# *Sample-specific repetitive learning for photo aesthetic auto-assessment and highlight elements analysis*

**Ying Dai**


**Multimedia Tools and Applications**
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-020-09426-z

Springer

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09426-z

# Sample-specific repetitive learning for photo aesthetic auto-assessment and highlight elements analysis

Ying Dai[1]

Received: 18 September 2019 / Revised: 10 June 2020 / Accepted: 21 July 2020
Published online: 08 September 2020


**Abstract**
Aesthetic assessment is subjective, and the distribution of the aesthetic grades is over-concentrated in the middle levels. In order to realize the auto-assessment of photo aesthetics, we focus on using repetitive self-revised learning (RSRL) to retrain the convolutional neural network (CNN)-based aesthetics prediction network repetitively by the transfer learning, so as to improve the performance of imbalanced learning caused by the overconcentration distribution of aesthetic scores utilized as learning data. As RSRL, the network is trained repetitively by dropping out the low likelihood photo samples at the middle levels of aesthetics from the training data set based on the previously trained network. Further, the two retained networks are used in extracting aesthetic highlight elements of the photos to analyze the relation of the photo composition with the aesthetic assessment. The objective and subjective experimental results show that the CNN-based RSRL is effective for improving the performances of the imbalanced scores prediction network for the photos aesthetic auto-assessment.

**Keywords** Photo aesthetic auto-assessment · Imbalanced learning · Repetitive self-revised learning · Dropping out sample · CNN

## 1 Introduction

In response to the growth of digital camera, more and more pictures are taken to upload the social media. Many people hope to improve aesthetic level of themselves by taking beautiful photographs. So, auto-assessment of photo aesthetics is challenging. Researches have been investigating methods for providing automated aesthetical evaluation and classification of photographs. Aesthetic assessment is subjective. One of the main difficulties in addressing this challenge is in developing formal models of human aesthetic preference [1]. In this paper, authors stated that such models would allow computer systems to predict the aesthetic taste of

✉ Ying Dai
dai@iwate–pu.ac.jp

[1] Iwate prefectural university, Takizawa, Japan

a human being or adapt to the aesthetic tendencies of a human group. For making aesthetics automatic evaluation and choices, the best way to proceed is to create datasets for training the model in collaboration psychology aesthetics (PA) researchers, because computational aesthetics (CA) research typically reposts results using a success rate, while psychologists are more likely to use correlation. Closer collaboration between CA and PA can give rise to results that advance both disciplines. In [11], recent computer vision techniques used in the assessment of image aesthetic quality were reviewed. The authors discussed the possibility of manipulating the aesthetics of images through computational approaches. The research reviewed in the paper generally aims at assessing the aesthetic quality of photos with aesthetic scores or distinguishing high-quality photos form low-quality photos, by training the photo aesthetic models based on the deep learning techniques. However, such models can't interpret which salient image composition features and highlight regions are correlated with the photo aesthetics. Moreover, who labeled the aesthetic scores of training data set for deep learning, professional photographer or amateur, is unclear. In [7], a set of features derived from both low- and high-level analysis of photo layout were exploited to perform the aesthetic quality evaluation by a support vector machine (SVM) classifier. In [2], authors designed a set of compact rule-based features based on photographic rules and aesthetic attributes, and used deep convolutional neural network (DCNN) descriptor to implicitly describe the photo quality. These approaches focused on extracting the handcrafted image features. However, the effectiveness is limited because extracting the features is based on the researchers' understanding on the aesthetic rules. In [5], the images were divided into three categories: "scene", "object" and "texture". Each category has an associated convolutional neural network (CNN) which learns the aesthetic features for the category classification. In [10], a scene convolutional layer was designed to learn specific aesthetic features for various scenes by deep learning model. In [9], a novel photograph aesthetic classifier with a deep and wide CNN for fine-granularity aesthetical quality prediction was introduced. However, the correlation of the extracted features with the photo aesthetic assessment was not interpreted in the view of PA in such research. In [6], the percentage distributions for orientation, curvature, color and global symmetry were extracted and fed to a deep neural network under the form of only 114 inputs. Differences in extracted features between aesthetically good and poor images were analyzed and some human aesthetic preferences in static two-dimensional scenes were observed. However, the issue whether the handcrafted features are generic for the photo aesthetic assessment is not involved. Moreover, all of the above approaches were not involved in the issue that the aesthetic rating is ambiguous and is different from person to person, which caused a highly imbalanced distribution of aesthetic ratings. Toward to tackling these issues, authors in [4] showed how to learn deep features for imbalanced data classification. Using the learned features, the classification was simply achieved by a fast cluster-wise kNN search followed by a local large margin decision. In [12], authors proposed an end-to-end CNN model which simultaneously implements aesthetic classification and understanding. A sample-specific classification method that re-weights samples' importance is implemented, and what is learned in the deep model was investigated. Ambiguous samples are given lower weights while clear samples are weighted high. However, the method to give the weight of every sample was not explicit, and the improvement for the imbalanced data classification was not salient from the experiment results. Further, the correlation of the learned deep features with the aesthetic assessment was not analyzed although deep activation map was visualized.

Motivated by the above research, the author aims at collecting the photos scored by professional photographers who could be considered as PA researchers. Such photos are used

as the training data set to learn the aesthetic assessment model. Now, about 3100 photos are scored aesthetically by a professional photographer. These photos were taken by the students of the photographer's class. The scores are in the range of [4, 11]. The photos with score 2 or less are aesthetically very poor; those with score 3 are poor; those with score 4 are fair; those with score 5 are good; those with 6 or more are excellent. The data set indeed exhibited a highly non-uniform distribution over score as illustrated in Fig. 1. The majority images concentrate on the values of 3 to 5 (more than 87%). The model could be overwhelmed by those general samples if the parameters are learned by treating all samples equally, and the more salient samples couldn't decide how the model is trained.

Accordingly, in order to solve the training data imbalance issue in aesthetic assessment, in this paper, the author focuses on using repetitive self-revised learning (RSRL) to retrain the CNN-based aesthetics prediction model repetitively by transfer learning, so as to improve the performance of imbalanced learning caused by the overconcentration distribution of aesthetic scores utilized as learning data. As the repetitive self-revised learning, the network is trained repetitively by dropping out the low likelihood photo samples scored in the range of [5, 7] from the training data set based on the previously trained model. Further, the two retained networks are used in extracting aesthetic highlight elements of the photos to analyze the correlation of the photo composition with the aesthetic assessment. The objective and subjective experimental results show that the CNN-based repetitive self-revised learning is effective for improving the performances of the imbalanced score prediction network for the photos aesthetic auto-assessment.

## 2 Related work

The photos' aesthetic level assessment exhibit highly-skewed score distribution as shown in Fig. 1. As described in [4], for such class-imbalanced data, the minority class often contains very few instances with high degree of visual variability. The scarcity and high variability make the genuine neighborhood of these instances easy to be invaded by other imposter nearest neighbors. To mitigate this issue, contemporary classification methods typically follow classic strategies such as re-sampling or cost-sensitive training. In [4], with validating the effectiveness of these classic schemes for representation learning on class-imbalanced data, the authors demonstrate that more discriminative deep representation can be learned by enforcing

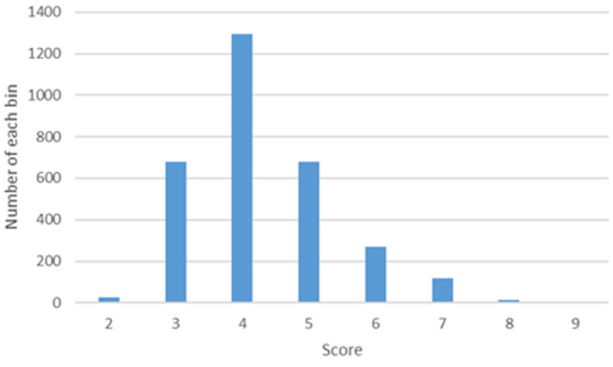


**Fig. 1** Score distribution of photo data set

a deep network to maintain bother inter-cluster and intra-cluster margins. This tighter constraint effectively reduces the class imbalance inherent in the local data neighborhood. However, the proposed method is suitable for the classification of objects such as faces or characters. Whether this method is available for the imbalanced prediction of the subjective aesthetic assessment is unclear. Further, the proposed method of a data structure-aware deep learning approach with build-in margins seems difficult to deploy in the photo aesthetics prediction.

In [3], a comprehensive literature survey to tackle the class data imbalance learning problem was reviewed. Generally, there are two groups of solutions: data re-sampling and cost-sensitive learning. The former group focuses on learning equally good classifiers by random under-sampling and over-sampling; informed under-sampling; synthetic sampling with data generation; adaptive synthetic sampling; sampling with data cleaning techniques; cluster-based sampling; and integration of sampling and boosting. The latter group operates at the algorithmic level by adjusting misclassification. It targets the imbalanced learning problem by using different cost matrices that describe the costs for misclassifying any particular data example. The cost matrix can be considered as a numerical representation of the penalty of classifying examples from one class to another.

A well-known issue with over-sampling is its tendency to overfitting. Therefore, under-sampling is often preferred, although potentially valuable information may be removed. Cost-sensitive alternatives avoid these problems by directly imposing heavier penalty on misclassifying the minority class. However, those literature methods mainly aim at the classification of the classes which are defined explicitly. For the scored prediction of the aesthetic assessment, it seems difficult to obtain the cost matrix used for imposing penalty on misclassifying the minority class.

## 3 Repetitive self-revised learning

In this paper, the author combines the ideas of random re-sampling and cost-sensitive learning to propose a scheme of solving the issue of imbalanced learning in the photos' score prediction for the aesthetic auto-assessment. The training data are not resampled randomly. The samples having low likelihoods of majority classes are dropped out from the training data set, while the likelihoods are calculated by the currently trained prediction model. The idea behind this scheme is the assumption that the sample calculated with the low likelihood to the majority classes is what is not scored precisely. These samples are easy to invade the genuine neighborhood of samples in the minority classes.

Based on this scheme, A CNN-based repetitive self-revised learning (RSRL) approach is considered by repetitively dropping out the low likelihood samples of majority classes defined by scores, so as to ameliorate the invasion of these samples to the minority classes, and prevent the loss of the samples with discriminative features in the majority classes. The scheme diagram is shown in Fig. 2.

The training data set is the score-labeled photo data set. The aesthetic scores in the range of [1, N] are given by the pro-photographer. The scores' distribution of samples in the data set is as shown in Fig. 1. The samples are almost concentrated in the fair classes with score 3 to 5. Score $n$ is handled as one class which is denoted $s_n$, while the number of classes are $N$.

The photos' imbalanced score prediction for the aesthetic assessment is tackled by the CNN-based network. The network's training begins from the pre-trained network such as

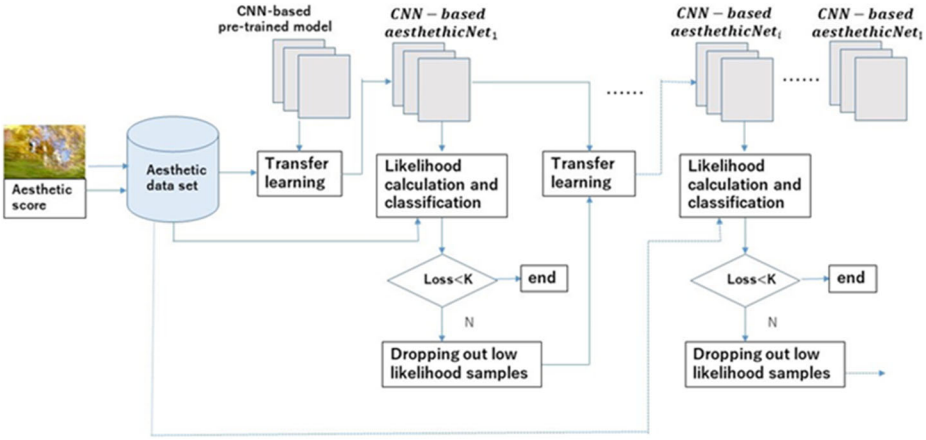


**Fig. 2** Diagram of training photo Aesthetic assessment model

alexNet by transfer learning. The last three layers of the pre-trained network are tuned to the score classes. By replacing the last three layers of the pre-trained network, the network to predicting photos' scores instead are fine-tuned by feeding the training data set. The initially learned network is called $aestheticNet_0$. The transferred network architecture is as the following.

*1--end-3 layers: pre-trained network layers' Transferring*

*end-2 layers 'fc': full connected layer, N nodes, each corresponding to one class of score*

*end-1 layer 'softmax': softmax layer, N nodes*

*end layer 'classoutput': prediction output*

Then, each node of 'fc' layer for a sample is activated to get the value, denoted as $x_{s_n}$. So, the corresponding Sigmoidal fuzzy membership value, denoted as $fc_{s_n}$, is calculated by the following Eq. (1).

$$fc_{s_n} = \frac{1}{1 + e^{-x_{s_n}}} \tag{1}$$

For the sample, the value of $fc_{s_n}$ could be treated as its likelihood belong to $s_n$. Based on $fc_{s_n}$, the samples in the majority classes, which have low likelihood to $s_n$, are dropped out from the training data set. Such picking out the low likelihood samples is considered to be a process of self-revision of the data set. The idea behind the method is that the samples labeled with the scores of the majority classes while having the low likelihood to them may become the imposter nearest neighbors of the samples in the minority classes to invade the genuine neighborhood of the minority classes. Accordingly, the conditions of dropping-out the low likelihood samples of the majority classes are expressed by the following equations, if the number of the majority classes is considered to be 3.

$$Droping\ out\ \text{a sample}\ i_{s_{max1}}, if\ its\ fc_{s_{max1}} < K1 \tag{2}$$

$$Droping\ out\ a\ \text{sample}\ i_{s_{max2}}, if\ its\ fc_{s_{max2}} < K2 \tag{3}$$

$$Droping\ out\ a\ \text{sample}\ i_{s_{max3}}, if\ its\ fc_{s_{max3}} < K3 \tag{4}$$

Where, the class having most samples is denoted $s_{max1}$, the next is $s_{max2}$, the third is as $s_{max3}$. The sample which is labeled with $s_{max1}$, denoted $i_{s_{max1}}$, is removed from the training data set if the corresponding $fc_{s_{max1}}$ is less than $K1$. And so on, the sample which is labeled with $s_{max2}$ or $s_{max3}$,denoted $i_{s_{max2}}$ or $i_{s_{max3}}$,is dropped out from the training data set if the corresponding $fc_{s_{max2}}$ is less than $K2$, or $fc_{s_{max3}}$ is less than $K3$.

Then, based on the previously trained network $aestheticNet_0$, the network is retrained with transfer learning using the self-revised training data set that dropped out the low likelihood samples to the majority classes based on $aestheticNet_0$. The correspondingly generated network is called the retrained network $aestheticNet_1$. Based on $aestheticNet_1$, the likelihoods of all samples of the original training data set belonging to each score class is calculated. The samples are dropped out if they apply to the above removing conditions (2), (3) or (4). It is noticed that the samples that were dropped out based on the previously trained network could be pulled back in the currently self-revised training data set based on the current network, so as to void in removing the samples with discriminative features for classification. We call such learning process as self-revised learning.

Repetitively, based on the latest retrained network $aestheticNet_{i-1}$, the network $aestheticNet_i$ is retrained by transfer learning with the latest training data set dropping out the low likelihood samples of major classes based on$aestheticNet_{i-1}$ . This learning procedure is continued until the total F-measure of all classes regarding the test data set reach the optimal value, while the F-measure of the class $F_{s_n}$, the total F-measure of all classes $F_{all}$ are calculated by the following Eqs. (5) and (6).

$$F_{s_n} = \frac{2*precision*recall}{precision + recall} \tag{5}$$

$$F_{all} = \sum_{n=1}^{N} F_{s_n} \tag{6}$$

Where, *precision* indicates the precision of a class, and *recall* indicates the recall of the class.

Then, the corresponding generated network is called $aestheticNet_I$, which is utilized as the optimal score prediction model. The index $I$ of this optimal network is determined by the following equation, which make the value of $F_{all}$ maximal.

$$\mathrm{I} = \underset{i \in N}{\operatorname{argmax}}\ F^i_{all} \tag{7}$$

## 4 Extracting aesthetic highlight elements

The retrained CNN-based networks are consided to extract photos' aesthetic highlight, so as to analyze how the photos are assessed by the photographer to investigate the composition of the

good photos, and the correlation of the salient objects in the photos with the background. The diagram of the proposed method is shown in Fig. 3.

For a photo image, feature maps of 'conv1' layer of two networks are activated. These two networks are the optimal retrained network *aestheticNet*$_I$and the previously retrained network *aestheticNet*$_{I-k}$.The feature maps are denoted by $\mathrm{conv1}_I^j$ and $\mathrm{conv1}_{I-k}^j$, respectively. *I-k* indicates the *(I-k)th* retrained network, and *j* indicates the *j*th feature map. It is assumed that the feature difference map of the two corresponding feature maps, which are of the minimal correlation, could reflect the aesthetic highlight elements. The idea behind is that the training data set is aesthetically labeled by the photographer who often focuses on the aesthetic highlight elements which embody photo's aesthetic level based on the essential principles, such as whether the object in the photo is distinctive, and whether the composition is concise. So, the activated values of the highlight elements should change more greatly if the network is retrained by removing the low likelihood samples. Therefore, the feature difference map could be used to extract such aesthetic highlight elements.

The index *J* of the feature maps which are of the minimal correlation is identified by the following equation.

$$\mathrm{J} = \underset{j \in N}{argmin}\left(corr\left[conv_I^j,\ conv_{I-k}^j\right]\right) \tag{8}$$

Where, *corr* indicates the manipulation of correlation. The feature difference map of the two feature maps with index *J* could be calculated by subtracting $\mathrm{conv1}_I^J$ and $\mathrm{conv1}_{I-k}^J$, expressed by the equation below.

$$diff_{I,I-k} = \mathrm{conv1}_I^J - \mathrm{conv1}_{I-k}^J \tag{9}$$

Figure 4 shows the feature difference maps of two photos. The left column are original images; the middle are feature maps regarding $\mathrm{conv1}_I^J$ and $\mathrm{conv1}_{I-1}^J$, respectively; the right are the feature difference maps of them, denoted $diff_{I,\,I-1}$. Here, I = 12.

For the upper example, the salient object bird as a highlight is emerged explicitly in the feature difference map; for the lower example, the mountain area is emerged in the feature difference map although the highlight of this sample is not obvious.

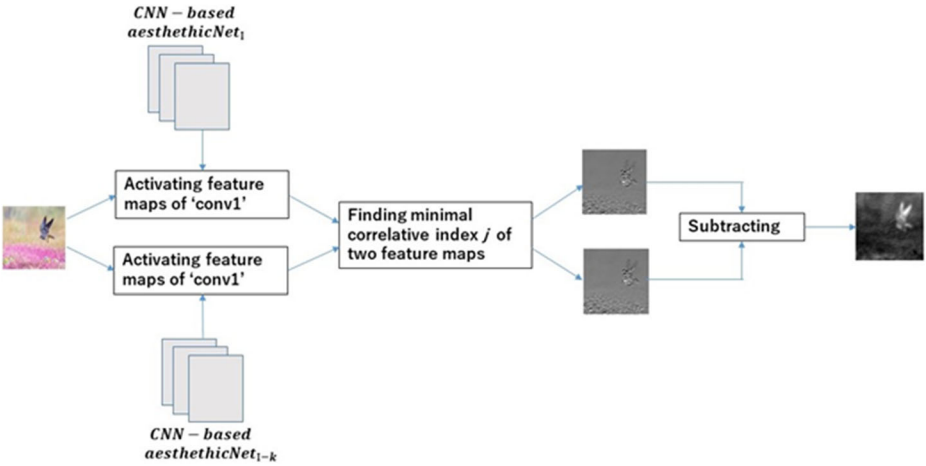


**Fig. 3** Diagram of extracting aesthetic highlight elements

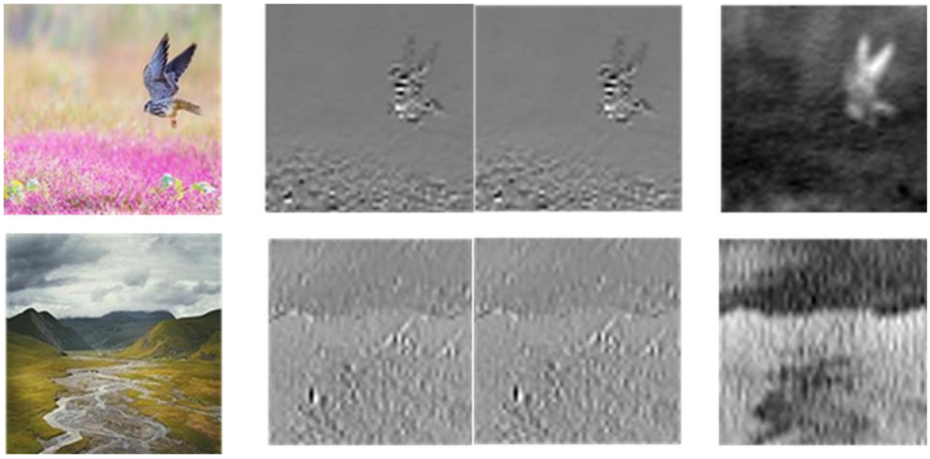

**Fig. 4** Calculating feature difference map

## 5 Experiments and analysis

### 5.1 Data set

Although the AVA data set [8] is the largest publicly available aesthetics dataset providing over 250,000 images in total, each image in which was aesthetically assessed by about 200 people with the rating score ranging from 1 to 10, all of the images were finally labeled with the mean score that lost the individual's aesthetic sense although the aesthetic tendencies of a human group could be reflected. However, embodying the aesthetic taste of a human being is important in training the imbalanced score prediction network for the photos' aesthetic auto-assessment. Therefore, we conduct our data set which contains 3100 photos assessed aesthetically by a professional photographer, which is called xiheAA. Those photos were taken by the students of the photographer's class. The scores range from 2 to 9. So, the number of classes $N = 8$. The distribution of the scores is shown in Fig. 1. The class having most samples is score 4; the next is score 3; the third is score 5.

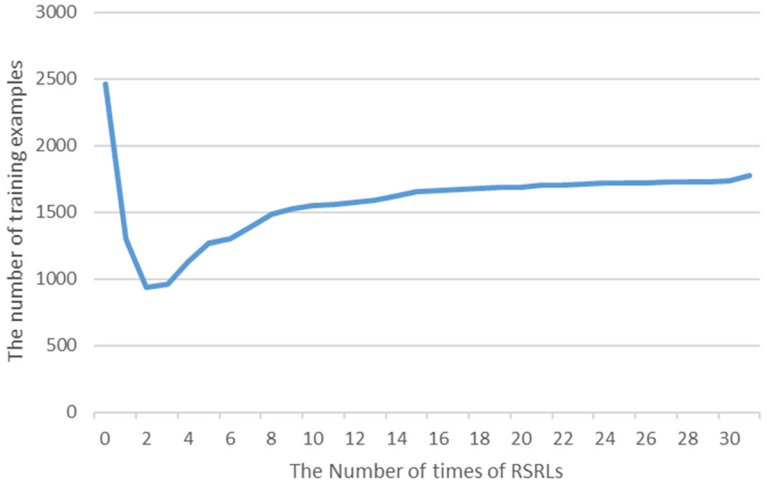


**Fig. 5** Change of sizes of training data set with RSRL

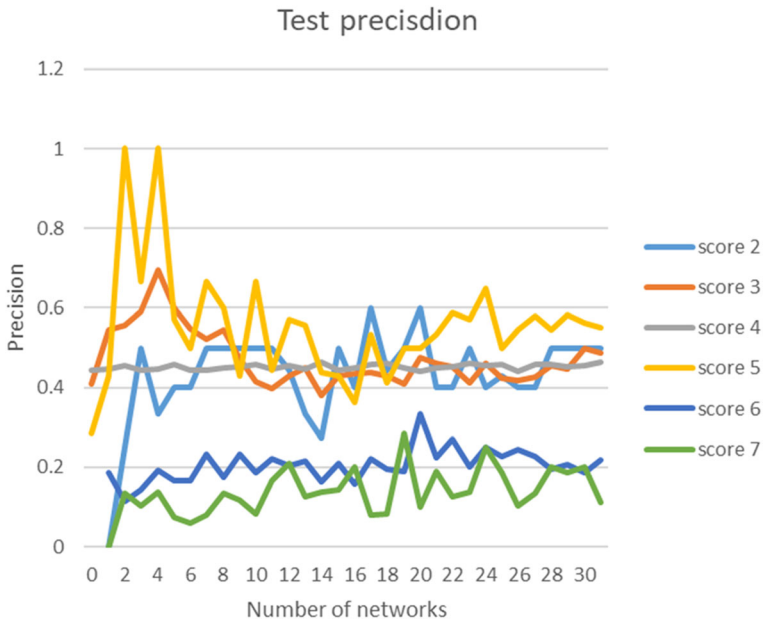


**Fig. 6** Test precision

For the 5-fold cross validation, four of fifth samples are selected randomly as the training dataset, and the rest samples are used as the test data set. That is, the size of training data set is 2480 samples, and the size of test data set is 620 samples.

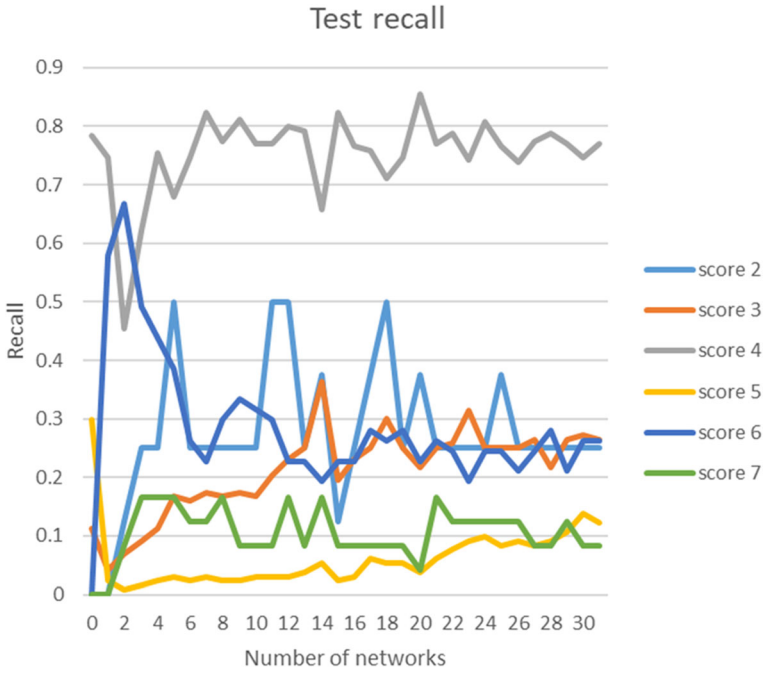


**Fig. 7** Test recall

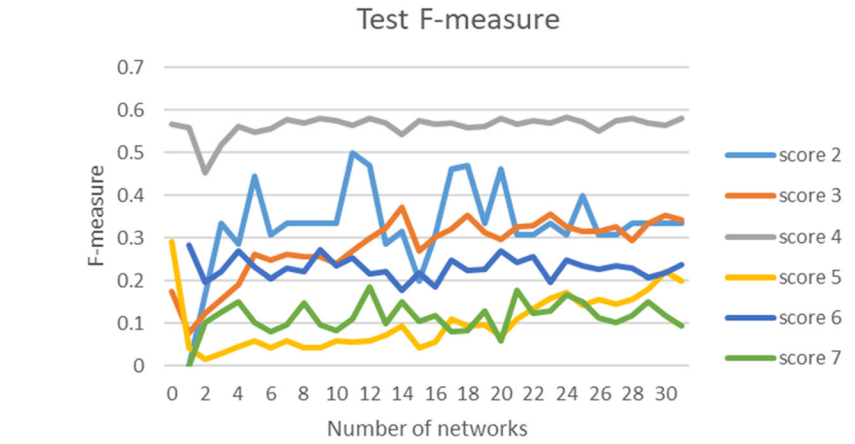


**Fig. 8** Test F-measure

### 5.2 Imbalanced score prediction

In this section, we evaluate how the RSRL improve the performance of the imbalanced score prediction for the photos' aesthetic assessment. The approach proposed in section 3 is implemented on Matlab. The alexNet neural network is used as the pre-trained network. The function *trainNetwork* is utilized to fine-tune the weights of the CNN-based pre-trained network by inputting the self-revised training data set to obtain the novel prediction network. The epoch is set as 5. The function *activation* is adopted to activate the nodes values of layer 'fc' to obtain the likelihoods of samples to each score class. The function *classify* is used to assign the test samples to the corresponding score classes based on the retrained prediction networks.

Figure 5 shows the change of sizes of the training data set caused by RSRL. The conditions dropping out the low likelihood samples are based on the expressions (2), (3), or (4). There, the

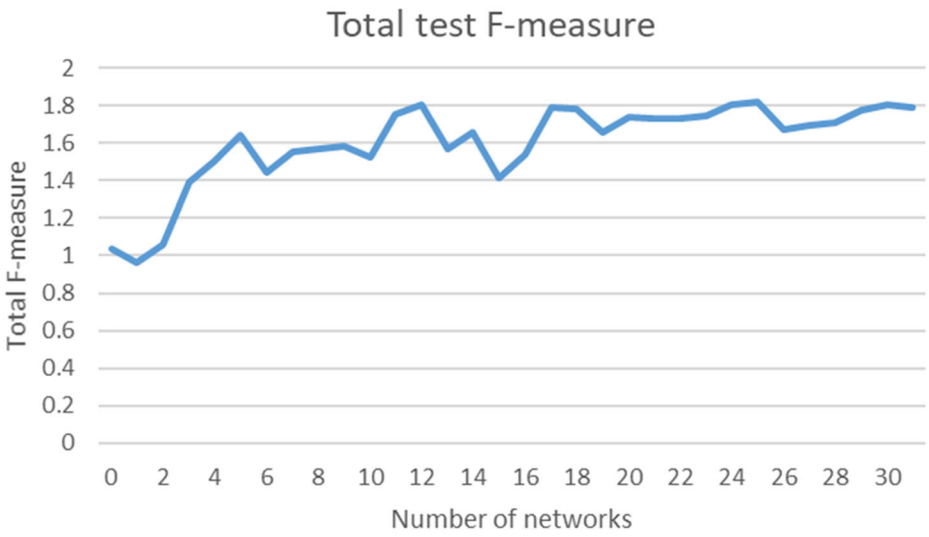


**Fig. 9** Change of total F-measures with RSRL

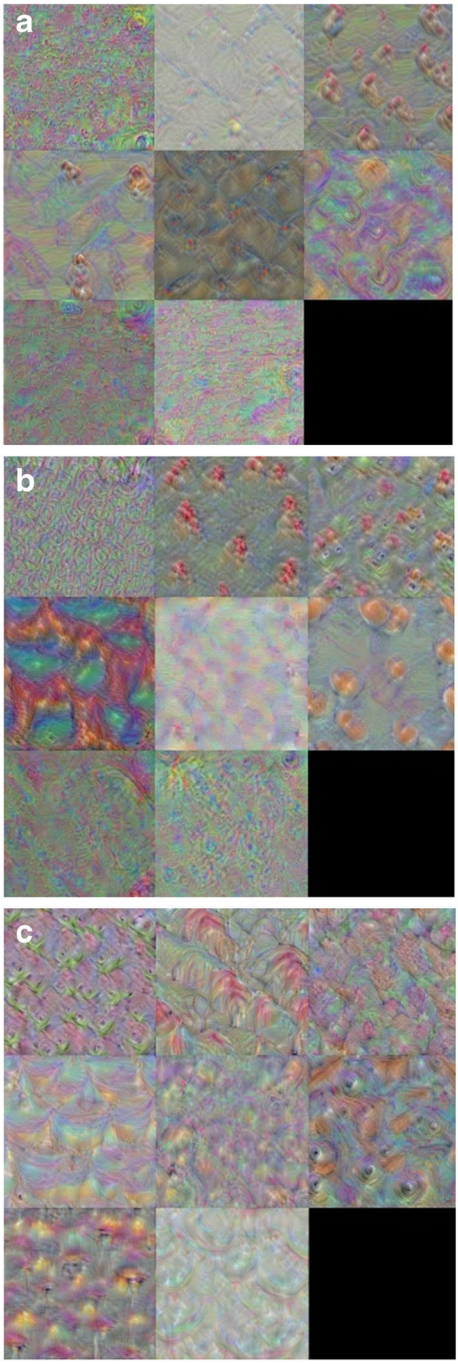


**Fig. 10** Feature maps of the end-2 layer 'fc'

values of K1, K2, and K3 are firstly set as 0.9, respectively. After the 2th round of RSRLs, the values of K1, K2, and K3 are adjusted to 0.95, respectively.

**Table 1** Total F-measure of alexBase and 2convCNN

| | Total F-measure | |
|---|---|---|
| | initial | optimal |
| alexBase | 1.02 | 1.81 |
| 2convCNN | 0.74 | 1.22 |

The size of the original training data set is about 2500 samples. The initial size of the self-revised training data set is about 1300 samples for the first round self-revised learning. About 1/2 samples in the original training data set are removed, including the samples with discriminative features. Then, the size decreases to about 950 samples for the second round RSRL, because the threshold of dropping out samples is varied from 0.9 to 0.95. After that, the sizes increase gradually with RSRL. That means, some samples with discriminative features not invading the genuine neighborhood of the minority classes come back in the self-revised training data set by RSRL. When the number of times of RSRLs are larger than 20, the sizes of the training data set are little changed, being stale at about 1800 samples. The sharing rate of the major classes with score 3 to score 5 becomes 83% for the 20th round training data set, while it is 87% for the original training data set. Accordingly, the interrupted samples of majority classes in the training data set for the classification could be removed, and the discriminative samples could be remained with RSRL.

The performances of the generated networks are evaluated by the precision, recall and F-measure.

For the test data set, the precision, recall and F-measure of thirty networks retrained iteratively by RSRL for each class of the test data set are show in Figs. 6, 7 and 8, respectively. Numeral 0 corresponds to the initial network generated by alexNet-based transfer learning using the original training data set. Numeral 1 indicates the first retrained network, and so on. There are not samples of score 9 in the test data set, because only 2 samples were collated in the xiheAA data set. Also, there are only 2 samples of score 8 in the test data set. So, the results with regard to the class score 8 and score 9 are not shown in the figures.

From Figs. 6, 7 and 8, it is seen that the precision, recall and F-measure of minority classes of score 2, score 6 and score 7 are 0 for the initial network. It means, there are no samples which are assigned to the minority classes of score 2, score 6 or score7 for the initial network, although there are the samples labelled to these classes in the test data set. However, for these minority classes, the precision, recall and F-measure of the retrained networks are improved greatly by RSRL. With RSRL, the retrained networks begin to assign the samples to the minority classes. For the score 2, the maximal values of the precision, recall, and F-measure could reach about 0.6, 0.5, and 0.5, respectively; for the score 6, the maximal values of those could reach about 0.35, 0.65, and 0.3, respectively; for the score 7, the maximal values of those could reach about 0.3, 0.15, and 0.2, respectively.

For the majority class score 4 which occupies about the 1/2 of the dada set, the precision, recall, and F-measure could maintain the levels of 0.45, 0.8, and 0.6 with RSRL, although there are a little fluctuation in the initial steps of RSRL for the recall and F-measure. For the majority class score 3 which occupies about the 1/6 of the training dada set, the recall and F-measure are declined from the first retrained network compared with the initial network, although the corresponding precision is increased. However, the tendency changes from the fourth retrained network. The precision, recall and F-measure of all retrained networks become larger than the initial network from that. The F-measure reaches maximal at the point of 13th


**Fig. 11** Examples assigned to each score class

retrained network, while the values of them are 0.42, 0.38, and 0.38, respectively. For the majority class score 5 which also occupies about 1/6 of the training data set, the changing tendencies are similar with the score 3, although the recall and F-measure are still higher for the initial network. The F-measure reaches maximal at the point of 29th retrained network, while the values of them are 0.55, 0.14, and 0.22, respectively.

Figure 9 shows the change of total F-measure regarding all of the classes. It is observed that the total F-measure is optimal for the 12th retrained network. Accordingly, this retrained network could be considered the optimal network for the photos' aesthetic score prediction. The total F-measure is 1.8 for this optimal network, while the total F-measure is 1.02 for the initial network. It is increased about 0.8 compared with the initial network.

For the majority classes of score 3, score 4 and score 5, the average precision, average recall and average F-measure of the optimal network are 0.52, 0.39 and 0.36, respectively, while those of the initial network are 0.38, 0.40, and 0.34, respectively.

For the minority classes of score 2, score 6 and score7, those of this optimal network are 0.3, 0.2 and 0.24, respectively, while there are no samples assigned to the minority classes by the initial network.

Figure 10 shows the deep feature maps of the end-2 layer 'fc' of the initial network (a), the first retrained network (b), and the optimal retrained network (c). One feature map corresponds to one score class. The training data are labelled in the range of score 2 to score 9, so, there are eight feature maps for the end-2 layer 'fc' of the network. The deep feature maps of the initial network for all scores are absurd. It is obvious that no aesthetic features are appeared. For the first retrained network, some deep features for the aesthetic assessment seem emerged. However, for the optimal retrained network, it is observed the deep features for the aesthetic assessment are emerged clearly. For each score, there are obviously different features used as the aesthetic assessment among the

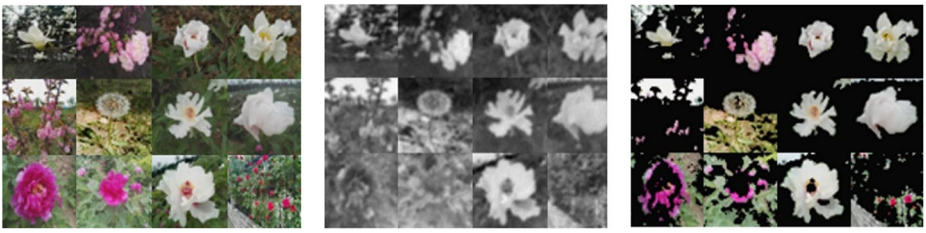

**Fig. 12** Different maps and highlight elements of some photos labeled with score 2

feature map. Accordingly, it is verified that the network generated by the alexNet-based transfer learning can be adapted to the aesthetic assessment task gradually by RSRL, and its performance of the imbalanced score prediction can be improved.

In order to demonstrate the effectiveness of RSRL in imbalanced learning, the two convolutional layers' CNN (2convCNN) are also designed and trained by the xiheAA training data set. The network's layers is as the below.

*Layers = [ 1 'imageinput': Image, Input227x227x3 images*

*2 'conv_1': Convolution, 1 1x1x3 convolutions with stride [8 8]*

*3 'relu_1': ReLU*

*4 'conv_2': Convolution, 25 3x3x1 convolutions with stride [1 1]*

*5 'relu_2': ReLU*

*6 'fc_1': Fully Connected layer, 3 nodes*

*7 'fc_2': Fully Connected layer, 8 nodes, each corresponding to one score*

*8 'softmax': Softmax, 8 nodes*

*9 'classoutput': prediction Output with '2' and 7 other scores ]*

The total F- measures of the initial network and the retrained optimal network are shown in Table 1, besides the previous results of the alexNet-based (alexBase) networks.

From Table 1, it is observed that either for the alexBase or for the 2convCNN, the total F-measure of the optimal retrained network is increased with RSRL. It gained 0.48 from 0.74 to 1.22 for the 2convCNN, while it does 0.79 from 1.02 to 1.81 for the alexBase. Accordingly, it could be said that the RSRL really improves the performance of the network's imbalance learning compared with the non-RSRL network, although the CNN-based deeper network seems to obtain the more improvement. It should be explored more in the future.

On the other hand, on the Matlab environment using the Dell PC with i7-9750H CPU and 16.0 GB RAM, the run times of alexBase and 2convCNN are the 0.0022 s, and 0.0016 s, respectively. The sizes of alexBase and 2convCNN and 1264 KB, and 202 KB, respectively. The computational time complexity of 2convCNN is better than alexBase, although the performance of F-measure is not good compared with the alexBase. With improving the effectiveness of the prediction network, the next work should be consider the trade-off between the effectiveness and efficiency.

Some examples and their predicted scores are shown in Fig. 11. These examples are downloaded from the 500px photo web site, which are not included in the xiheAA data set.

It is obviously that the visual aesthetic quality of these examples is coincided with the level of the predicted scores, and seems to meet the common techniques for composing a good photo. So, the subjective visual assessment verified the reliableness of the score prediction results.

Accordingly, it can be said that RSRL focusing on solving the issue of the imbalance learning improves the performances of the imbalanced score prediction for the phots' aesthetic auto-assessment. The experimental results verify that the issue of imbalanced learning could be improved by RSRL. However, with RSRL, the severe overfitting is occurred. It is necessary to research more in the future.

**Fig. 13** Different maps and highlight elements of some photos labeled with score 4

### 5.3 Aesthetic analysis of highlight elements

In section 4, extracting aesthetic highlight elements from the photo image by using two repetitively trained networks was proposed. In this section, we focus on analyzing the correlation of the extracted elements with the aesthetic assessment, so as to illustrate how to improve the photo's aesthetic quality.

Figure 12 shows the highlight elements of some photos labeled with score 2 in the xiheAA data set. The left is the set of some original images; the middle is the set of the corresponding different maps calculated by the optimal network $aestheticNet_{12}$ and $aestheticNet_{11}$; the right is the set of the extracted highlight regions based on the different maps using the method of k-means.

It is obviously that the extracted parts are messed and cluttered. The salient objects look ugly. Of course, the photo's highlight elements with such features make the aesthetic assessment bad.

Fig. 13 shows the highlight elements of some photos labeled with score 4 in the xiheAA data set.

It is observed that the extracted parts look plain and dull. There are not the salient objects in the extracted region. So, it is seen that the photos without salient objects often obtain the fair assessment.

Figure 14 shows the highlight elements of some photos labeled with score 7 in the xiheAA data set.

It is observed that the extracted parts are very clear. The salient objects are distinctive and made outstanding, and look pretty. So, it is seen that the distinctive object with the clear highlight region make the photo have the high aesthetic assessment.

Accordingly, it could be said that the photos' aesthetics highlight elements extracted by using the repetitively trained aesthetic assessment network reveal the photo's aesthetic qualities. By analyzing the compositions of the extracted elements with the aesthetic scores assigned to the photos, it is possible to learn how to arrange the elements in the photos to make up good photos.

**Fig. 14** Different maps and highlight elements of some photos labeled with score 7

## 6 Conclusion

In this paper, for training the photos' aesthetics prediction network, we proposed a scheme of CNN-based RSRL to solve the issue of imbalanced learning. The pre-trained network are retrained repetitively by the self-revised training data set. Self-revision is done by dropping out the low likelihood photo samples scored in the middle levels from the training data set based on the previously trained network. Experimental results verified that the proposed method is effective of the imbalanced score prediction for the photos' aesthetic auto-assessment, and could be expected to extract the photos' highlight elements related with the aesthetic assessment by the repetitively retrained networks.

Moreover, we think that the proposed method is also available for other domains which are relevant to the imbalanced learning concerned with the subjective auto-assessment.

## References


1. Colin G, McCormack JJ, Santos I, Romero J (2019) Understanding aesthetics and fitness measures in evolutionary art systems. Hidawi Complexity 3495962:1–14. https://doi.org/10.1155/2019/3495962
2. Dong Z, Tian X (November 2015) Multi-level photo quality assessment with multi-view features. Neurocomputing 168(30):308–319
3. He H, Garcia EA (2009) Learning from imbalanced data. TKDE 21(9):1263–1284
4. Huang C, et al. (2016) Learning deep representation for imbalanced classification", 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), https://doi.org/10.1109/CVPR.2016.580, USA
5. Kao Y et al (September 2016) Hierarchical aesthetic quality assessment using deep convolutional neural networks. Signal Process Image Commun 47:500–510
6. Lemarchand F (September 2018) Fundamental visual features for aesthetic classification of photographs across datasets. Pattern Recognition Letters 112(1):9–17
7. Mavridaki E, Mezaris V (2015) A comprehensive aesthetic quality assessment method for natural images using basic rules of photography, 2015 IEEE international conference on image processing (ICIP) , Canada
8. Murray N, et al. (2012) AVA: a large-scale database for aesthetic visual analysis. Proceedings of the IEEE conference on computer vision and pattern recognition, Pages 2408–2415
9. Tan Y, Tang P, Zhou Y, Luo W, Kang Y, Li G (March 2017) Photograph aesthetical evaluation and classification with deep convolutional neural networks. Neurocomputing 228(8):165–175
10. Wang W et al (September 2016) A multi-scene deep learning model for image aesthetic evaluation. Signal Process Image Commun 47:511–518
11. Yubin, Chen Change Loy, and Xiaoou Tang, "Image aesthetic assessment: an experimental survey", arXiv: 1610.00838v2
12. Zhang C et al (September 2018) Visual aesthetic understanding: sample-specific aesthetic classification and deep activation map visualization. Signal Process Image Commun 67:12–21