Study on relationship between composition and prediction of photo aesthetics using CNN
Daichi Sakaguchi¹, Hironori Takimoto²* and Akihiro Kanagawa²

Abstract: The purpose of image aesthetics assessment is to automatically predict the perceived quality of an image. Convolutional neural network (CNN) based on deep learning has been used for aesthetics assessment and has displayed potential results. Our final objective is to identify features that contribute to the estimation of aesthetic quality by using explainable AI. In this study, we focused on composition as the first step. By applying clustering to the attention maps obtained by CNN and Grad-CAM++, it was experimentally verified whether the CNN model considers composition for aesthetics assessment. In addition, we verified whether the aesthetic quality features that humans pay attention to differ according to photographic categories such as landscape or portrait.

Keywords: Prediction of photo aesthetics; Convolutional neural network; Explainable AI

1. Introduction
The need to capture attractive photographs is increasing because of the more numerous opportunities to share photographs with the spread of social networking services. To capture an attractive photograph, the main subject matter in the photograph should be defined clearly (Datta et al., 2006; Dhar et al., 2011; Ke et al., 2006; Luo et al., 2011; Marchesotti et al., 2011; Nishiyama et al., 2011). Composition is one of the most important elements for clarifying the subject of a photograph. The main subject of a photograph becomes clearer when the elements and spaces that constitute the photograph are arranged in a more balanced manner by considering composition. For a subject or background that is photographed, the impression would vary with variation in the arrangement. Therefore, it is feasible to enhance the attractiveness of a photograph by considering its composition.

Meanwhile, automated techniques for evaluating the aesthetic quality of photographs have been proposed in the field of computer vision. The automatic aesthetic assessment of photographs has attracted attention owing to its wide range of potential applications. Aesthetic assessment is a quantification of whether an individual likes a photograph sensibly. In the early research into the aesthetic quality assessment of photographs, several hand-designed feature-based methods were proposed to realize image aesthetic assessment (Luo et al., 2011). In these methods, visual features that contribute to the human perception of aesthetics (such as colorfulness, contrast, composition, lighting, and subject) are modeled manually as low-level features. However, it is difficult for hand-designed feature-based methods to precisely estimate it because only a few visual elements contribute to human perception of aesthetics that are explainable as explicit knowledge.

During the past years, convolutional neural network (CNN) models based on deep learning (LeCun et al., 2015) have become the most influential architecture for addressing visual recognition. Hence, many studies based on deep learning for aesthetic assessment of photographs have
been proposed (Ma et al., 2017; Kong et al., 2016; Lu et al., 2015; Mai et al., 2016; Omori et al., 2019; Takimoto et al., 2021; Talebi & Milanfar, 2018). However, these CNN models are generally referred to as black boxes because of the low understanding of their internal workings. There are no research studies that explain how CNN models perform aesthetic estimation of photographs.

Recently, eXplainable AI (XAI) has been attracting attention for visually explaining predictions made by CNN models to make these more transparent and explainable. In particular, several techniques have been proposed to estimate the input regions that affect the output of CNN by analyzing the process from input to output in the reverse direction (Chattopadhay et al., 2018; Selvaraju et al., 2017; Zhou et al., 2016). The features and areas that contribute to the estimation of aesthetic quality are likely to be revealed through the application of these XAI techniques of aesthetic quality assessment. The identification of the factors that contribute to the aesthetic quality of photographs would facilitate their enhanced evaluation. Furthermore, it is expected to discover not only formalized elements but also tacit knowledge that has not yet been revealed.

Our final objective is to identify important features that contribute to aesthetic quality estimation using XAI. In this study, we focus on compositional information as the first step. We analyze whether composition is considered in photo aesthetic assessment based on CNN by using an XAI. First, we construct a CNN that achieves aesthetics estimation with high accuracy. Next, the image regions that contribute to aesthetic estimation are obtained as attention maps by using Grad-CAM++ (Chattopadhay et al., 2018) as a technique to visualize the learned concepts of the CNN. Clustering is performed on the attention map obtained by Grad-CAM++ using x-means. Then, representative regions of interest are obtained from the clustering results to examine whether a composition is considered in aesthetic quality estimation. In addition, we consider whether the aesthetic quality features that humans pay attention to differ according to photograph categories such as landscape and portrait.

2. Composition
Composition refers to the manner in which the various elements in a scene are arranged within the frame. There are a number of established composition guidelines that can be applied in almost all scenarios to enhance the impact of a scene. These guidelines would aid you in capturing more compelling photographs by lending these a natural balance, drawing attention to the important parts of the scene, or leading the viewer’s eye through the image. Shizuno et al. (Shizuno & Hamada, 2014) reported that the typical compositions in photography are rule of thirds, center composition (Hinomaru), diagonal composition, and symmetry composition.
Examples of these typical compositions are shown in Figure 1, and their respective characteristics are described below.

**Rule of thirds**

The rule of thirds is a compositional technique that leverages asymmetry to create dynamic and visually exciting compositions. It involves a uniform division of your frame into nine parts using two horizontal lines and two vertical lines. The concept is to position your focal point in conjunction with one of the lines or one of the four intersections. This naturally places the focal point off-center, thereby creating aesthetically pleasing images through asymmetry. The two-part composition with different numbers of segments is also popular. Here, the main subject is the distinctive dividing line in the photograph, such as the horizon or the horizon line. First, the vertical or horizontal sides of the camera screen are divided uniformly into two. Then, the horizontal or vertical lines are aligned on the dividing line. The image becomes a balanced and stable composition when the photograph is uniformly divided into two parts with a dividing line.

**Central composition**

Central composition is a straightforward and common form of composition. Here, an important subject such as a prominent object or vanishing point is placed in the center of the photograph. The central position of an image is highly important and draws the viewer’s attention. Central composition works because it allows for negative space in all the corners of the frame. Essentially, the background surrounds the subject on all the sides, thereby functioning as a natural frame.

**Diagonal compositions**

In diagonal composition, the elements in the image are organized based on a diagonal line. This composition can emphasize perspective, impart a sense of depth to the image, and add dynamism. In addition, the tension of the diagonal line (rather than a horizontal or vertical line) makes the composition more engaging.

**Symmetry composition**

In photography, symmetry appears when parts of your composition mirror other parts. It is obtained when two halves of your scene look identical and balance each other. Horizontal symmetry occurs when the line passes through the scene from right to left and divides the scene into equal halves. This creates a horizontal line of symmetry. In addition, vertical symmetry and radial symmetry are highly popular as symmetry compositions.

**3. Realization of estimation of aesthetic value of photographs using CNN and visualization of basis of judgment**

The purpose of this study is to verify whether CNN-based estimation of the aesthetic quality of photographs considers compositional information by using XAI. The procedure of the proposed method is described below:

**Step 1.** Construction of a CNN model for image aesthetic assessment.

**Step 2.** Calculation of an attention map that visualizes the decision basis of the CNN model by using Grad-CAM++.

**Step 3.** Consideration of the relationship between aesthetics and composition based on representative regions of interest obtained by clustering attention maps.
3.1. CNN Model for image aesthetic assessment

First, a CNN is constructed to estimate the aesthetics of individual photographs with high accuracy. In this study, the VGG-19 architecture is used for the aesthetics estimation model. The architecture of Grad CAM++ based on VGG19 with global average pooling (GAP) is shown in Figure 1. VGG is designed based on the fundamental concept that deeper networks are better. VGG has filters that are smaller than those of AlexNet. Here, each filter has a size of \(3 \times 3\) albeit with a lower stride of one, and effectively captures a receptive field identical to that captured by a \(7 \times 7\) filter with four strides (Karen & Zisserman, 2014).

In the VGG model used in the proposed method, image features related to aesthetics estimation are extracted for the input image, mainly in the convolutional layers (Conv1-1–Conv5-4). Then, an output layer (FC3) is incorporated in addition to the two full convolutional layers (FC1 and FC2), to perform a two-class classification that outputs either “high quality” or “low quality” for the input image. Note that a \(224 \times 224\) pix. RGB 3ch color image is used as the input image.

3.2. Calculation of attention map using Grad-CAM++

It is generally crucial that machine learning models are interpretable and explainable. Interpretability and explainability help the developer to debug and improve the model, enhance trust in it, validate model predictions, and gain insights. However, it is difficult for a user to know what criteria are used to make decisions of models based on deep learning because models that can identify and discriminate with high accuracy generally have a highly complex structure. XAI techniques are being developed to increase the explanatory power and interpretability of these deep learning models. Several methods have been proposed to visually explain the basis of CNN decisions by acquiring attention maps of inputted images for image recognition tasks. To better interpret the decision of the CNN model, we need to determine which feature in our inputs contributed the most to that decision. The class activation mapping (CAM; Zhou et al., 2016) method and its variations have been investigated for XAI since 2016. These have been subsequently used for digital healthcare, particularly in the areas of medical image analysis. (Yang et al., 2022)

Selvaraju et al. proposed Grad-CAM (Selvaraju et al., 2017) to visualize the region of interest of CNN. Grad-CAM is a method to pseudo-represent the degree of the output of each region of a convolutional layer to be visualized in a CNN as a heat map. Note that the obtained heat map is called an attention map. When a CNN as a classifier outputs that the input image is in class \(c\), the attention map \(L^c\) is defined as follows:

\[
L^c = \sum_k w_k^c A^k
\]  

(1)

where \(A^k\) is the feature map obtained as the output of the final convolutional layer of the CNN. \(w_k^c\) is the weight from the \(k\) th node to class \(c\) after global average pooling (GAP) is applied to the feature map \(A^k\):

\[
w_k^c = \frac{1}{Z} \sum_i \frac{\partial Y^c}{\partial A^k_i}
\]  

(2)

where \(Y^c\) is the classification score when class \(c\) is determined and \(Z\) is the number of pixels in the feature map.

Chattopadhyay et al. proposed Grad-CAM++ (Chattopadhyay et al., 2018), which is a generalization of Grad-CAM. In Grad-CAM++, \(w_k^c\) is defined as follows by using the value \(\alpha\) for the area of features in the feature map:
\[ w_k^c = \sum_i \sum_j \alpha_{ij}^c \text{ReLU} \left( \frac{\partial Y^c}{\partial A_{ij}^k} \right) \]  

\[ \alpha_{ij}^c = \frac{2 \left( \frac{\partial^2 Y^c}{\partial A_{ij}^k} \right)^2 + \sum_{a} \sum_{b} A_{ab}^k \left( \frac{\partial^2 Y^c}{\partial A_{ab}^k} \right)^2 - \sum_{a} \sum_{b} A_{ab}^k \left( \frac{\partial Y^c}{\partial A_{ab}^k} \right)^2}{\sum_{a} \sum_{b} A_{ab}^k \left( \frac{\partial Y^c}{\partial A_{ab}^k} \right)^2} \]  

\[ \text{Table 1. Structure of CNN model for image aesthetic assessment} \]

| Layer type   | Kernel\Stride | Activation | Output size     |
|--------------|----------------|------------|-----------------|
| Conv1-1      | 3 x 3\1       | ReLU       | 224 x 224 x 64  |
| Conv1-2      | 3 x 3\1       | ReLU       | 224 x 224 x 64  |
| Max Pooling  | 2 x 2\1       | —          | 112 x 112 x 64  |
| Conv2-1      | 3 x 3\1       | ReLU       | 112 x 112 x 128 |
| Conv2-2      | 3 x 3\1       | ReLU       | 112 x 112 x 128 |
| Max Pooling  | 2 x 2\1       | —          | 56 x 56 x 128   |
| Conv3-1      | 3 x 3\1       | ReLU       | 56 x 56 x 256   |
| Conv3-2      | 3 x 3\1       | ReLU       | 56 x 56 x 256   |
| Conv3-3      | 3 x 3\1       | ReLU       | 56 x 56 x 256   |
| Conv3-4      | 3 x 3\1       | ReLU       | 56 x 56 x 256   |
| Max Pooling  | 2 x 2\1       | —          | 28 x 28 x 256   |
| Conv4-1      | 3 x 3\1       | ReLU       | 28 x 28 x 512   |
| Conv4-2      | 3 x 3\1       | ReLU       | 28 x 28 x 512   |
| Conv4-3      | 3 x 3\1       | ReLU       | 28 x 28 x 512   |
| Conv4-4      | 3 x 3\1       | ReLU       | 28 x 28 x 512   |
| Max Pooling  | 2 x 2\1       | —          | 14 x 14 x 512   |
| Conv5-1      | 3 x 3\1       | ReLU       | 14 x 14 x 512   |
| Conv5-2      | 3 x 3\1       | ReLU       | 14 x 14 x 512   |
| Conv5-3      | 3 x 3\1       | ReLU       | 14 x 14 x 512   |
| Conv5-4      | 3 x 3\1       | ReLU       | 14 x 14 x 512   |
| Max Pooling  | 2 x 2\1       | —          | 7 x 7 x 512     |
| Flatten      | —              | —          | 1 x 1 x 25088   |
| FC1          | —              | ReLU       | 1 x 1 x 4096    |
| FC2          | —              | ReLU       | 1 x 1 x 4096    |
| FC3          | —              | Softmax    | 1 x 1 x 2       |

\[ S_{\text{FC3}} = \sum_{ij} \alpha_{ij}^c \text{ReLU} \left( \frac{\partial Y}{\partial A_{ij}} \right) \]

\[ \text{Table 2. Structure of CNN model with Grad-CAM++ for image aesthetic assessment (after Conv5-4 layer)} \]

| Layer type   | Kernel\Stride | Activation | Output size     |
|--------------|----------------|------------|-----------------|
| Conv5-4      | 3 x 3\1       | ReLU       | 14 x 14 x 512   |
| Global Average Pooling | —            | —          | 1 x 1 x 512     |
| FC1          | —              | ReLU       | 1 x 1 x 512     |
| FC2          | —              | ReLU       | 1 x 1 x 512     |
| FC3          | —              | Softmax    | 1 x 1 x 2       |
This study introduces Grad-CAM++ to the CNN model with the aesthetics estimation function constructed in Section 2. In the proposed method, the target convolutional layer for Grad-CAM++ is the Conv5-4 layer in Table 1, and the size of the obtained attention map is $14 \times 14$ pix. For the CNN model with the GAP layer applied to the output of the final convolutional layer (Conv5-4 layer), the structure after the target final convolutional layer (Conv5-4 layer) is shown in Table 2.

### 3.3. Clustering of attention maps

As mentioned in the previous section, the attention map obtained by Grad-CAM++ extracts the regions that CNN considers important for estimating a photograph as being of high quality. In this study, we use Grad-CAM++ to obtain attention maps for images ($n$ images) that are estimated to be of high quality by the CNN model. Let $L_i$ be the attention map for the $i$th image.

First, the attention map of $14 \times 14$ pix. is resized to a size identical to that of the original image ($224 \times 224$ pix.). Next, we apply clustering to the $n$ attention maps $L_i$ ($i = 1, 2, \ldots, n$) obtained by Grad-CAM++. In proposed analysis method, the number of clusterings is very important. It is not subjective and should be determined by clear criteria. In this study, we use the $k$-means method to obtain the number of clusters that can be used as a criterion for analyzing the relationship between the attention map and the composition. The $x$-means method is based on the $k$-means method and is a type of non-hierarchical clustering (Pelleg & Moore, 2000). The $x$-means method automatically determines the optimal number of clusters $k$ by repeating the two-partite $k$-means algorithm based on the information criterion. The Bayesian information criterion (BIC) is used in this study as the information criterion.

The center of gravity of each of the $k$ clusters obtained by the $x$-means method is obtained and visualized. The center of gravity of each cluster can be considered as a representative region that the CNN considers important in estimating the quality of the photograph. Finally, we analyze the relationship between the obtained clusters and composition. There are two ways to determine which compositional template a attention map fits: observation by the human eye and quantitative methods. Because there are certain differences in human observations, it is better to use quantitative methods to determine. However, such an objective method has not yet been established. Therefore, the judgment of composition matching is subjective.

We conducted a subjective evaluation experiment to analyze a composition that matches the attention map. Ten subjects (8 male, 2 female, 22–41 years old, $\bar{x} = 23.5$) participated in the subjective evaluation experiment. For each cluster, the subjects were presented with the original images and heat maps included in that cluster. At the same time, a group of composition templates containing representative auxiliary lines, as shown in Figure 1, were illustrated to the subjects. Subjects were asked which of the composition best matched the highlighted regions of the heatmap included in each cluster. Subjects responded from five options, including four representative compositions. The presented options were: “rule of thirds”, “two-part composition”, “central composition”, “diagonal compositions” and “none of them”. The most frequently selected composition for each cluster was then chosen as the matched composition.

### 4. Experiment

#### 4.1. Experimental setup

We employ the AVA-dataset, which is a large-scale database for aesthetic visual analysis (Murray et al., 2012). This dataset contains approximately 255,000 photographs that have been aesthetically evaluated. Each photograph is scored by at least 200 subjects. The photograph ratings range from 1 to 10, with 10 indicating the highest quality. The number of votes for each score of each photograph is defined as $V = [v_1, v_2, \ldots, v_{10}]$. $v_i$ represents the number of votes at score $i$. The mean score is represented as Score. The Score of each photograph is defined as follows:
Score = \left( \frac{\sum_{i=1}^{10} i \cdot v_i}{\sum_{i=1}^{10} v_i} \right) 

For the categories of the photographs, two of the 66 tags are attached to each photograph. All the images are resized to 224 × 224 pixels. We use two datasets for the experiment. Each dataset consists of images tagged with “landscape” and those tagged with “portrait.” The dataset contains 21,945 landscape images, and the median Score is 5.56. Meanwhile, there are 17,515 portrait images with a median Score of 5.58. Note that each portrait image contains a person or face.

Here, the distribution of Score in the AVA-dataset is non-uniform because the AVA-dataset contains many images with Score near the median. The aesthetics estimation model constructed in this study classifies the input images into high quality and low quality as aesthetics estimation. Therefore, the image is not used in the experiment if the Score of an image is within median ± 0.5. As a result, 1,500 images in order of highest Score and 1,500 images in order of lowest Score are
used as high and low quality images, respectively. Ninety percent of the image data is used for training, and the remaining is used for validation.

The proposed model is trained by Adam in 100 epochs. Then, the weight of the epoch with the lowest validation loss is used in subsequent experiments. The initial learning rate and the batch size are set to 0.000001 and 128, respectively. The ReLU function is used as the activation function except in the output layer, and the Softmax function is used in the output layer. Categorical cross entropy is used as the loss function. The weights of the VGG model are initialized by training on ImageNet (Deng et al., 2009). Experiments are performed on an NVIDIA Titan RTX GPU by using Keras with the Tensorflow backend to implement all the models.

4.2. Experimental results and discussion
First, we show the experimental results for landscape images. We input 1,500 training images with high-quality labels to the trained CNN model for aesthetics estimation. Then, attention maps are obtained using Grad-CAM++ for 982 images estimated to be of high quality. The 982 attention maps
are classified into 16 clusters by applying k-means. As mentioned earlier, the number of clusters ($k = 16$) for landscape images was determined by the k-means method. The cluster centers of the obtained attention maps are shown in Figure 2. In the figure, the numbers in parentheses refer to the assigned cluster IDs, and the regions that contributed significantly to the decision are highlighted. Meanwhile, examples of typical images belonging to each cluster are shown in Figure 3. Figure 4 shows the heat map output by overlapping the attention map on each image in Figure 3. Figures 3 and 4 verify that the region around the characteristic object in the image contributes to the decision. Here, characteristic objects are man-made objects in the landscape, natural objects that are relatively small in area and conspicuous, and the boundaries of the sky, land, and sea.

We discuss the relationship between the obtained attention map and the composition. Cluster (1) contains many images with prominent boundaries such as horizons. These boundaries are located near the region that divides the image into the upper two-thirds and lower one-third. Therefore, it is indicated that cluster (1) is a group of images whose composition is strongly influenced by the rule of thirds. Similarly, in cluster (2), typical boundaries such as the horizon or the boundary between a lake and mountain are located in the horizontal line that divides the
Table 3. Details of each cluster and the associated composition for landscape

| Cluster ID | # of images | Composition          | Detail                                                                 |
|------------|-------------|----------------------|------------------------------------------------------------------------|
| 1          | 55          | Rule of thirds       | Boundary or object on the lower dividing line                          |
| 2          | 79          | Two-part composition | Boundary or object on the dividing line                               |
| 3          | 65          | Central composition  | The object or vanishing point in the center                            |
| 4          | 58          | Rule of thirds       | Object in the center of the right vertical dividing line              |
| 5          | 58          | Diagonal compositions| Objects or boundaries extend diagonally from the upper right          |
| 6          | 60          | Rule of thirds       | Object on the left vertical dividing line                             |
| 7          | 77          | Rule of thirds       | Object on the left vertical dividing line (main) and on the right dividing line |
| 8          | 61          | Rule of thirds       | Object on the right vertical dividing line (main) and on the left dividing line |
| 9          | 42          | Rule of thirds       | Boundary or object on the upper dividing line                         |
| 10         | 57          | Rule of thirds       | Boundary line on the lower horizontal dividing line, object at the lower right intersection |
| 11         | 59          | Rule of thirds       | Boundary line on the lower horizontal dividing line, object at the lower left intersection |
| 12         | 82          | Central composition  | Large object in the center                                            |
| 13         | 61          | Rule of thirds       | Object on lower left intersection                                      |
| 14         | 80          | —                    | —                                                                      |
| 15         | 56          | Rule of thirds       | Object on lower right intersection                                     |
| 16         | 62          | Diagonal compositions| Objects and boundaries extend diagonally from the upper left          |

image into two parts. Cluster (2) is a group of images that are influenced strongly by the two-part composition. Cluster (3) contains images in which the unique and salient object or vanishing point in the image is located in its center. This is a typical central composition.

Table 3 summarizes the number of images in each cluster and the details of the associated composition. Note that the “Detail” in Table 3 are verbalized interpretations by several subjects including the authors based on the results of the matching compositions by the subjective experiments. In the columns [Composition] and [Detail] of the table, “—” indicates the absence of matching. The regions used as the basis for the assessment is classified approximately by applying x-means to the attention maps obtained from the 982 landscape images determined to be of high quality. Thereby, we verify that many of the clusters obtained correspond to the three typical compositions, namely, rule of thirds (including the two-part composition), central composition, and diagonal composition.
Meanwhile, we show the experimental results for the portrait images. We input 1,500 training images with high-quality labels to the trained CNN model for aesthetics estimation. Then, attention maps are obtained using Grad-CAM++ for 1,238 images estimated to be of high quality. The 1,238 attention maps are classified into 12 clusters by applying x-means. As mentioned earlier, the number of clusters ($k = 12$) for portrait images was determined by the x-means method. The cluster centers of the obtained attention maps are shown in Figure 5. Figures 6 and 7 show typical example images belonging to each cluster and heatmaps overlaying attention maps, respectively. Table 4 summarizes the number of images in each cluster and the details of the associated composition. Note that the “Detail” in Table 4 are verbalized interpretations by several subjects including the authors based on the results of the matching compositions by the subjective experiments.

We discuss the relationship between the obtained attention map and the composition for portrait images. In clusters (3), (6), and (7), the characteristic objects in portrait photographs, such as the face and body, are located on the tripartite line or in the center of the image. Therefore, it can be inferred that these may have been evaluated by the rule of thirds or the center composition. However, although salient objects such as faces, bodies, and eyes are focused on in many clusters, it is difficult to identify a relationship with typical compositions, unlike in the case of landscape images.

In the environment we live in, the face or human body are one of many objects. However, the capability of humans to identify faces and human bodies is special. It is known to process significantly fine details compared with other object recognition. In addition, the face plays a vital role in human life. It is empirically evident and psychologically verified. Therefore, in portrait images containing individuals, the face (including the body) is a prominent and particular object.
Figure 6. Examples of typical images belonging to each cluster for portrait.

Figure 7. Heat map by overlapping the attention map for portrait.
that attracts strong attention. Hence, it can be considered that the manner in which the face or human body appears in an image has a more significant impact on its aesthetic evaluation than the position (composition) of the face or individual in it.

5. Conclusions
In this study, we used XAI to analyze whether the CNN model based on deep learning models for estimating aesthetics considers compositional information. This was performed with the objective of analyzing the crucial factors that contribute to the aesthetic quality of photographs. We constructed a CNN model for aesthetics estimation to extract image features for aesthetics from input images. Next, as a technique to visualize the learned concepts of CNN, we proposed a method to obtain attention maps using Grad-CAM++ and classify these using x-means. Experiments using AVA-dataset showed that compositional information is vital in aesthetic estimation for landscape images using the CNN model. Furthermore, we observed that the features considered differed depending on the genre of the photograph.

Several CNN models have been proposed, such as ResNet(He et al., 2016) and Xception (Chollet, 2017), that achieve better performance than VGG in ImageNet tasks. Therefore, in future work, further discoveries about the aesthetics and composition of photographs may be obtained by using the latest CNN models than the VGG as a feature extractor.

This research was supported in part by a grant-in-aid for Scientific Research (C) from the Japan Society for the Promotion of Science (Grant No. 21K11940).

Funding
This work was supported by the JSPS [21K11940].

Author details
Daichi Sakaguchi1
E-mail: sc188777@comwork.coka-pu.ac.jp
Hironori Takimoto2
E-mail: takimoto@coka-pu.ac.jp
ORCID ID: http://orcid.org/0000-0002-8795-7109
Akihiro Kanagawa2
E-mail: kanagawa@coka-pu.ac.jp

1 Graduate School of Computer Science and Systems Engineering, Okayama Prefectural University, Soja, Japan.
2 Faculty of Computer Science and Systems Engineering, Okayama Prefectural University, Soja, Japan.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Citation information
Cite this article as: Study on relationship between composition and prediction of photo aesthetics using CNN,
Daichi Sakaguchi, Hironori Takimoto & Akihiro Kanagawa, Cogent Engineering (2022), 9: 2107472.

References
Chattopadhyay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018). Grad-CAM++. Generalized gradient-based visual explanations for deep convolutional networks. Proc. of IEEE winter Conf. on applications of Comput. vision, 839–847.
Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proc. of 2017 IEEE Conf. on Comput. Vision and Pattern Recog, 1800–1807.
Datta, R., Joshi, D., Li, J., & Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. Proc. of the 9th European Conf. on Comput. Vision, 288–301.
Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. Proc. of IEEE Int. Conf. on Comput. Vision and Pattern Recog, 248–255.
Dhar, S., Ordonez, V., & Berg, T. L. (2011). High level describable attributes for predicting aesthetics and interestingness. Proc. of 2011 IEEE Conf. on Comput. Vision and Pattern Recog, 1657–1664.
He, K., Zhang, S. R., & Sun, J. (2016). Deep residual learning for image recognition. Proc. of IEEE Conf. Comput. Vision and Pattern Recog, 770–778.
Karen, S., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. Proc. of the 3rd Int. Conf. on Learning Representations.
Ke, Y., Tang, X., & Jing, F. (2006). The design of high-level features for photo quality assessment. Proc. of IEEE Conf. Comput. Vision and Pattern Recog, 419–426.
Kong, S., Shen, X., Lin, Z., Mech, R., & Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. Proc. European Conf. on Comput. Vision, 662–679.
LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep Learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539
Lu, X., Lin, Z., Shen, X., Mech, R., & Wang, J. Z. (2019). Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. Proc. of the IEEE Int. Conf. on Comput. Vision, 990–998.
Luo, Y., & Tang, X. (2008). Photo and video quality evaluation: focusing on the subject. Proc. of European Conf. on Comput. Vision, 386–399.
Luo, W., Wang, X., & Tang, X. (2011). Content-based photo quality assessment. Proc. IEEE Int. Conf. on Comput. Vision, 2206–2213.
Ma, S., Liu, J., & Chen, C. W. (2017). A-lamp: Adaptive layout-aware multipath deep convolutional neural network for photo aesthetic assessment. Proc. of IEEE Int. Conf. on Comput. Vision and Pattern Recog, 722–731.
Mai, L., Le, H., Niu, Y., & Liu, F. (2011). Rule of thirds detection from photograph. Proc. of IEEE Int. Symp. on Multimedia, 91–96.
Mai, L., Jin, H., & Liu, F. (2016). Composition-preserving deep photo aesthetics assessment. Proc. of IEEE Int. Conf. on Comput. Vision and Pattern Recog, 497–506.
Marchesotti, L., Perronnin, F., Larlus, D., & Csurka, G. (2011). Assessing the aesthetic quality of photographs using generic image descriptors. Proc. of IEEE Int. Conf. on Comput. Vision, 1784–1791.
Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. Proc. of IEEE Int. Conf. on Comput. Vision and Pattern Recog., 2408–2415.
Nishiyama, M., Okabe, T., Sato, I., & Sato, Y. (2011). Aesthetic quality classification of photographs based on color harmony. Proc. of IEEE Conf. on Comput. Vision and Pattern Recog., 33–40.
Omori, F., Takimoto, H., Yamauchi, H., Kanagawa, A., Iwasaki, T., & Ombe, M. (2019). Aesthetic quality evaluation using convolutional neural network. Asia-Pacific Journal of Industrial Management, 8(1), 71–77.
Pelleg, D., & Moore, A. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. Proc. of the 17th Int. Conf. on Machine Learning, 727–734.
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proc. of IEEE Int. Conf. on Comput. Vision, 618–626.
Shizuno, Y., & Hamasu, R. (2014). The automatic composition determination method at the time of the photography using the composition matching technique. Proc. of multimedia distributed cooperative and mobile symposium, 646–656.
Takimoto, H., Omori, F., & Kanagawa, A. (2021). Image aesthetics assessment based on multi-stream CNN architecture and saliency features. Applied Artificial Intelligence, 35(1), 25–40. https://doi.org/10.1080/08839514.2020.1839197
Talebi, H., & Milanfar, P. (2018). NIMA: Neural image assessment. IEEE Trans. on Image Processing, 27(8), 3998–4011. https://doi.org/10.1109/TIP.2018.2831899
Yang, G., Ye, Q., & Xia, J. (2022). Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion, 77, 29–52. https://doi.org/10.1016/j.inffus.2021.07.016
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. Proc. of the IEEE Conf. on Comput. Vision and Pattern Recog., 2921–2929.
