IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Measuring photography aesthetics with deep CNNs

ISSN 1751-9659
Received on 9th October 2019
Revised 13th February 2020
Accepted on 2nd March 2020
E-First on 14th May 2020
doi: 10.1049/iet-ipr.2019.1300
www.ietdl.org

Gajjala Viswanatha Reddy[1], Snehasis Mukherjee[1] ✉, Mainak Thakur[1]

[1]Computer Vision Research Group, Indian Institute of Information Technology SriCity, Chittoor, India

✉ *E-mail: snehasismukho@gmail.com*

Abstract: In spite of the recent advancements of deep learning based techniques, automatic photo aesthetic assessment still remains a challenging computer vision task. Existing approaches used to focus on providing a single aesthetic score or category ("good" or "bad") of photograph, rather than quantifying "goodness" or "badness". The existing algorithms often ignore the importance of different attributes contributing to the artistic quality of the photograph. To obtain the human-interpretability of aesthetic score of photo, we advocate learning the aesthetic attributes alongwith the prediction of the general aesthetic score. We propose a multi-task deep CNN, that collectively learns aesthetic attributes alongwith a general aesthetic score for the photograph. To understand the mathematical representation of the attributes in the proposed model, a visualization technique is proposed using back propagation of gradients. These visualization of attributes correspond to the location of objects in the images in order to find out which part of an image "triggers" the classification outcome, thus providing the insights about the model's understanding of these attributes. This paper proposes an aesthetic feature vector based on the relative foreground position of the object in the image. The proposed aesthetic features outperform the state-of-art methods especially for Rule of Thirds attribute.

1 Introduction

Photography aesthetics deal with the nature of art, beauty, and taste. The aesthetic quality of the photograph is deeply connected with the creation or appreciation of art. More broadly, Riedel [1] defines aesthetics as 'critical reflection on art, culture and nature'. Judging beauty and other aesthetic qualities in photographs is a highly subjective task as illustrated in [2]. In addition to the subjective nature of the process of aesthetic quality assessment, some aspects related to the assessment process can be articulated through the standard photography practices and some visual design rules of photography such as the rule of thirds, the golden ratio, balancing elements, symmetry, depth of field etc. With the ever-increasing trend of uploading photographs in social media, automatic aesthetic assessment of the uploaded photographs has become a topic of interest to the research community due to its usefulness in a wide range of applications such as building a personal photo-assistant, photo manager, photo quality enhancement, evaluating sharing media, image retrieval, and many more [3, 4].

Conventional approaches for automatic aesthetic quality assessment of photos have been either modelled as a two-class classification problem (aesthetically bad or good photograph) [2, 3, 5], or treated as a regression problem (a single aesthetic score for the photograph) [6–8].

The goal of this study is to provide a quality score for a photograph so that the score correlates with human perception and artistic evaluations. Researchers tried to measure photographic quality in several ways [2–4, 6, 9–13]. Efforts have been made to find out attributes that are associated with image aesthetic quality and describe them with a mathematical model that can be computationally analysed to extract features. The way authors select the image attributes is especially supported by the user (i.e. expert) intuition and photography information, such as colourfulness [2, 13–15], rule of thirds [2, 9, 13], simplicity [14, 15], and so on. Some researchers adopted generic image features, which are originally designed for recognition (e.g. scale-invariant feature transform [16] and Fisher vector [17, 18]). These generic features are observed to outperform the methods based on rule-based features [19]. Recently, more complex models based on deep convolutional neural network (DCNN) are being used for photography aesthetic assessment [3, 7, 10–12, 20]. In a typical DCNN approach, weights are initialised by training on classification datasets (e.g. ImageNet [21]), and then fine-tuned on annotated data for perceptual quality assessment tasks, due to the unavailability of datasets with a huge number of photographs required for training a convolutional neural network (CNN).

Although the state-of-the-art approaches can provide near-human performance in two-class classification of photographs ('good' or 'bad'), they fail to provide critical insights or rationalisation in support of such classifications. For instance, if a photograph gets a poor rating by a computational method, it is difficult to get an insight into the aesthetic attributes (e.g. rule of thirds, uninteresting (or) dull colours etc.), which led to the poor rating.

This study aims to propose a method to measure a score for a photo with respect to the individual attributes (e.g. colour harmony, rule of thirds, symmetry, depth of field etc.) along with a general aesthetic score. We use a multi-task DCNN to simultaneously learn the aesthetic attributes along with a general aesthetic score. We use the aesthetics and attribute database (AADB) [22] for training and testing our model, as this is the only available dataset where different attributes are provided along with the photographs and the ground truth aesthetic scores. The aesthetic attributes considered in this dataset are as follows: balancing elements, colour harmony, content, depth of field, light, motion blur, object emphasis, repetition, rule of thirds, symmetry, vivid colour, and general aesthetic score. Fig. 1 shows examples of some photographs taken from the AADB dataset where each column of Figs. 1*a* and *b* shows sample photographs related to an attribute. The top rows of each column of Figs. 1*a* and *b* represent photographs with high aesthetic ground truth-value according to the specific attribute specified in the column. The bottom rows of Figs. 1*a* and *b* show photos with low aesthetic scores. The proposed approach for assessment of the aesthetic quality of photographs has three major contributions:

• We apply a multi-task DCNN architecture based on EfficientNet [23] architecture to learn the aesthetic score for different attributes

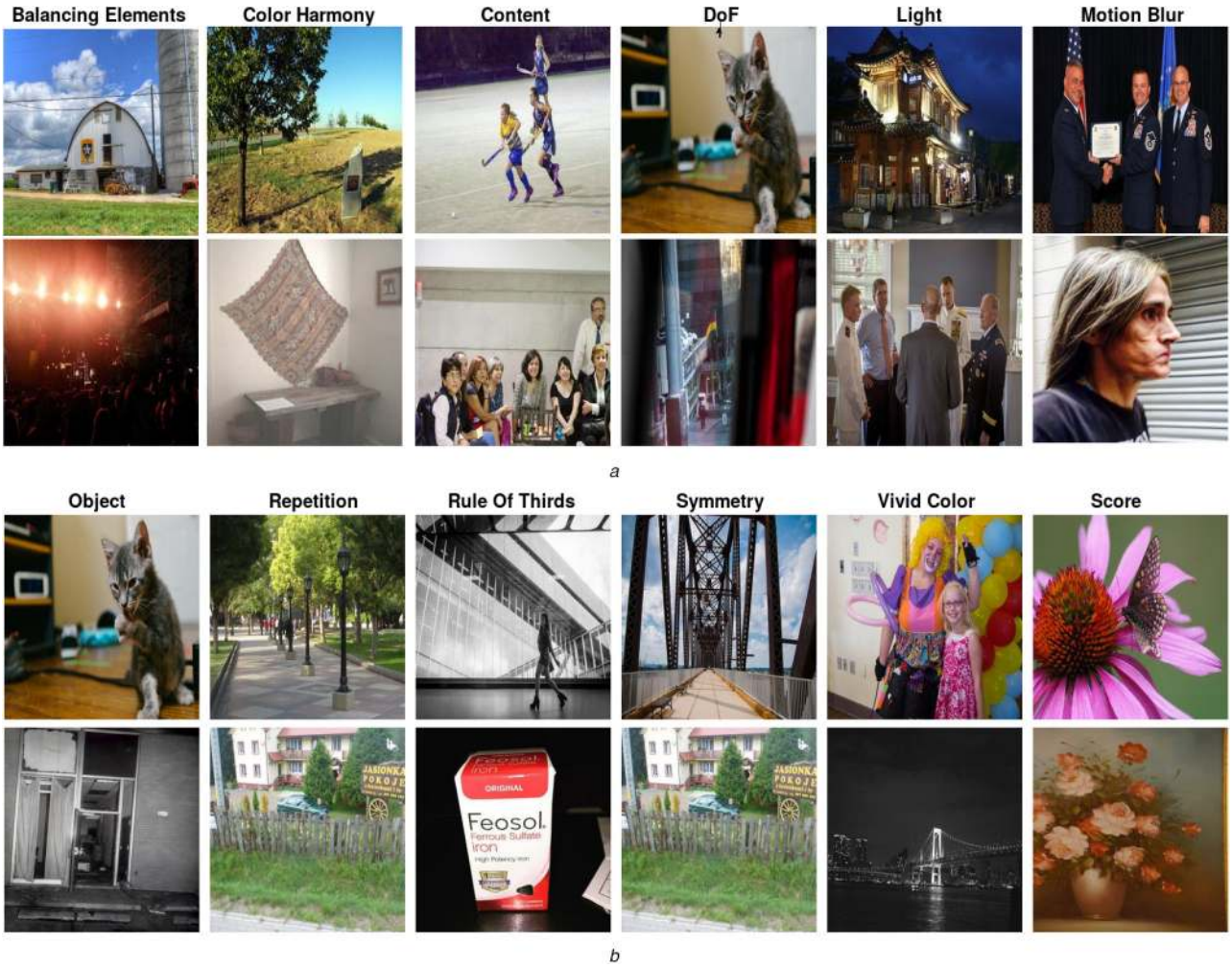

Fig. 1 *Sample images taken from the AADB dataset*
(a, b) Each column shows sample photographs related to an attribute, mentioned over the corresponding columns. Top rows of each column represent photographs with high aesthetic ground truth-value according to the specific attribute specified in the respective column. Bottom rows show photos with low aesthetic scores for the same attributes

for a photo. This study is the first attempt to apply EfficientNet architecture for aesthetic quality estimation of photographs.
• We propose a novel loss function that suits the proposed architecture for learning the aesthetic scores for different attributes.
• To further improve the score for the rule of thirds attribute, we propose a handcrafted feature vector based on the relative foreground position of the object in the photograph. This study is the first attempt to measure the aesthetic quality of photographs especially based on the rule of thirds attribute.

Next, we provide a survey of literature in the area of aesthetic photo quality assessment.

2 Literature survey

Photo aesthetic quality assessment using computational techniques has been a subject of interest to researchers during the last few decades [24]. Most of the earlier methods used to combine low-level image features such as gradient or Laplacian, hue, dark channel prior (DCP) etc., to design high-level features related to specific aesthetic attributes, and trained aesthetic classifier over the high level features [2, 8, 25]. Based on the standard rules to measure the quality of photography and visual design, Datta *et al.* [2] proposed 56 visual features to encapsulate low-level image features to measure aesthetic attributes of the photo. In [25], aesthetic attributes were divided into three categories by Dhar *et al.* as follows:

(1) Compositional attributes (e.g. salient objects, low depth of field, rule of thirds, and opposing colours).
(2) Content attributes (e.g. faces, portraits, presence of animals, and 15 scene types).
(3) Sky-illumination attributes (e.g. clear sky, cloudy sky, and sunset sky).

Classifiers were trained for each of these attributes separately from low-level features (e.g. haar features, spatial pyramid of shape features, colour histograms, and centre surrounding wavelets). Outputs of these classifiers are used as input features for another classifier for measurement of aesthetic quality. Lahrache *et al.* [26] relied on some basic image features along with image layout and colour combination of the image. The combination of features is trained using a series of classifiers.

Sun *et al.* [27] emphasised on image complexity features such as composition and distribution of colour and shapes for measuring aesthetics quality. Verma *et al.* [28] proposed a multi-layer perceptron model to combine low-level image features such as colour hue, DCP, Laplacian, and illumination and provide a score between 1 and 10 determining the aesthetic quality of the photograph. To capture the inherent subjectivity of the photo quality measure, Wu *et al.* [8] proposed a model to predict the distribution of the measures of the aesthetic quality of photos, instead of a single aesthetic measure, where the distribution of aesthetic measure is learned by a support vector regressor. In [13], the photo quality is measured by the prominence of the object of interest in the photograph, where the prominence is measured from image contrast around the object boundary.

After the introduction of deep learning-based techniques, automatic extraction of relevant features became more convenient and boosted the efficiency of any classification techniques of any domain. Subsequently, deep learning has shown promising success

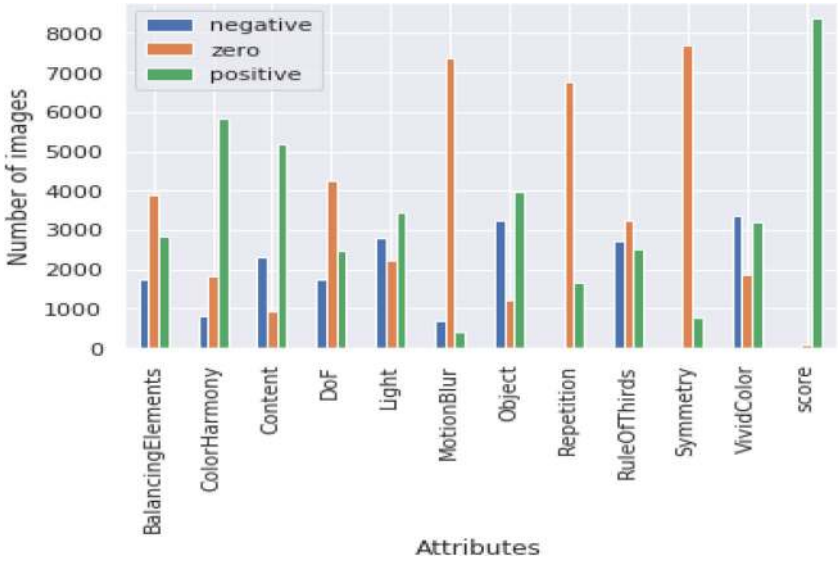

Fig. 2 *The distribution of the number of training photographs (along the y-axis) of the AADB dataset, rated (annotated) with respect to the different aesthetic attributes (mentioned along the x-axis)*

in predicting the technical quality of images. Deep learning techniques have shown much better performance compared to the traditional machine learning-based approaches for assessment of the aesthetic quality of photos [3, 7, 20, 22, 29]. Lu *et al.* [3] decomposed the photograph into a bag of orderless patches, and then apply a DCNN to extract features from the patches and aggregate them to find the feature vector for classification of the photograph as 'good' or 'bad'. Kao *et al.* [7] applied a DCNN to extract the aesthetic features from the photograph, and then apply a regression model to predict a continuous aesthetic score.

All the datasets available in the literature, for predicting the aesthetic score of photographs, are unbalanced across classes (e.g. from 'bad' to 'good' photographs), which is a big concern for CNN-based approaches as the learning model may tend to be biased towards or against some particular classes when trained on such an unbalanced dataset. For instance, Fig. 2 shows the distribution of a number of photographs across different categories with respect to different aesthetic attributes for the AADB dataset [22] used in this study. Clearly, the AADB dataset is highly biased with respect to most of the aesthetic attributes. To overcome the problem of the lack of balance of dataset across classes, Jin *et al.* [20] proposed a weighted CNN to extract the aesthetic features from the photograph, followed by a regression model to train the aesthetic score.

Most of the state-of-the-art techniques for assessment of the aesthetic quality of photographs either classify the photographs as 'good' or 'bad' photographs or provide an overall score for the photograph based on its aesthetic quality. A few approaches try to provide a range of aesthetic scores of the photograph, instead of providing a single score. However, only a few attempts have been made to predict the aesthetic quality of a photograph with respect to the individual aesthetic attributes. Kong *et al.* [22] made the first attempt in this direction, by proposing a DCNN for aesthetic score prediction, by learning the scores with respect to individual aesthetic attributes. They proposed a Siamese network to unify the photo content and aesthetic attributes to provide the aesthetic scores of the photo. Malu *et al.* [29] proposed a multi-task DCNN model based on ResNet 50 architecture [30] to jointly learn nine aesthetic attributes (out of the 12 attributes mentioned in Section 1) from the photographs. Motivated by Malu *et al*. [29], we propose a novel multi-task DCNN architecture for jointly learning the aesthetic attributes. The proposed architecture has lesser parameters compared to [29] and provide a better score (i.e. closer to human interpretation) for the attributes. Moreover, unlike [29], the proposed architecture can dynamically learn the loss weights, enabling better adaptability of the network for learning different aesthetic attributes. Furthermore, observing the lack of significant efforts in the literature to emphasise on the score for the rule of thirds attribute of photography (except Bhattacharya *et al.* [9]), we propose a feature vector to provide a better score of photos for the rule of thirds attribute compared to the state-of-the-art.

In multi-task learning, shared network features are often affected by all the associated attributes. Out of the 12 aesthetic attributes specified in the AADB dataset, we omit symmetry, repetition, and motion blur attributes for training the model as most of the photographs in the AADB dataset were rated neutral (zero) for the attributes (Fig. 2). We model the proposed DCNN for the remaining eight attributes along with the general aesthetic score as a regression problem. The following section describes the proposed architecture.

3 Proposed DCNN architecture

The proposed DCNN architecture for extracting the image features for predicting the aesthetic quality is influenced by the EfficientNet architecture [23]. Tan and Le [23] proposed a model scaling method that uses a simple yet highly effective compound coefficient to scale up CNNs in a structured manner. Unlike conventional approaches that arbitrarily scale network dimensions, such as width, depth, and resolution, the EfficientNet uniformly scales each dimension with a fixed set of scaling coefficient. EfficientNet scales up the baseline network EfficientNet-B0 using compound scaling to obtain a family of models, called EfficientNets (B0–B7) [23].

We experimented with the family of EfficientNet models from B0 to B7. We observed that due to the inadequacy of training samples (8500) in the dataset for training a deep neural network, the models with a higher level of scaling (B5, B6, and B7) are affected by overfitting problem. The amount of overfitting, in this study, is measured by the difference between training and test accuracies. On the other hand, model B4 can capture the in-depth aesthetic features from the photograph, compared to the simpler models (B0, B1, B2, and B3). Hence, we use the EfficientNet-B4 network to train the aesthetic attributes along with the overall aesthetic score altogether, in the same way as in [22, 29]. The EfficientNet-B4 network computes a feature hierarchy layer by layer, and with sub-sampling layers, the feature hierarchy has an inherent multi-scale, pyramidal shape. This in-network feature hierarchy produces feature maps of different spatial resolutions. In the EfficientNet-B4 network, early-stage feature maps are larger with low-level features that describe spatial details, while late-stage feature maps are smaller with high-level features that are more discriminative. In general, localisation is sensitive to low-level features while high-level features are crucial for classification.

A number of recent approaches have improved detection and segmentation by using different (or) multiple layers in CNN architectures [31–33]. In our CNN, we reuse the multi-scale feature maps from different layers computed in the forward pass and thus

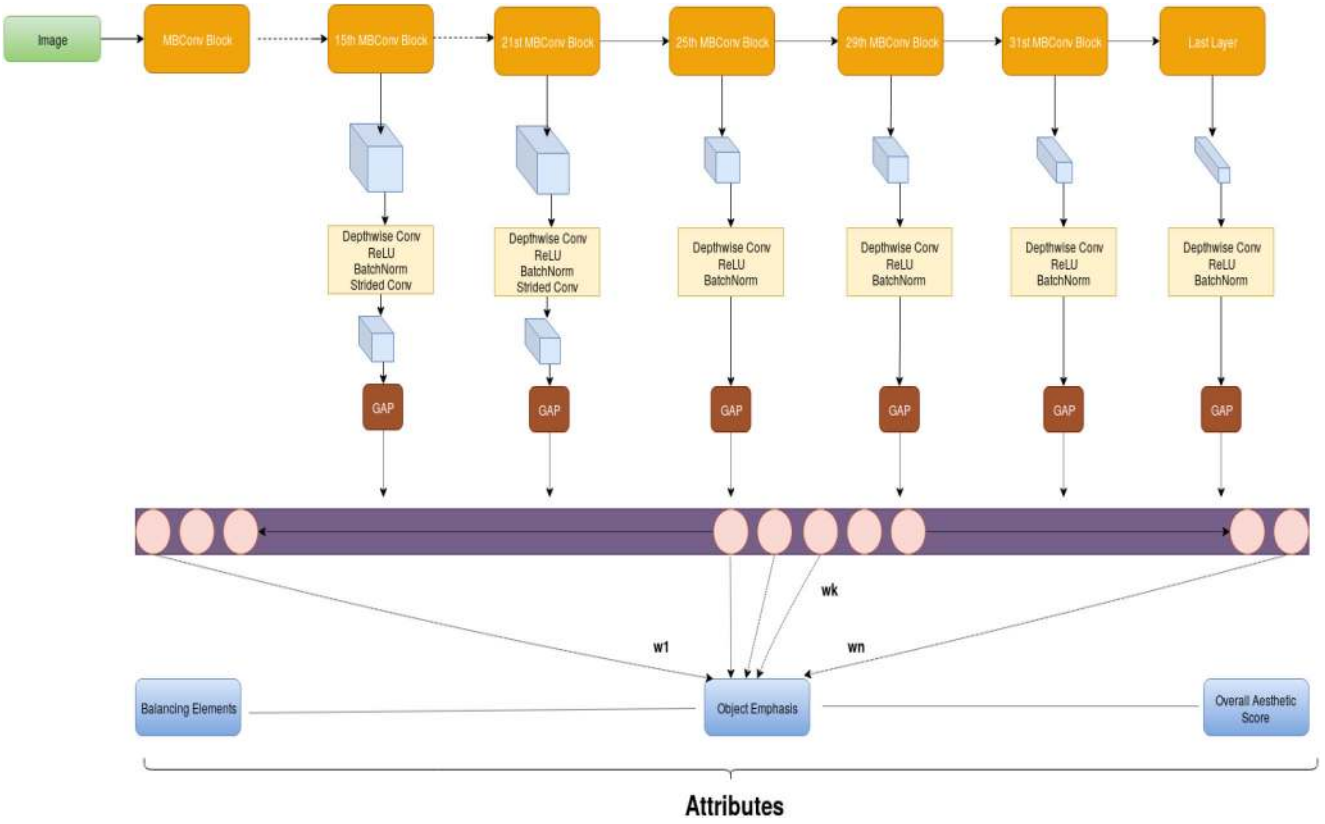

Fig. 3 *A diagram of the proposed CNN architecture of the proposed model. As depicted in the diagram, the proposed CNN learns the respective attributes separately, followed by a weighted sum of the different components of the CNN*

come free of cost (expense of parameters). The high-resolution maps have low-level features having less representational capacity. We mix up the low-resolution, semantically strong features with the high-resolution, semantically weak features instead of using just the final low-resolution layer.

The EfficientNet-B4 architecture is divided into 31 successive mobile inverted bottleneck convolution blocks (MBConvBlock) [23]. Each MBConvBlock comprises a convolutional layer, batch normalisation layer, Swish activation function, depthwise convolution, batch normalisation, Swish activation function, squeeze excitation block followed by a convolutional layer and batch normalisation.

Feature maps are extracted from the output of the MBConvBlocks number 15, 21, 25, 29, and 31 (i.e. the C1, C2, C3, C4, and C5 convolutional blocks, respectively), along with the final low-resolution layer (C6). A depthwise convolution is performed on feature maps C1, C2, C3, C4, C5, and C6 followed by Relu and batch normalisation layers. We use feature dimensions as 256, 256, 256, 256, 512, and 1024 for the six convolutional layers, respectively. More channels (or feature dimensions) are used for extracting high-level features because of the discriminative nature of the features. Instead of down-sampling, strided convolution is performed on C1 and C2 feature maps to maintain the same resolution. Features are pooled from each of these six feature maps with a global average pooling (GAP) layer. The GAP layer provides the average of the rectified convolution maps in the spatial domain. The pooled features are concatenated and used as an input to a fully connected layer. The dropout p is kept as 0.4. The proposed model architecture is shown in Fig. 3.

Malu *et al.* [29] used the following loss function for extracting aesthetic features using the ResNet-based architecture by summing up all attribute losses:

$$\mathscr{L}(f(x), y) = \sum_{i=1}^{n} w_i \times L_i(f_i(x), y_i), \tag{1}$$

where n corresponds to the total number of attributes, w_i corresponds to the weight given for the ith attribute, y_i corresponds to the ground truth value for the ith attribute, $f_i(x)$ is the predicted value for the ith attribute, x represents input image and $L_i(y_i, f_i(x))$ corresponds to the mean squared error for the ith attribute.

In the above loss function, weights are pre-defined or fixed for each attribute. Instead of using the fixed loss weights for all attributes, a dynamic weighting scheme is employed in the proposed loss function, which automatically and dynamically learns the loss weights as follows:

$$\mathscr{L}(f(x), y) = w_1 \times [(y_1 - \hat{y}_1)^2] + \cdots + w_n \times [(y_n - \hat{y}_n)^2], \tag{2}$$

where $w_1, w_2, \ldots, w_n$ are the weights for the features corresponding to the aesthetic attributes and $\hat{y}_i$ corresponds to the predicted value for the ith attribute. Weights are learnable parameters. Here the weights may not add up to one. So, we have used the weighted mean. The modified loss function is as follows:

$$\mathscr{L}(f(x), y) = \frac{\sum_{i=1}^{n} w_i \times [(y_i - \hat{y}_i)^2]}{\sum_{i=1}^{n} w_i}, \tag{3}$$

where n is the total number of attributes. The equations for calculating the derivatives with respect to w_j(jth attribute) are shown as

$$\frac{\partial \mathscr{L}(f(x), y)}{\partial w_j} = \frac{\sum_{i=1}^{n} w_i \times ((y_j - \hat{y}_j)^2)] - \sum_{i=1}^{n} w_i \times [(y_i - \hat{y}_i)^2]]}{(\sum_{i=1}^{n} w_i)^2}, \tag{4}$$

where $j \in \{1, 2, \ldots, n\}$.

The weight w_j is updated as

$$w_j = w_j - \alpha \times \frac{\partial \mathscr{L}(f(x), y)}{\partial w_j}, \tag{5}$$

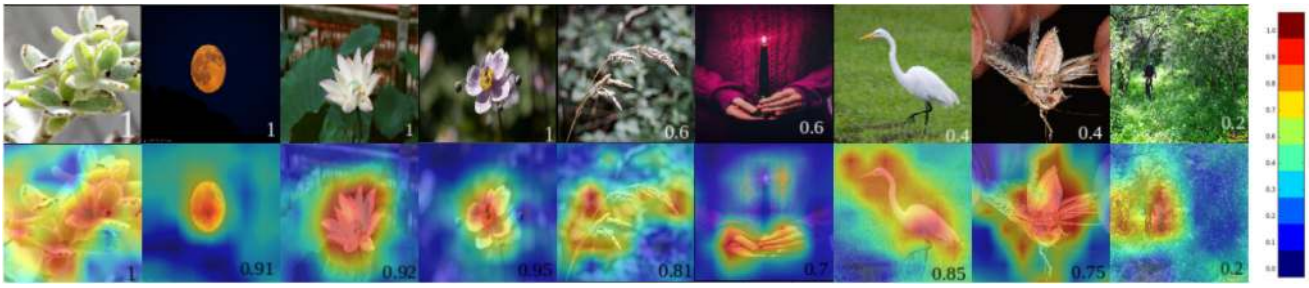

Fig. 4 *Activation maps corresponding to the object emphasis attribute for a few sample test photographs of the AADB dataset. The first row shows the original images (ground truth scores for the same attribute are marked at the bottom right corner of the respective images). The second row shows the activation maps for the corresponding images given by our model (the predicted scores by the proposed method are marked at the bottom right corners of each image). Colour-bar at the right indicates the colour encoding of activation map*

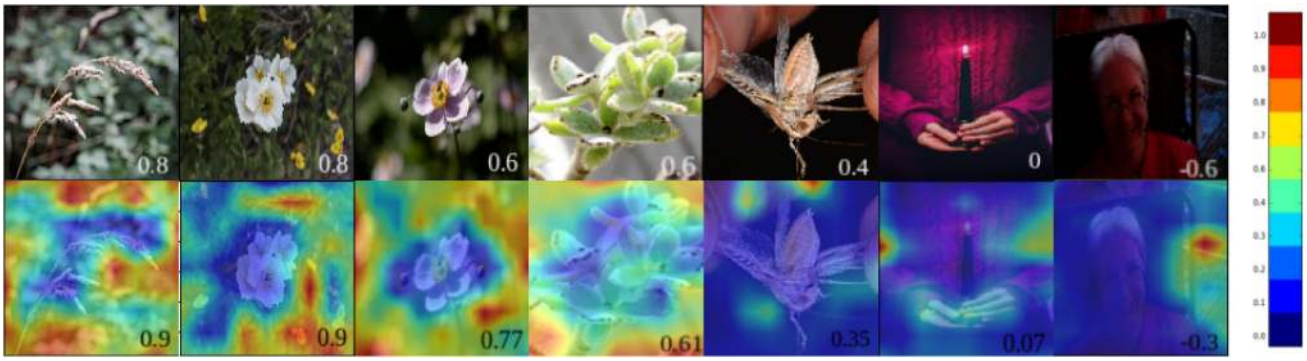

Fig. 5 *Activation maps corresponding to the depth of field attribute for a few sample test photographs of the AADB dataset. The first row shows the original images (ground truth scores for the same attribute are marked at the bottom right corner of the respective images). The second row shows the activation maps for the corresponding images given by our model (the predicted scores by the proposed method are marked at the bottom right corners of each image). Colour-bar at the right indicates the colour encoding of activation map*

where α denotes the learning rate. Next, we illustrate the process of visualising the performance of the proposed method for extracting the most challenging aesthetic features.

4 Aesthetic features and extraction

Unlike [22], we generate class activation maps for visualising the performance of the proposed approach and a better understanding of the attributes, using [34], which is illustrated next.

4.1 Visualisation

We generate class activation maps using [34], corresponding to four important aesthetic attributes to highlight the most important regions in the image. The four aesthetic attributes are object emphasis, depth of field, use of light, and content. The attribute activation maps corresponding to an attribute depicts how the proposed model can identify and emphasise the most important regions in the photograph. In this section, we show some sample photographs taken from the test samples, and their corresponding activation maps with respect to the four attributes. We call the activation map corresponding to an attribute as 'gaze' of the model. Note that, in the AADB dataset, different images are given and annotated according to different aesthetic attributes. It is difficult to get the same set of images annotated according to the different aesthetic attributes. Hence, in our study, we have shown the visualisation of the aesthetic features for the images which are annotated according to the particular aesthetic attribute.

4.1.1 Object emphasis: The activation maps corresponding to the attribute object emphasis for some sample test photographs of the AADB dataset are shown in Fig. 4. Fig. 4 depicts that the proposed model can learn the object emphasis attribute, as we can observe a high concentration of gaze around the object regions of the images.

4.1.2 Depth of field: We represent the shallow depth of field measures for a few sample test photographs from the AADB dataset, as provided by the proposed model. Fig. 5 shows the sample images and corresponding shallow depth of field measure given by our model. We observe in Fig. 5 that the proposed model looks for blurry regions (represented by lighter/reddish colour) outside the object of interest (represented by blueish colour) while making the judgement based on the depth of field attribute. We can further observe that the activation maps are more active towards the corner of the images, supporting the rules of photography.

4.1.3 Use of light: The assessment of photograph based on the use of light while capturing the photo is challenging as the use of light attribute depends not only on the quantity of light in the photo but also an assessment on how the light enhances the whole composition. As depicted by Fig. 6, in most of the cases the proposed model tends to give higher emphasis (more reddish) on brighter parts of the photograph. In columns 2–4 of Fig. 6, the proposed model is trying to look at the source of the light in the photograph. We can also conclude that the proposed model is excluding darker regions in the photograph while making predictions about score related to the attribute use of light.

4.1.4 Content: The aesthetic quality of content in a photograph is significantly subjective and is dependent on the context of the photo. In photography, the context of a photo is described by the objects of interest and its relative position with respect to the other objects in the photograph. Hence, a model is expected to devote more activation around the objects of interest of the photo, for better performance in an assessment of the content attribute. The proposed model shows satisfactory performance on this aspect, as depicted in Fig. 7, especially in the first and second columns. Furthermore, we can observe from Fig. 7 that the proposed approach is performing well at identifying the content. As shown in Fig. 7, activation maps provided by the proposed model are most active at the contents of the image.

In Fig. 5, the activation map in column 1 shows that the proposed model is looking for blurry regions in the image and column 3 of Fig. 7 model activation maps are most active at the object (content). In Fig. 5, the activation map shown in column 6 shows that the proposed model is looking for blurring regions and

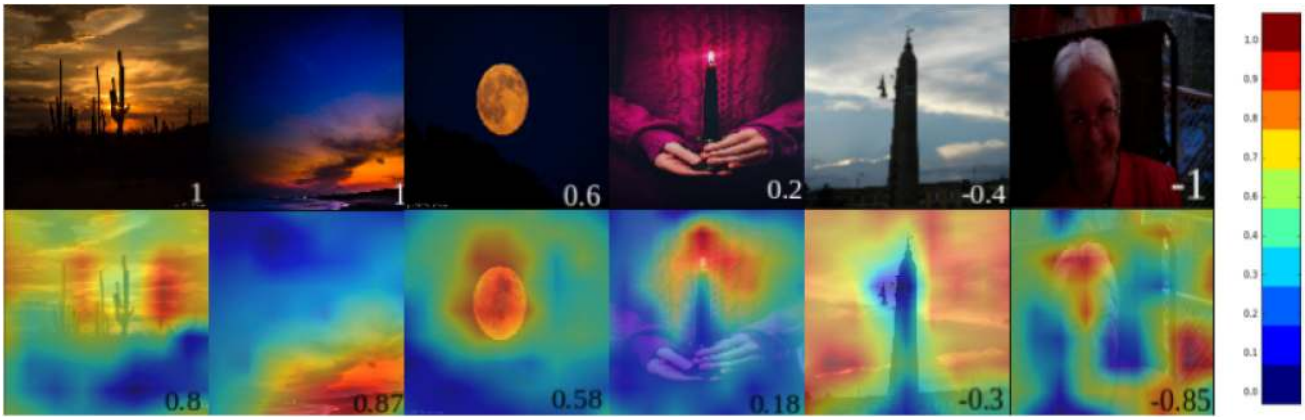

Fig. 6 *Activation maps corresponding to the use of light attribute for a few sample test photographs of the AADB dataset. The first row shows the original images (ground truth scores for the same attribute are marked at the bottom right corner of the respective images). The second row shows the activation maps for the corresponding images given by our model (the predicted scores by the proposed method are marked at the bottom right corners of each image). Colour-bar at the right indicates the colour encoding of activation map*

Fig. 7 *Activation maps corresponding to the photograph content attribute for a few sample test photographs of the AADB dataset. The first row shows the original images (ground truth scores for the same attribute are marked at the bottom right corner of the respective images). The second row shows the activation maps for the corresponding images given by our model (the predicted scores by the proposed method are marked at the bottom right corners of each image). Colour-bar at the right indicates the colour encoding of activation map*

in Fig. 6 column 4 the model activation map is maximally active at the source of the light in the image. From the above illustrations, we can conclude that even though the proposed model is based on multi-task learning, the model has learned task specific feature representations. The object's emphasis and content attributes are closely related to each other. Both attributes try to look at the object of interest in the image. The activation maps and predicted scores (0.91 and 0.92) of Fig. 4 column 3 and Fig. 7 column 6 are almost similar. Similarly, the activation maps of Fig. 4 column 8 and Fig. 7 column 7 are alike and the predicted scores (0.75 and 0.7) are almost similar. This clearly shows that our model has captured the relation between these two attributes.

We perform further experiments with the proposed model to observe the activation maps provided by the proposed model when applied on over- and under-exposed images taken from the MIT-Adobe FiveK dataset [35]. Motivated by the studies conducted in [36] on images with wrong colour constancy and in [37] on under-exposed images, we obtain the activation maps from a few under-exposed and the corresponding over-exposed images for three different attributes: objectness, depth of field, and content and show them in Fig. 8. From Fig. 8, we can observe that the proposed model can provide a similar activation map for both under- and over-exposed images with the same visual content, with respect to all the three attributes.

4.2 Rule of thirds

The rule of thirds is a ‘rule of thumb’ or guideline which applies to the process of composing visual images such as designs, films, paintings, and photographs [38]. The rule of thirds is the most well-known rule of photographic composition. The basic principle behind the rule of thirds is to break an image down into thirds (both horizontally and vertically) so that we have nine parts. With this grid, the ‘rule of thirds’ now identifies four important parts of the image that one should consider for placing points of interest in as we frame an image. Apart from this, it also gives us four ‘lines’ that are also useful positions for elements in a photo.

The theory behind the rule of thirds attribute is that if we place points of interest in the intersections or along the lines that the photo becomes more balanced and will enable a viewer of the image to interact with them more naturally. Studies have shown that while viewing photographs human attention is usually concentrated at one of the intersection points most naturally rather than the centre of the shot. The ‘rule of thirds’ can help us to create well balanced and interesting shots.

To formulate photographic quality assessment based on the aesthetic attributes, in the context of a machine learning problem, we need to associate the users' notions of aesthetics to well defined, attribute-specific features from an image. To this end, we extract a relative foreground position feature for images with single-foreground compositions. This feature is based on elementary rules of photographic composition. Details about the aesthetic feature are discussed next.

4.2.1 Aesthetic features: Relative foreground position in a photograph is defined as the normalised Euclidean distance between the foreground's centre of mass, also called the visual attention centre, to each of the four symmetric stress points or points of interest in the image frame as shown in Fig. 9. In photographic literature [39], the stress points are the strongest focal points in a photographic frame (indicated by sky-blue dots in Fig. 9). To attract the viewer's attention to a foreground object, the photographer is often advised to adjust the frame in such a way that the visual attention centre coincides with one of the four stress points. The clause, ‘one of the stress points’, is of particular interest in this context. If the visual attention centre is positioned

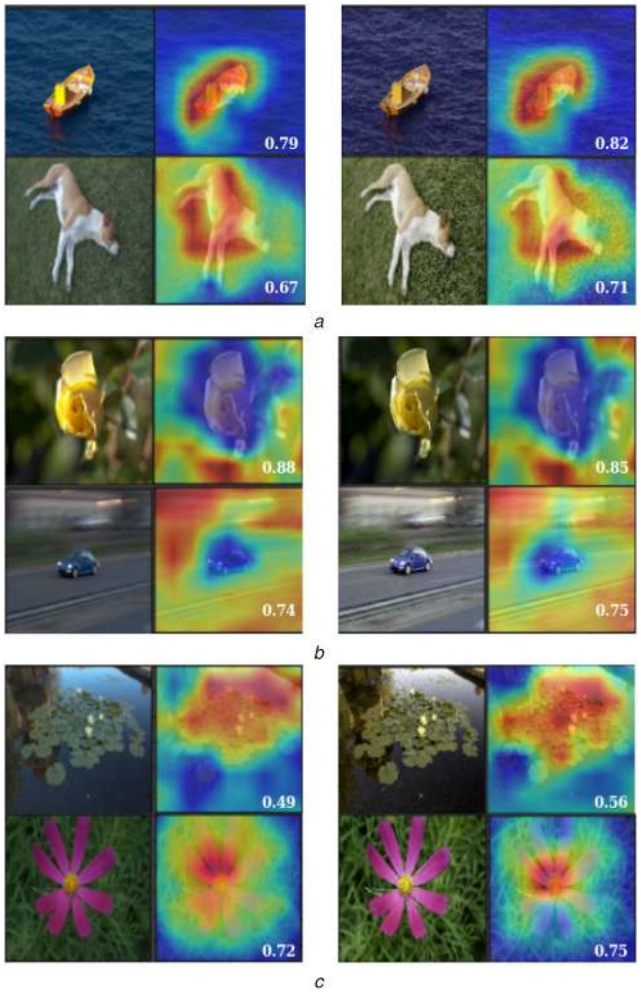

Fig. 8 *Activation maps for the proposed model when applied on under-exposed (left side) and over-exposed (right side) images, corresponding to the*
(a) Objectness, ***(b)*** Depth of field, ***(c)*** Photograph content attributes for a few sample test photographs of the MIT-Adobe FiveK dataset [35]

Fig. 9 *A sample image to show the relationship between the visual attention centre and the four stress points (adapted from the rule of thirds attribute) for a photograph. The black lines divide the frame into nine identical parts. Each intersection of the black lines generates a stress point indicated by sky-blue dots. In this image the stress point or point of interest coincides with the foreground object's visual attention centre which depicts a high aesthetic score for the image according to the rule of thirds attribute*

equidistant from all the stress points during capturing the photo, the viewers' attention gets equally divided across these four points. This causes the viewer to lose interest in the photograph, thereby reducing its artistic value. This observation is also confirmed by the study where participants for ranking the photos tend to rank the photos with foreground aligned near a stress point higher than those with foreground centred in the frame [9].

Thus, every photograph containing a single subject composition can be uniquely characterised by a four-dimensional feature vector ($\mathcal{F}$) as follows:

$$\mathcal{F} = 1/(h \times w)[\ \| x_0 - s_1 \|\ ,\ \| x_0 - s_2 \|\ ,\ \| x_0 - s_3 \|\ ,\ \| x_0 - s_4 \|\ ], \quad (6)$$

where h and w are the height and width of the image, x_0 is the visual attention centre and s_i are the stress points starting from top-left, in clockwise direction. Fig. 9 shows a single subject composition from our dataset with its respective visual attention centre and stress point locations.

4.2.2 Extracting visual attention center: The next task is to extract the visual attention centre for a given photo. To find the visual centre, we need to know the relative position of the object in the given image. We use YOLOv3 to find the object in a given photo [40]. Fig. 9 shows the generated object proposals from YOLO. If there are multiple objects in a photo, then the object with high probability value provided by the YOLO is chosen for extracting the visual centre. The aesthetic feature vector is generated for the images in the dataset. A linear support vector machine is trained on these feature vectors. The model trained using the proposed handcrafted aesthetic feature obtained a value of 0.455 (Spearman's correlation (ρ)) on the test dataset for the rule of thirds attribute where the state-of-art accuracy is 0.225. Next, we illustrate some more experiments and results.

5 Experiments and results

We first discuss the experimental set-up made for the experiments carried out on this study, followed by the results.

5.1 Experimental set up

We use the aesthetic and attributes database (AADB) [22] for experimenting on the proposed approach. The AADB contains a balanced distribution of professional and consumer photos, with a total of 10,000 images. Eleven aesthetic attributes and annotator's IDs are provided with the dataset. A standard partition with 8500 images for training, 500 images for validation, and 1000 images for testing are proposed in the dataset [22]. The distribution of the number of photographs annotated with respect to all the attributes in the training data of the AADB dataset is shown in Fig. 2. All the aesthetic attributes except the repetition and symmetry attributes in the AADB dataset are normalised to $[-1, 1]$. Repetition and symmetry attributes are normalised to [0, 1], as negative values are not considered for these two attributes. The overall aesthetic score is normalised to [0, 1].

Owing to the unavailability of training images, we used a pre-trained EfficientNet-B4. The model was trained on the ImageNet classification dataset with about 1.2 million images of 1000 classes. The input image size is set to 299×299. Following [23], the red (R)–green (G)–blue (B) values of the images are normalised to standard normal variates with means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225] for the R, G, and B intensity values. For example, if r is the intensity value, m_r is the mean and s_r is the standard deviation, then the normalised intensity value $s = (r - m_r)/s_r$. We applied data augmentation on the input training images with techniques such as horizontal flip. The last residual block of the proposed network provides convolution maps of size 9×9. We reduce the sizes of the feature maps from each layer to the next layer maintaining the size. We fix the batch size to 16 and train the model for 20 epochs in our experiments.

A learning rate of 1×10^{-6} is applied for the pre-trained layers. A learning rate of 0.001 is used for the newly added layers with a weight decay of 0.01. The learning rate of 0.00001 is applied for the loss function learnable parameters. A lower learning rate is used for the loss function's learnable parameters because higher values are leading to unstable training. Adam optimiser is used to update the network parameters.

Table 1 Spearman's rank correlations measure obtained by the proposed CNN model compared to the state-of-the-art, for all the aesthetic attributes separately

Attribute	Kong *et al*. [22]	Malu *et al*. [29]	Malu *et al*. [29] using dynamic weighting scheme	EfficientNet fine tune	Our method
balancing elements	0.220	0.186	0.205	0.1682	**0.3314**
content	0.508	0.584	0.590	0.3814	**0.5985**
colour harmony	0.471	0.475	0.490	0.2799	**0.5165**
depth of field	0.479	0.495	0.553	0.3858	**0.677**
light	0.443	0.399	0.453	0.2647	**0.5146**
object emphasis	0.602	0.666	0.660	0.4931	**0.6772**
rule of thirds	0.225	0.178	0.221	0.1807	**0.2733**
vivid colours	0.648	0.681	0.685	0.4867	**0.7057**
aesthetic score	0.678	0.689	0.693	0.4699	**0.7059**

Bold values indicate the highest (best) measurement obtained among the competing methods with respect to the attribute specified in the row.

5.2 Results and discussions

To evaluate the scores for photographs with respect to the aesthetic attributes provided by our model, we report the Spearman's ranking parametric statistic (ρ) between the computable score of the individual aesthetic attribute and the corresponding ground truth score for the test data. The ranking parametric statistic (ρ) evaluates the score based on the monotonic relationship between computable scores and ground truth scores. Hence explicit calibration between the ground truth and computed scores is not needed. The parametric statistic lies within the range of $[-1, 1]$, with larger values similar to higher correlation and vice-versa. Table 1 shows the performance of the proposed CNN model on the AADB dataset using the proposed loss function. We compare the performance of the proposed model with Kong *et al.* [22] and Malu *et al.* [29] to establish the efficacy of the proposed method. The values presented in Table 1 depict the performances of all the competing methods according to our own experimental set up.

Although from Table 1 we observe that Malu *et al*. [29] provides slightly better result compared to EfficientNet [23] baseline model, still we have worked on [23] because of the following two reasons. First, EfficientNet has fewer parameters compared to ResNet, which is the backbone of [29]. Second, the EfficientNet is computationally much cheaper than other networks, and hence, is easily deployable in a photographic camera.

To show the effectiveness of our proposed dynamic weighting scheme, we compare the proposed model with the dynamic weighting scheme proposed in [29], along with [22], as these two are the only methods in the literature, aiming for predicting aesthetic scores for different attributes separately. From the results shown in Table 1 we can infer that the EfficientNet trained using the proposed dynamic weighting scheme performed better compared to the model trained with fixed loss weights in the loss function. Moreover, we apply the proposed loss function on the state-of-the-art method [29] and still the proposed method shows better performance as depicted in Table 1. The proposed method outperformed the models in [22, 29] for almost all attributes expect for object emphasis which is marginally less (0.660) compared to model in [29] (0.666). From Table 1 we can conclude that there is a huge improvement in terms of the correlation values for low performing attributes such as light (from 0.399 to 0.453), rule of thirds (from 0.178 to 0.221), and depth of field (from 0.495 to 0.553). Marginal improvements are observed in the correlation values for better performing attributes such as vivid colours (from 0.681 to 0.685) and overall aesthetic score (from 0.689 to 0.693). This shows that the proposed weighting scheme concentrates on predicting the more complicated attributes.

Apart from the proposed dynamic weighted scheme, we also tried with fine-tuning the EfficientNet-B4 model pre-trained on the ImageNet dataset, for more experimentation. The fine-tuning is performed by modifying the last fully connected layer of the pre-trained EfficienNet-B4 and training it on the AADB training dataset for aesthetic attribute prediction. Table 1 shows the performance of the two different training approaches on AADB. The results in Table 1 shows the effectiveness of the EfficienNet-B4 architecture with the proposed dynamic weighting scheme.

The proposed CNN model obtains less correlation value (ρ) for the rule of thirds and balancing elements attributes, compared to the other attributes. Most of the images have a rating in the range of $[-0.4, 0.4]$ for these two attributes. The distributions of the rule of thirds and balancing elements attributes are shown in Figs. 10 and 11, respectively. From Figs. 10 and 11, we can infer that the scores in the AADB dataset are highly imbalanced for the rule of thirds and balancing elements attributes. Only nine images are available in the dataset with the highest rating based on the rule of thirds attribute. Only ten images are there in the dataset with the lowest rating (-0.8) based on the rule of thirds attribute. Owing to less number of positive and negative samples, it is difficult for a CNN model to extract useful features that might lead to high Spearman's correlation.

The proposed CNN-based method manages to outperform the state-of-the-art approaches in [22, 29] for all the aesthetic attributes. However, for balancing elements and rule of third attributes the score provided by the proposed CNN model is low. In fact, these two attributes are location sensitive. Balancing elements attribute deals with the relative positioning of the objects of interest, with each other and the frame. The rule of thirds attribute deals with the positioning of the salient elements in the given frame. However, the proposed handcrafted feature for the rule of thirds attribute outperforms all the previous methods with a huge difference in terms of the correlation values. The correlation values for the rule of thirds attribute, obtained by the proposed handcrafted feature are tabulated in Table 2.

The judgements of photographs based on the aesthetic attributes are very subjective in nature. To quantify this perspicacity, in the AADB dataset the ground-truth score is obtained by calculating the mean score of ratings given by completely different human beings. To quantify the efficacy of the computational models between ratings given by the model and the ground truth, ρ between each individual's ratings and the ground-truth scores for each attribute were calculated [22]. The average values of ρ are reported in Table 3. From Table 3, it clearly indicates that the different human raters annotate the images consistently, and when labelling more images, raters contribute more stable rankings of the aesthetic scores. From Table 3, we can see that the proposed model outperforms the human performances consistently (as measured by ρ) averaged across all raters. However, while considering the individual raters who have annotated more than 200 images, human evaluator's consistently surpass the proposed model's performance in terms of rank correlation (ρ).

6 Conclusion

This study presents a multi-task DCNN trained with a dynamic weighting scheme to automatically and dynamically learn the loss weights for learning aesthetic attributes of photographs. Results show that the scores of six aesthetic attributes (content, colour harmony, depth of field, light, object emphasis, vivid colour)

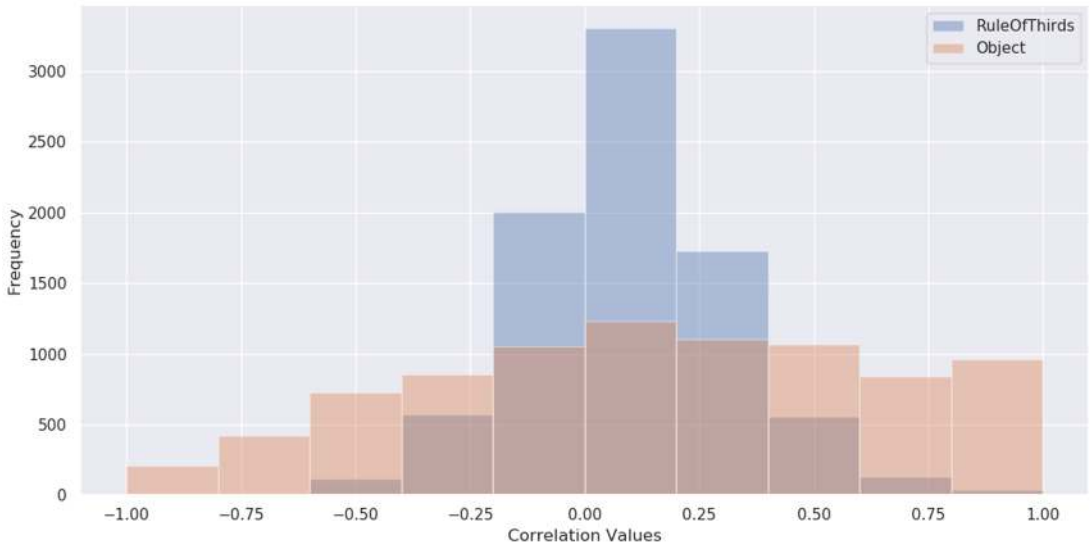

Fig. 10 *Histogram showing the distribution of the number of images with respect to the different correlation values according to the rule of thirds and object emphasis attributes for training set*

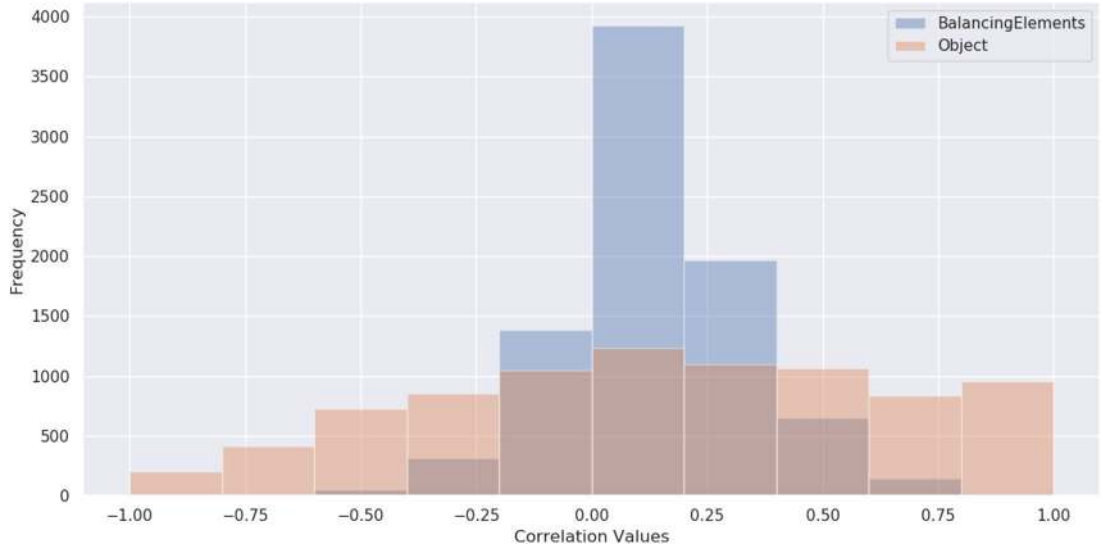

Fig. 11 *Histogram showing the distribution of the number of images with respect to the different correlation values according to the balancing elements and object emphasis attributes for training set*

Table 2 Spearman's rank correlations for the rule of thirds attribute

Attribute	Kong *et al.* [22]	Malu *et al.* [29]	Our method
rule of thirds	0.225	0.178	0.455

Table 3 Human performance on AADB

Number of images rated	Number of raters	Correlation value (ρ)
> 0	195	0.6738
> 100	65	0.7013
> 200	42	0.7112
our approach	—	0.7059

estimated by the proposed approach correlate considerably with their individual ground truth scores. Whereas in the case of attributes such as the balancing elements and the rule of thirds, the correlation is less. The activation maps corresponding to the learned individual aesthetic attributes show that the proposed model can capture the inherent representation of the aesthetic attributes suitable to highlight the attributes automatically. Furthermore, we have proposed a handcrafted aesthetic feature vector based on the relative foreground position of the object in the image to perform well on the rule of thirds attribute. The obtained correlation values show that the proposed handcrafted feature has learned the inherent representation well for the rule of thirds attribute for a given photograph. In the future, the proposed DCNN model, with the support of an adequate number of photographs, may be extended to work for the better prediction of the rule of thirds and balancing elements attributes.

7 Acknowledgment

The authors wish to thank NVIDIA for providing a Titan × GPU as a research grant, which is used in this research for experimentations.

8 References

[1] Riedel, T.: 'Review of encyclopedia of aesthetics 4 vols. Michael Kelly', *Art Documentation: J. Art Libr. Soc. North Am.*, 1999, **18**, (2), pp. 48, doi:10.1086/adx.18.2.27949030

[2] Datta, R., Joshi, D., Li, J., *et al.*: 'Studying aesthetics in photographic images using a computational approach'. Proc. European Conf. on Computer Vision, Graz, Austria, 2006

[3] Lu, X., Lin, Z., Shen, X., *et al.*: 'Deep multi-patch aggregation network for image style, aesthetics, and quality estimation'. Proc. IEEE Int. Conf. on Computer Vision, Santiago, Chile, 2015, pp. 990–998

[4] Marchesotti, L., Murray, N., Perronnin, F.: 'Discovering beautiful attributes for aesthetic image analysis', *Int. J. Comput. Vis.*, 2014, **113**, pp. 1–21

[5] Lu, X., Lin, Z., Jin, H., *et al.*: 'Rapid: rating pictorial aesthetics using deep learning'. Proc. 22nd ACM Int. Conf. on Multimedia, Orlando, FL, USA., 2014, pp. 457–466

[6] Kao, Y., Huang, K., Maybank, S.: 'Hierarchical aesthetic quality assessment using deep convolutional neural networks', *Signal Process., Image Commun.*, 2016, **47**, pp. 500–510

[7] Kao, Y., Wang, C., Huang, K.: 'Visual aesthetic quality assessment with a regression model'. 2015 IEEE Int. Conf. on Image Processing (ICIP), Quebec City, Canada, 2015, pp. 1583–1587

[8] Wu, O., Hu, W., Gao, J.: 'Learning to predict the perceived visual quality of photos'. IEEE Int. Conf. on Computer Vision (ICCV), Barcelona, Spain, 2011, pp. 225–232

[9] Bhattacharya, S., Sukthankar, R., Shah, M.: 'A framework for photo-quality assessment and enhancement based on visual aesthetics'. Proc. 18th ACM Int. Conf. on Multimedia, Firenze, Italy, 2010

[10] Xue, W., Zhang, L., Mou, X.: 'Learning without human scores for blind image quality assessment'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Portland, OR, USA., 2013, pp. 995–1002

[11] Kang, L., Ye, P., Li, Y., *et al.*: 'Convolutional neural networks for no-reference image quality assessment'. Proc. IEEE Conf. on Computer Vision and Pattern Recognition, Columbus, OH, USA., 2014, pp. 1733–1740

[12] Bosse, S., Maniry, D., Wiegand, T., *et al.*: 'A deep neural network for image quality assessment'. 2016 IEEE Int. Conf. on Image Processing (ICIP), Phoenix, AZ, USA., 2016, pp. 3773–3777

[13] Luo, Y., Tang, X.: 'Photo and video quality evaluation: focusing on the subject'. European Conf. on Computer Vision (ECCV), Marseille, France, 2008, pp. 386–399

[14] Lo, K.Y., Liu, K.H., Chen, C.S.: 'Assessment of photo aesthetics with efficiency'. Int. Conf. on Pattern Recognition (ICPR), Tsukuba, Japan, 2012, pp. 2186–2189

[15] Ke, Y., Tang, X., Jing, F.: 'The design of high-level features for photo quality assessment'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Graz, Austria, 2006, pp. 419–426

[16] Wong, L.K., Low, K.L.: 'Saliency-enhanced image aesthetics class prediction'. Int. Conf. on Pattern Recognition (ICPR), ICIP 2009, Cairo, Egypt, 2009, pp. 997–1000

[17] Caffe, Y.J.: 'An open source convolutional architecture for fast feature embedding'. Available at http://cafe.berkeleyvision.org/, 2013

[18] Liu, W., Tao, D.: 'Multiview Hessian regularization for image annotation', *IEEE Trans. Image Process.*, 2013, **22**, (7), pp. 2676–2687

[19] Le Callet, P., Viard-Gaudin, C., Barba, D.: 'A convolutional neural network approach for objective video quality assessment', *IEEE Trans. Neural Netw.*, 2006, **17**, (5), pp. 1316–1327

[20] Jin, B., Segovia, M.V.O., Susstrunk, S.: 'Image aesthetic predictors based on weighted CNNs'. 2016 IEEE Int. Conf. on Image Processing (ICIP), Phoenix, AZ, USA., 2016, pp. 2291–2295

[21] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'ImageNet classification with deep convolutional neural networks'. Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA., 2012, pp. 1097–1105. 1, 3, 5, 6

[22] Kong, S., Shen, X., Lin, Z., *et al.*: 'Photo aesthetics ranking network with attributes and content adaptation'. Proc. European Conf. on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016, pp. 662–679

[23] Tan, M., Le, Q.V.: 'EfficientNet: rethinking model scaling for convolutional neural networks'. Int. Conf. on Machine Learning (ICML), Long Beach, CA, USA., 2019, pp. 1–10

[24] Deng, Y., Loy, C.C., Tang, X.: 'Image aesthetic assessment: an experimental survey', *IEEE Signal Process. Mag.*, 2017, **34**, (4), pp. 80–106

[25] Dhar, S., Ordonez, V., Berg, T.L.: 'High level describable attributes for predicting aesthetics and interestingness'. 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA., 2011, pp. 1657–1664

[26] Lahrache, S., El-Ouazzani, R., El-Qadi, A.: 'Rules of photography for image memorability analysis', *IET Image Process.*, 2018, **12**, (7), pp. 1228–1236

[27] Sun, L., Yamasaki, T., Aizawa, K.: 'Photo aesthetic quality estimation using visual complexity features', *Multimedia Tools Appl.*, 2018, **77**, (5), pp. 5189–5213

[28] Verma, A., Koukuntla, K., Varma, R., *et al.*: 'Automatic assessment of artistic quality of photos', arXiv:1804.06124, 2018

[29] Malu, G., Bapi, R.S., Indurkhy, B.: 'Learning photography aesthetics with deep CNNs'. Proc. Midwest Artificial Intelligence and Cognitive Science Conf. (MAICS), Fort Wayne, IN, USA., 2017

[30] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA., 2016, pp. 770–778

[31] Liu, W., Anguelov, D., Erhan, D., *et al.*: 'SSD: single shot multibox detector'. European Conf. on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016

[32] Lin, T., Goyal, P., Girshick, R., *et al.*: 'Focal loss for dense object detection'. IEEE Int. Conf. on Computer Vision IEEE Int. Conf. on Computer Vision (ICCV), Venice, Italy, 2017

[33] Lin, T.-Y., Dollar, P., Girshick, R., *et al.*: 'Feature pyramid networks for object detection'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Hawaii, USA., 2017, pp. 2117–2125

[34] Selvaraju, R.R., Das, A., Vedantam, R., *et al.*: "Grad-cam: why did you say that? Visual explanations from deep networks via gradient-based localization', arXiv preprint arXiv:1610.02391, 2016

[35] Bychkovsky, V., Paris, S., Chan, E., *et al.*: 'Learning photographic global tonal adjustment with a database of input/Output image pairs'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA., 2011

[36] Afifi, M., Price, B., Cohen, S., *et al.*: 'When color constancy goes wrong: correcting improperly white-balanced images'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA., 2019, pp. 1535–1544

[37] Wang, R., Zhang, Q., Fu, C.-W., *et al.*: 'Underexposed photo enhancement using deep illumination estimation'. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA., 2019, pp. 6849–6857

[38] Meech, S.: '*Contemporary quilts: design, surface and stitch*' (Sterling Publishing, Batsford, UK., 2007), ISBN 0-7134-8987-1

[39] Jonas, P.: '*Photographic composition simplified*' (Amphoto Publishers, Prentice Hall, USA., 1976)

[40] Redmon, J., Farhadi, A.: YOLOv3: an incremental improvement, Available at https://arxiv.org/abs/1804.02767, 2018