment

Xin Sun(✉) and Jun Zhou

Institute of Image Communication and Network Engineering,
Shanghai JiaoTong University, Shanghai 200240, China
379349408@qq.com
zhoujun@sjtu.edu.cn

Abstract. Image aesthetic assessment is a challenging problem in the field of computer vision. Recently, the input size of images is often limited by the network of aesthetic problems. The methods of cropping, wrapping and padding unify images to the same size, which will destroy the aesthetic quality of the images and affect their aesthetic rating labels. In this paper, we present an end-to-end deep Multi-Task Spatial Pyramid Pooling Fully Convolutional Neural NasNet (MTP-NasNet) method for image aesthetic assessment that can directly manipulate the original size of the image without destroying its beauty. Our method is developed based on Fully Convolutional Network (FCN) and Spatial Pyramid Pooling (SPP). In addition, existing studies regards aesthetic assessment as a two-category task, a distribution predicting task or a style predicting task, but ignore the correlation between these tasks. To address this issue, we adopt the multi-task learning method that fuses two-category task, style task and score distribution task. Moreover, this paper also explores the reference of information such as variance in the score distribution for image reliability. Our experiment results show that our approach has significant performance on the large-scale aesthetic assessment datasets (AVA [1]), and demonstrate the importance of multi-task learning and size preserving. Our study provides a powerful tool for image aesthetic assessment, which can be applied to photography and image optimization field.

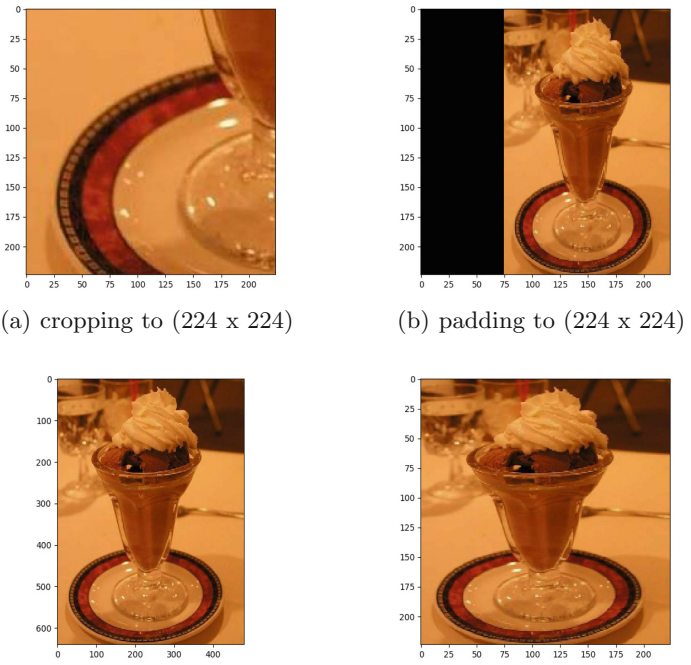
Keywords: Multi-task learning · Image aesthetic assessment · Fully convolutional neural networks · Spatial pooling layer

1 Introduction

Image aesthetic assessment is a challenging issue in the field of computer vision in recent years, and has a wide range of application scenarios. For example, image aesthetic quality assessment can give certain guidance and help to photography [2]; it can be used as a loss function for image beautification or optimization [3]; iterative artificial intelligence can make pictures and optimization [4].

H. Gao et al. (Eds.): ChinaCom 2019, LNICST 313, pp. 3–19, 2020.
https://doi.org/10.1007/978-3-030-41117-6_1

Recently, the most models for aesthetic assessment have to fix the size of the input image, thus destroying the aesthetic elements of the image, affecting its aesthetic score, and affecting the subsequent training. Figure 1 shows the three common methods that divert images to the fixed size. The original image (640×480) is taken by three operations: cropping, wrapping and padding that transforms the image to the specified size (224×224). These operations will obviously damage the beauty of the picture. To address this problem, this paper proposes the MTP-NasNet for images of arbitrary size input, which has achieved outstanding experimental results. The main work are introducing the modified FCN of image segmentation and the SPP layer of image classification to our aesthetic models. In addition, for the convenience and efficiency of training, three different treatments were performed to speed up the training and increase the convergence speed for three different aspect ratio pictures.



(a) cropping to (224 x 224) (b) padding to (224 x 224)

(c) the original image (640 x 480) (d) wrapping to (224 x 224)

Fig. 1. The original image and three images that transformed to (224×224)

In addition, early image aesthetic assessments were judged based on two classifications, that is, the image was divided into high and low quality according to the threshold of the rating score (usually 5 points). Usually, a single and simple label (i.e.,good or bad) is attached to image to indicate its aesthetic quality [5]. However, due to the subjectivity of aesthetic assessment, a simple label might be

insufficient to indicate the divergence among different rater' aesthetic. In recent years, the researches on aesthetic assessment have gradually changed from simple two-category prediction to complex score distribution prediction and image style prediction [6–9]. Nevertheless, the correlation between these tasks is ignored. In fact, these tasks can promote each other and learn together.

In this paper, we introduce a method based on multi-task learning that learned score distributed prediction, two-category prediction and style prediction at the same time. The two classification of images reflects the overall aesthetic quality of a photo, but for aesthetic problems, the amount of information in training times is too small, it is difficult for computer learning to get better results, and it cannot reflect the difference of human views. At present, the relationship between the two-category prediction task and the score distribution prediction task is not effectively utilized. In addition, style prediction is also helpful for the two-category prediction tasks and score distribution prediction tasks. Above all, multi-task learning is an inductive migration mechanism that uses additional sources of information to improve the learning performance of current tasks [10–12]. Based on the above points, this paper proposes a new multi-task method combining distributed prediction, two-category prediction and style prediction.

Furthermore, we discovery that the distribution of each image rating score of our AVA dataset has different variances. Figure 2 is the variance histogram and standard deviation histogram of all the image scores distributions of the AVA dataset. The abscissa is the magnitude of the variance and the standard deviation, and the ordinate is the number of images. The large variance distribution indicates that the scorers' opinions have large divergence and therefore its reference is quite doubtful. So the significance of such training data onto predicting the high and low quality binary classification is relative smaller. Based on this problem, this paper provides a corresponding weight for each training data to indicate its reliability.

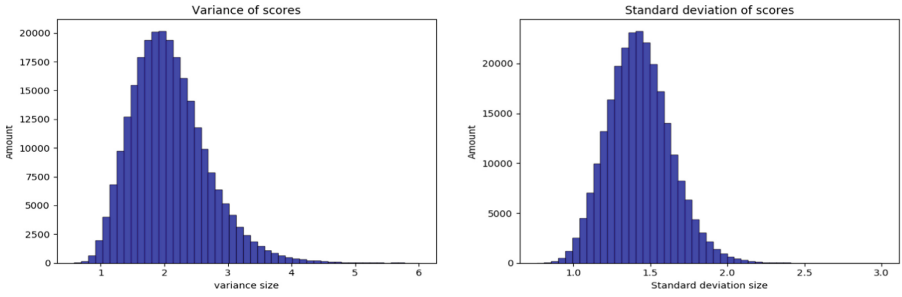


Fig. 2. Variance and Standard deviation histogram of the score distribution.

Our main contributions can be summarized as the following three points:

In order to solve the problem that the picture input into the network is unified to a fixed size and thus the image aesthetics is damaged, we introduce the full

convolutional network and the spatial pyramid pooling layer, and propose a new aesthetic evaluation network that can process image input of any size.

In order to improve the experimental results, we adopted the multi-task learning method, and jointly studied the three tasks of two-category prediction, style prediction and score distribution prediction.

Finally, considering the variance information in the score distribution to help determine the reliability of the score, we added the weight of the distribution variance to each image to optimize the learning efficiency.

The rest of this paper is organized as follows: The second part reviews the related work in this field in recent years, and the third part elaborates the details and implementation details of the new model proposed in this paper. The fourth part shows the experimental results and experimental analysis. The final fifth section summarizes the overall work and looks ahead to future work.

2 Related Work

In this section, we first review the existing research on image aesthetic assessment and then demonstrate a short overview of two interrelated of our work: multi-task learning and our MTP-NasNet.

2.1 Image Aesthetic Assessment

Detecting and measuring different distortions, e.g. block, noise and blur to measure image quality might be the earliest methods for the image quality assessment [13]. Although these methods always have an outstanding effect on those issues caused by storage, transmission and acquisition, they often reflect people's subjective perception of image aesthetic quality not well [14].

Image aesthetic quality evaluation has attracted the attention of many researchers because of its wide application [4]. The common image aesthetic quality evaluation system is divided into two stages: feature extraction and decision making. Feature extraction is divided into manual design features and deep learning automatic learning features [3]. Classifiers or regressions such as Bayesian classifiers, support vector regression or convolutional neural networks are generally used in the decision process. Initial research often designs aesthetic features that match it by analyzing some of the adopted photography rules and common perceptual criteria. Datta et al. [15] was the first to start researching this aspect. Sun et al. [16] estimates the distribution of the focus of the visual person based on the global significance region. Luo et al. [5] extracts geometric composition, color harmony, texture definition, illumination and other features in the extracted target area to represent the image. Most of the subsequent manual extraction features are based on content and significantly improve the accuracy of image aesthetic quality assessment.

With the development of deep learning the research work on image aesthetic classification and scoring has entered a new era: the aesthetic characteristics of images are automatically extracted. The researchers applied a variety of volumes

and neural networks for image recognition to aesthetic scores, and the accuracy was much higher than that of hand-designed features. Peng et al. [17] proposed to improve the network structure of AlexNet for emotional classification, style classification and other tasks.

But most deep neural network models require fixed-size inputs, and recent studies have tried to solve this problem. Lu et al. [8] proposes the method that the images of the same picture with different cuts are input into the network in order to obtain global features and local features. Argyriou et al. [18] implements the prediction of any size input by applying a full convolutional neural network, but the training is still fixed in two sizes. Although some recent studies want to eliminate the effects of fixed size, their improvement actually hasn't sufficient effect.

For this issue, we propose our MTP-NasNet method for image aesthetic assessment that can directly manipulate the original size of the image without destroying its beauty. In addition, we introduced multitasking and variance weights into our model to improve predictions. Specifically, our approach modified the NasNet to a double-column CNN that one column handle the full-convolution network and another column adopts the spatial pyramid pooling layer. In order to improve accuracy, we adopted multi-task learning, joint learning of two-category prediction task, style prediction task and scoring distribution prediction task. Furthermore, we consider the influence of variance information on the distribution reliability, and add variance weights to each group of distributions. The experimental results also illustrate the effectiveness of our work.

2.2 Multi-task Learning

For complex problems, they can be divided into simple and independent subproblems, and then merged to finally get the results of complex problems. But in the real world, many problems cannot be broken down into multiple subproblems. In addition, if we treat a real problem as a stand-alone single task, we will ignore the rich information between the questions. Multitasking is born to solve this problem [19]. The associated multitasking learning is better than the single task learning. Since all tasks have more or less noise, for example, when we train the model on task A, our goal is to get a good representation of task A, ignoring data-related noise and generalization performance [19–21]. Since different tasks have different noise modes, learning different tasks at the same time can get a more general representation. Tang et al. [12] proposes a deep identification-verification features for joint training of face recognition loss and face classification loss. Lu et al. [8] has found that there is a close relationship between style and image classification. In this paper, we propose to use multi-task learning to learn the two-category, score distribution and style simultaneously and achieved good results.

2.3 FCN and SPP

In order to avoid the influence of image sizes change, this paper proposes a deep MTP-NasNet method that can handle the issue of different pixel images, thus retain the quality of original images. The design of our MTP-NasNet is inspired by the success of the FCN [13,22] in image semantic segmentation filed and SPP [23] in visual recognition filed.

At present, FCN [13] in the field of image segmentation and SPP [23] in the field of image recognition can theoretically process images of any size input, but they all have their own problems. By increasing the deconvolution layer of the data size, the FCN can output fine results and can feel the details of the picture. However, since FCN only classifies individual pixels and does not fully consider the relationship between pixels and pixels, it lacks spatial consistency. Since SPP uses a plurality of windows to extract features, it is possible to effectively consider spatial information of an image. However, since SPP divides the image from the fine space to the coarse space, it is easy to lack the perception of the detailed information of the image (the theoretical division is very fine and the detailed information can be perceived, but the calculation amount is too large to be realized [22]). In this paper, in order to achieve arbitrary size input and at the same time take into account the spatial information and detail information of the image, we propose a new two-column aesthetic network. Both column can receive image input of any size, one column network uses FCN to extract detail information of the image, and the other column network uses SPP to extract spatial information. We refer to the idea of the two-column network in [8], but our work is very different from his work: first, although their network and our network are both two-column network, but their network input still fixed size to compromise the aesthetics of the image. Second, the network structure inside each channel is different totally. The network proposed in this paper uses SPP to extract spatial information and uses FCN to extract other information.

3 Framework

3.1 Multi-task Learning

we adopts the multi-task learning method to improve our tasks' performance, where each level of the supervised information is formulated as a learning task. This allows our model to share learned features between multiple tasks, making it possible to learn more deep image features by using the additional low-level supervision [19]. This greatly facilitates convergence of the task for aesthetic quality prediction.

The method of multi-task learning we propose is shown in Fig. 3. We share hidden layers between image two-category prediction tasks, score distribution prediction tasks, and style prediction tasks, and retain their respective output layers. By sharing parameters between shared layers, the risk of overfitting can be reduced, and individual parameters of different specified tasks can be trained by separate training of specific output layers. In this experiment, because the

current research gradually shifts from the two-category prediction to the score distribution prediction, we pay more attention to the effect of the score distribution prediction. We refer to the classical method [24–26], taking the score distribution prediction as the core task. In the experiment, three tasks are jointly trained. During the test, only the two-category prediction and the score distribution prediction were tested. The three tasks of our model have a certain hierarchy and range from simple two-category prediction to complex rating distribution prediction. This hierarchy essentially follows the basic procedures of the person to judge the beauty of the picture. People should first be able to have an intuitive judgment of the style of the image, and then have a high or low judgment on its aesthetics, which in turn can have a rough division of the specific score distribution.

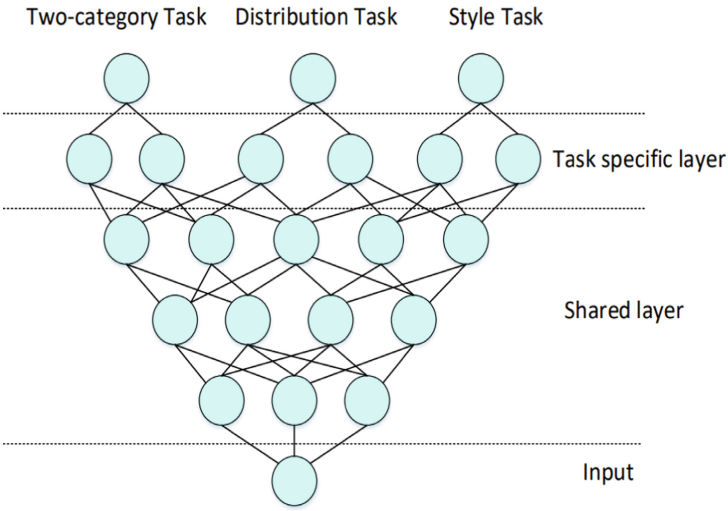


Fig. 3. Our multi-task learning method.

In the jointly training of three tasks, we use the implicit sharing of hidden layer parameters. The most important problem in the specific implementation is the loss function. In the traditional multi-task learning's training process, the importance of all tasks is considered the same [27], but in this model, it is obvious that the importance of the three tasks is different. For example, the aesthetic quality two-category prediction is much more complicated than the style prediction, which leads to learning difficulty and convergence rate are different. We initially tried to add different losses simply. But soon we found that although one task would converge to get good result, others would perform poorly. We found a good method [27], which proposes to introduce uncertainty to determine the weight loss in multi-task learning: learn another noise parameters in the loss function of each task. In this way, we can directly add up to the total loss as

before. In addition, we refer to the proposed multitasking loss function by Liu et al. [28]. Finally our multitasking total loss function and their separate loss functions are as follows:

$$Loss_{total} = \frac{1}{2\sigma_1^2} \cdot L_{dis} + \frac{1}{2\sigma_2^2} \cdot \lambda_1 L_{two} + \frac{1}{2\sigma_3^2} \cdot \lambda_2 L_{sty} + \log \sigma_1 + \log \sigma_2 + \log \sigma_3 \quad (1)$$

where L_{dis}, L_{two} and L_{sty} are the distribution functions of fractional distribution prediction, two-category prediction and style prediction. σ_1, σ_2 and σ_3 are the observation noises of three task scalars respectively, and the range of values is $-1.0 < \log \sigma < 2.5$. λ_1 and λ_2 are the weighting factor of the auxiliary task respectively. After a lot of experiments and repeated tests, their values are finally set as: $\lambda_1 = 0.15, \lambda_2 = 0.06, \log \sigma_1 = 0.5, \log \sigma_2 = 0.9, \log \sigma_3 = 0.6$.

$$\mathrm{L_{dis}}(y, \widehat{y}) = \left(\frac{1}{N} \sum_{k=1}^{N} \left| \mathrm{CDF}_y(k) - \mathrm{CDF}_{\widehat{y}}(k) \right|^r \right)^{1/r} \quad (2)$$

where y and $\hat{y}$ are the truth distribution and predictive distribution, with N ordered classes of distance $\|s_i - s_j\|_r$, $\mathrm{CDF}_{\mathbf{y}}(k)$ is the cumulative distribution function as $\sum_{i=1}^{k} \mathbf{y}_{s_i}$

$$\mathrm{L_{two}}\left(y_i, \hat{y}_i\right) = -\log \left| Softmax\left(y_i, \hat{y}_i\right) \right| \quad (3)$$

$$\mathrm{L_{sty}}\left(y_i, \hat{y}_i\right) = -\log \left| Softmax\left(y_i, \hat{y}_i\right) \right| \quad (4)$$

Adjusting the learning rate in a neural network is one of the most important hyper parameters. So we try to adjust the learning rate. However, there is a particularly suitable learning rate for task A, while the rate of learning is different for another task B. If the learning rate is too large, the gradient will disappear in training. We adjust the learning rate separately in the sub-network of each task, and use another learning rate in the shared network part. In the specific experiment, we set the learning rate to 0.01 in the shared network part, and set the learning rate to 0.001 for the specific task part.

3.2 Double-Column Network

In order to solve the problem of different size input, we have proposed a double-column MTP network to directly manipulate input images of any size without having to change to a fixed size. Figure 4 are our double-column network structure. The first column of the network draws on the ideas of the FCN network. CNN usually consists mainly of two types of layers with weight parameters: convolutional layer and fully connected layer. Among them, the convolutional layer uses the filter sliding serial port to obtain the convolution, and does not require a fixed-size input. However, the fully connected layer requires a fixed length vector as input, resulting in constraints on the fixed size input of the CNN. Inspired by

this, for the first column, we removed the fully connected layer in the original networks, transformed it into a full convolutional network structure, and then replaced the original full layer with a convolution layer with filters that size of 1×1.

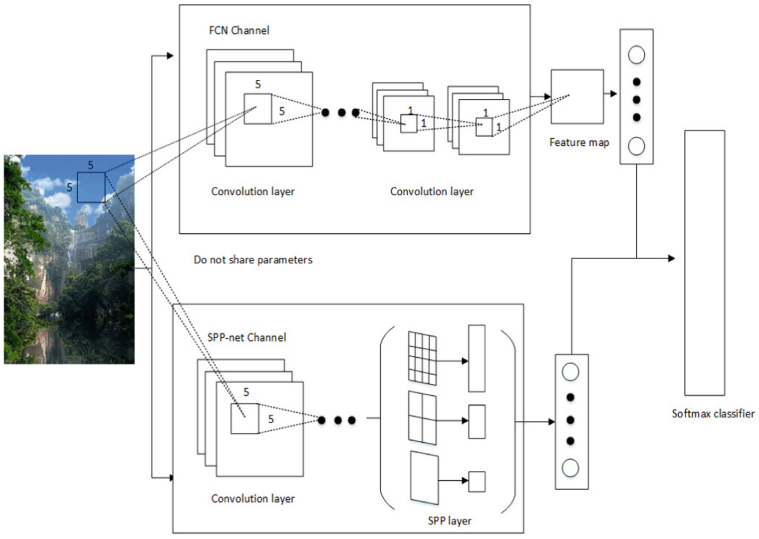


Fig. 4. Our double-column NasNet structure.

Another column of networks refers to the idea of SPP. We add the spatial pyramid pooling layer behind the feature map of the convolutional layer output and use the softmax classifier to get the final probability distribution. Specifically, our experiment uses 1×1, 3×3, and 4×4 three pooling windows to pool the convolved feature maps, merge the results, and then pass the fully connected layer to get the output. Then we combine the outputs of the two columns of channels and get the final prediction result through the softmax classifier. We identify the improved networks based on the original Nasnet as MTF-Nasnet. It is worth noting that in the experiment we used the pre-trained weights on Imagenet as the initial weights. This is because the weight of the original network and the new network we proposed are the same, which makes our training easier and more efficient.

3.3 Variance and Distribution-Aware

In the previous work, we treated all sample images fairly. In reality, however, the score distribution for each image has different variances and medians to indicate the degree of score divergence and the concentration score for most scores, respectively [39]. For the distribution prediction task, if the variance of

the score distribution of a picture is larger, the difference of the score is larger, which indicates that the aesthetic distribution is less credible. Conversely, the smaller the variance, the higher the credibility. For the high and low quality two-category classification task, if the score of a picture is more concentrated at about 5 points, it means that most people think that the aesthetic quality of this picture is medium, then the importance of this image for the classification task is small. Based on the above two points, we add corresponding weights to the training samples in the distributed prediction task and the two-category prediction task to indicate their referability.

4 Experiment

We conducted the tasks of aesthetic distribution prediction and aesthetic quality classification, and compared them with the existing learning methods in these two fields. For the aesthetic distribution prediction task, we mainly compare with the kNN [29], LDSVR [30], SANE [7], IIS-LDL [31] and SVDR [32] methods. The aesthetic distribution prediction task mainly evaluates the performance of different methods by measuring the distance between the predicted distribution and the true distribution of all images. In this experiment, we used several measures: Probability of Euclidean Distance (PED), Chebyshev distance (Cheb), cosine distance (Cos), Probability of Kullback-Leibler divergence (PKL) and Earth Mover's Distance EMD. For the aesthetic quality classification task, we mainly compare with DCNN [8], DMA-Net [6], MNA-CNN [43], and SANE [7]. The measure of performance is the accuracy of the two classifications. Based on the NasNet network, we modified it and obtain the Multi-tasking NasNet (MT-Nasnet) that only adopt the multi-tasking method, Spatial pooling and Fully-convolutional NasNet (SF-NasNet) that only adopt the two column convolutional NasNet, Variance and Distribution-aware NasNet (VD-NasNet) that add corresponding weights and Multi-Task Spatial Pooling Fully Convolutional Neural NasNet (MTP-NasNet) respectively. The experimental results and corresponding comparative analysis are described in detail below.

PED: The loss function using the Euclidean distance of the two probability distribution functions is defined as:

$$l^{PED}(y, \hat{y}) = \sum_{i=1}^{Z} (y(i) - \hat{y}(i))^2 \tag{5}$$

PKL (Wang et al. [23]): The loss function using the symmetrical version of the KullbackLeibler divergence of the two probability distribution functions is defined as:

$$l^{PKL}(y, \hat{y}) = \frac{1}{2} \left[\sum_{i=1}^{Z} y(i) \log \frac{y(i)}{\hat{y}(i)} + \sum_{i=1}^{Z} \hat{y}(i) \log \frac{\hat{y}(i)}{y(i)} \right] \tag{6}$$

Cheb: The Chebyshev distance is a measure derived from a uniform norm (or upper bound norm) and is also a type of injective metric space. It is defined as:

$$l^{Cheb}(y, \hat{y}) = \max_i \left(|y(i) - \hat{y}(i)|\right) \tag{7}$$

Cos: For two n-dimensional sample points a (y1, y2, ..., yn) and b ($\hat{y1}$, $\hat{y2}$, ..., $\hat{yn}$), a similarity to the cosine of the angle can be used to measure the degree of similarity between them. It is defined as:

$$\mathrm{sim}(y, \hat{y}) = \cos\theta = \frac{\boldsymbol{y} \cdot \hat{\boldsymbol{y}}}{\|y\| \cdot \|\hat{y}\|} \tag{8}$$

EMD [9]: EMD is defined as the minimum cost to move the mass of one distribution to another. Given the ground truth and estimated probability mass functions y and $\hat{y}$, with N ordered classes of distance $\|s_i - s_j\|_r$, the normalized Earth Mover's Distance can be defined as:

$$\mathrm{EMD}(\mathbf{y}, \widehat{\mathbf{y}}) = \left(\frac{1}{N} \sum_{k=1}^{N} \left| \mathrm{CDF}_{\mathbf{y}}(k) - \mathrm{CDF}_{\widehat{\mathbf{y}}}(k) \right|^r \right)^{1/r} \tag{9}$$

where $\mathrm{CDF}_{\mathbf{y}}(k)$ is the cumulative distribution function as $\sum_{i=1}^{k} \mathbf{y}_{s_i}$.

4.1 Datasets

We trained our different models on the AVA dataset. The AVA dataset is a large-scale image aesthetic quality dataset from Murray, which contains 255,530 images downloaded from the online image sharing scoring website (dpchallenge.com). This datasets is a recognized benchmark set in the field of image aesthetic evaluation. In this experiment, 200,000 picture were randomly selected as the training sets, the 25,000 of rest pictures were value sets and the rest 25,000 pictures were test sets. So it is set to 80% training, 10% valuing and 10% testing.

4.2 Distribution Predicting Results

On the basis of NasNet [33], we have proposed the MT-NasNet, SF-NasNe, VD-NasNet and the MTP-NasNet. Our improved models were tested on the AVA dataset and compared with the work related to the distribution prediction. The experimental results are shown in Table 1. The evaluation indicators evaluated were PED, Cheb, Cos, PKL and EMD. Among them, the smaller the PED, Cheb, PKL and EMD, the better the performance of the distribution prediction, and the larger the Cos, the better the performance of the distribution prediction.

It can be seen that, obviously, the network we designed is superior to other competitors in all evaluations. The best-performing MTP-NasNet increased by 18.11%, 17.44%,26.6% 12.77%, 1.57%, and 39.24% on PED, Cheb, PKL, Cos, and EMDrespectively. Therefore, we can confirm that MTP–NasNet has achieved

the best results on the aesthetic distribution forecast. In addition, through observation, it can be found that SANE has also achieved good results as a method that can also accept multi-scale input. SVDR performs poorly in most cases, indicating that its defined loss function does not effectively distinguish the aesthetic distribution of images. This finding suggests that this computationally complex structured learning is less suitable for aesthetic distribution prediction. IIS-LDL is not very effective due to its difficulty in convergence. Recent studies have also shown that this algorithm is an extremely low-efficiency entropy model in parameter estimation. KNN directly minimizes the distance between the predicted distribution and the real distribution, and has achieved good results, but is not effective than LDSVR.

Table 1. The result of different methods for aesthetic distribution prediction on AVA datasets

Model	PED	Cheb	PKL	Cos	EMD
IIS-LDL	0.215	0.154	0.213	0.886	0.132
KNN	0.172	0.113	0.176	0.918	0.095
LDSVR	0.153	0.100	0.139	0.934	0.083
SVDR	0.381	0.294	0.143	0.820	0.126
SANE	0.127	0.086	0.094	0.958	0.079
MT-NasNet	0.118	0.081	0.091	0.963	0.064
SF-NasNet	0.116	0.079	0.086	0.966	0.057
VD-NasNet	0.123	0.084	0.092	0.959	0.072
MTP-NasNet	**0.104**	**0.071**	**0.082**	**0.973**	**0.048**

4.3 Two-Category Prediction

Our improved models were tested on the AVA dataset and compared with the work related to the two-category quality prediction. The experimental results are shown in Table 2. The evaluation index of the evaluation is the two-category accuracy rate. We use the trained model to predict the aesthetic distribution of the image, and then use the average of the aesthetic distribution as its quality score. In our work, we judge photos with scores less than $5 - \delta$ points as low-quality photos and photos above $5 + \delta$ as high-quality photos. In our experiment, δ is set to 1. By comparing the prediction category with the category of the real distribution, we can get the accuracy of data sets.

It can be seen that, obviously, the network we designed is also slightly better than other competitors in the accuracy evaluation. It can be seen that the current methods have achieved quite high accuracy, and the method proposed in this paper has only achieved a small improvement. The network structure of SANE

Table 2. The result of different methods for aesthetic two-category prediction on AVA datasets

Model	Accuracy
DMA-Net	80.12%
DCNN	88.01%
SANE	96.71%
MNA-CNN	95.76%
MT-NasNet	96.84%
SF-NasNet	96.95%
VD-NasNet	96.79%
MTP-NasNet	**97.34%**

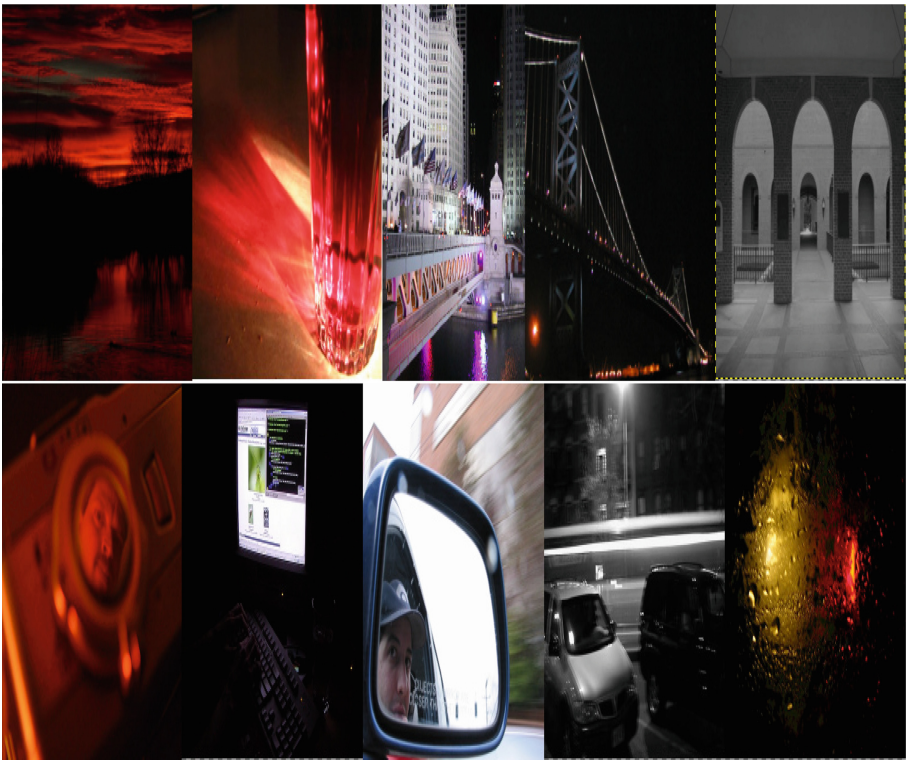


Fig. 5. Images predicted with the higher and lower aesthetic rating in the testing set.

and MNA-CNN still has high reference value. Figure 5 shows examples of high-quality images and low-quality images of our proposed model on the test set. It can be seen that high-quality images have finer quality and more aesthetic layout than low-quality images. It also illustrates the effectiveness of our model.

4.4 Effect of Input Size Reserving

In our experiments, we used NasNet as the initial network and get our new MTP-NasNet by transforming it. In order to evaluate the effectiveness of our network structure for any size input, we compared the network with three operations of cropping, warpping and padding. Among them, the input of the NasNet-crop network is clipped to a fixed size of 224×224; the input of the NasNet-wrap network is scaled to a fixed size of 224×224. The input to the NasNet-pad network is scaled down to 224 on the long side and then zeroed to 224×224. Table 3 shows the comparison between our experimental results and these three methods. The experimental results show that these three operations do have a certain negative effect on the experimental results. In addition, we found that the effect of NasNet-wrap is the best of the three networks, probably because direct scaling does not reduce the information of the original image, and the retention is relatively complete.

Table 3. Compare between our method and the baseline methods with fixed-sized inputs

Model	Euc	Cheb	KL	Cos	EMD
NasNet-Crop	0.137	0.097	0.131	0.933	0.089
NasNet-Wrap	0.134	0.092	0.125	0.949	0.087
NasNet-Pad	0.141	0.101	0.136	0.930	0.094
MTP-NasNet	**0.104**	**0.071**	**0.082**	**0.973**	**0.048**

4.5 Implementation Details

We used the deep learning platform Keras to implement network training and testing. Our network uses the original Nasnet to pre-train the weights on Imagenet for initialization. All experiments were performed on a workstation equipped with a 16-core 2.8 GHz Intel Xeon processor, two Nvidia GTX 1080Ti GPUs, and 256 G RAM. The implementation details is below:

Training size: In theory, our method can accept images of any size as input, but in fact, one is too much calculation for training, and the other is not easy to optimize and parameter transfer. Therefore, we have adopted a multi-scale training method for training to simulate original results. We counted the aspect ratios of all the images in the dataset and found that they can be roughly divided into 1:1.5, 1.5:1, and 1:1. Therefore, we have selected 224×336, 336×224, and 224×224 as the predetermined sizes. Different epoch turns to unify the pictures to different sizes in training, so that the network can also learn the concept of variable size. In the test phase, we handle images of any size directly.

Regularization: In our experiment, Adam was used as the optimization function, and the batch size was set to 128. The baseline NasNet weights are initialized by training on the ImageNet [34], and the last fully-connected layer is randomly initialized. Mostly, the learning rate was set to 0.001. In the experiment, our training generally converged around 70 epoches and took nearly 3 days.

Data enhancement: At the beginning we did not adopt a data enhancement method and produced an overfitting. Later, we adopted a data enhancement method of horizontal flipping, vertical flipping, and rotation, which expanded the scale of the data set and achieved better results.

5 Conclusion

Transforming the input image to a fixed size that causes aesthetic damage is an important issue in the field of aesthetic quality evaluation. To solve this problem, this paper proposes a new end-to-end deep double-column network structure. Through this double-column network structure based on SPP and FCN, we can not only operate input images of any size, but also extract spatial information and detailed information of images at the same time. In addition, we have effectively improved the prediction effect through multi-task learning. Further, we consider the information such as the variance in the score distribution, and enhance the learning effect by weighting the samples. The results on AVA's large datasets illustrate the effectiveness of our improvements and the importance of arbitrary size input and multitasking learning. Next we will delve into the important factors that affect the aesthetics of the image and introduce it into our network to optimize the model.

Acknowledgment. The paper was supported by NSFC under Grant 61471234, 61771303, and Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 18DZ1200102

References

1. Murray, N., Marchesotti, L., Perronnin, F.: Ava: a large-scale database for aesthetic visual analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2408–2415. IEEE (2012)
2. Marchesotti, L., Perronnin, F., Larlus, D., Csurka, G.: Assessing the aesthetic quality of photographs using generic image descriptors. In: 2011 International Conference on Computer Vision, pp. 1784–1791. IEEE (2011)
3. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. J. Electron. Imaging **19**(1), 011006 (2010)
4. Yin, W., Mei, T., Chen, C.W., Li, S.: Socialized mobile photography: learning to photograph with social context via mobile devices. IEEE Trans. Multimed. **16**(1), 184–200 (2013)
5. Luo, Y., Tang, X.: Photo and video quality evaluation: focusing on the subject. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5304, pp. 386–399. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88690-7_29
6. 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, pp. 990–998 (2015)

7. Cui, C., Fang, H., Deng, X., Nie, X., Dai, H., Yin, Y.: Distribution-oriented aesthetics assessment for image search. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1013–1016. ACM (2017)
8. Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z.: Rating image aesthetics using deep learning. IEEE Trans. Multimed. **17**(11), 2021–2034 (2015)
9. Talebi, H., Milanfar, P.: Nima: Neural image assessment. IEEE Trans. Image Process. **27**(8), 3998–4011 (2018)
10. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments (2008)
11. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 94–108. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_7
12. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, pp. 1988–1996 (2014)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
14. Brandão, T., Queluz, M.P.: No-reference quality assessment of H. 264/AVC encoded video. IEEE Trans. Circuits Syst. Video Technol. **20**(11), 1437–1447 (2010)
15. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Studying aesthetics in photographic images using a computational approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 288–301. Springer, Heidelberg (2006). https://doi.org/10.1007/11744078_23
16. Sun, X., Yao, H., Ji, R., Liu, S.: Photo assessment based on computational visual attention model. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 541–544. ACM (2009)
17. Peng, K.C., Chen, T.: Toward correlating and solving abstract tasks using convolutional neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. IEEE (2016)
18. Cui, C., Liu, H., Lian, T., Nie, L., Zhu, L., Yin, Y.: Distribution-oriented aesthetics assessment with semantic-aware hybrid network. IEEE Trans. Multimed. **21**(5), 1209–1220 (2018)
19. Evgeniou, T., Pontil, M.: Regularized multi-task learning. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 109–117. ACM (2004)
20. Jebara, T.: Multitask sparsity via maximum entropy discrimination. J. Mach. Learn. Res. **12**, 75–110 (2011)
21. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Mach. Learn. **73**(3), 243–272 (2008)
22. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: the all convolutional net. arXiv preprint arXiv:1412.6806 (2014)
23. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(9), 1904–1916 (2015)
24. Liu, X., Gao, J., He, X., Deng, L., Duh, K., Wang, Y.Y.: Representation learning using multi-task deep neural networks for semantic classification and information retrieval (2015)

25. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
26. Yang, Y., Hospedales, T.M.: Trace norm regularised deep multi-task learning. arXiv preprint arXiv:1606.04038 (2016)
27. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7482–7491 (2018)
28. Liu, P., Qiu, X., Huang, X.: Adversarial multi-task learning for text classification. arXiv preprint arXiv:1704.05742 (2017)
29. Geng, X.: Label distribution learning. IEEE Trans. Knowl. Data Eng. **28**(7), 1734–1748 (2016)
30. Geng, X., Hou, P.: Pre-release prediction of crowd opinion on movies by label distribution learning. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
31. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE Trans. Pattern Anal. Mach. Intell. **35**(10), 2401–2412 (2013)
32. Mai, L., Jin, H., Liu, F.: Composition-preserving deep photo aesthetics assessment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 497–506 (2016)
33. Wu, O., Hu, W., Gao, J.: Learning to predict the perceived visual quality of photos. In: 2011 International Conference on Computer Vision, pp. 225–232. IEEE (2011)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)