PhotoHelper: Portrait Photographing Guidance Via Deep Feature Retrieval and Fusion

Nan Jiang, *Member, IEEE*, Bin Sheng, *Member, IEEE*, Ping Li, *Member, IEEE*, and Tong-Yee Lee, *Senior Member, IEEE*

***Abstract*—We introduce a new photographing guidance (PhotoHelper) for amateur photographers to enhance their portrait photo quality using deep feature retrieval and fusion. In our model, we comprehensively integrate empirical aesthetic rules, traditional machine learning algorithms and deep neural networks to extract different kinds of features in both color and space aspects. With these features, we build a modified random forest with a structured photograph collection to identify types of photos. We also define the composition matching score to measure the similarity between the given photo and the reference photo. By combining all of the above processes, a one-stop deep portrait photographing guidance is constructed to provide users with professional reference photographs that are similar to the current scene and automatically generate spatial composition guidance according to the user-selected reference photo. Experiments and evaluations show that the aesthetic quality of portrait photos can be significantly improved via the composition guidance of our photographing guidance approach.**

***Index Terms*—Aesthetic assessment, deep feature fusion, image retrieval, photographing guidance, spatial composition rule.**

I. Introduction

With the development of smart-phone camera technology, it has become easier than ever before for a common person to take a photograph. Among the general public, portrait photos are the most popular type. However, amateur photographers usually lack skills to take photos of high quality. Common people are often incapable of designing a harmonious photo composition including the background and the person in the foreground. Conversely, experienced cameramen always handle these problems expertly based on their professional skills and long-accumulated experience, which are almost impossible to learn from textbooks or teaching videos. To help novices improve the quality of their portrait photos, the most direct method is retrieving professional photographs that are similar to the current scene and imitating them when taking pictures. However, it is almost impossible in practice to manually find similar photos in real time when photographing even if the photographs are indexed by category for fast access. In addition, learning the artistic style of professional photographs is also a difficult task for amateurs. The image search engine is another convenient choice since users can obtain similar photographs quickly [1], [2]. However, the retrieved results are sometimes of low quality. The search engine also cannot teach photographers how to reproduce a better photograph with the search results. Other studies tried to obtain the best composition by cropping out a smaller frame from the original photograph to help users obtain a better picture [3], [4]. However, the limitation of this method is also obvious. Since photographs taken by common people are usually of low quality, the subregion in the photo with high aesthetic value may also be very small; thus, the cropped picture will lose much of the information and feeling that the original photo intended to convey. With the knowledge that existing techniques are not proper for current situations, we sought to design a method to address these problems.

Here, we propose a one-stop portrait photographing assistant for retrieving professional reference photos and instructing users to improve the layout of their framing. Along with the idea of combining both scene type and spatial composition similarity [3], our model is designed in a feature-based manner [5] and is mainly composed of two parts: reference photo retrieval and photographing guidance generation. We combine empirical rules, existing traditional algorithms, deep neural networks and statistical methods to extract several kinds of image features. According to their functions, these features are divided into three types: learned features, global features and similarity matching features. In the reference photo retrieval step, learned and global features are used to train a modified random forest that can predict the scene type of the given photo. After that, the similarity matching features are used to calculate the composition matching score for comparing the similarity between each retrieved photo and the given photo. Finally, we show some top-ranked-score photos on the screen, and once the user selects one of them as his or her reference, our method will automatically provide visualized spatial composition guidance to the photographer. We show this process in a flowchart in Fig. 1. Notably, if users want to obtain photos particularly similar to the

Manuscript received 1 March 2021; revised 15 October 2021 and 8 January 2022; accepted 16 January 2022. Date of publication 25 January 2022; date of current version 7 June 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 61872241 and 61572316, in part by The Hong Kong Polytechnic University under Grants P0030419, P0030929, and P0035358, and in part by the Ministry of Science and Technology under Grant 110-2221-E-006-135-MY3, Taiwan. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Liangliang Cao. *(Corresponding author: Bin Sheng.)*

Nan Jiang and Bin Sheng are with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: njiang2021@sjtu.edu.cn; shengbin@cs.sjtu.edu.cn).

Ping Li is with the Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: p.li@polyu.edu.hk).

Tong-Yee Lee is with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan 70101, Taiwan (e-mail: tonylee@mail.ncku.edu.tw).

Digital Object Identifier 10.1109/TMM.2022.3144890

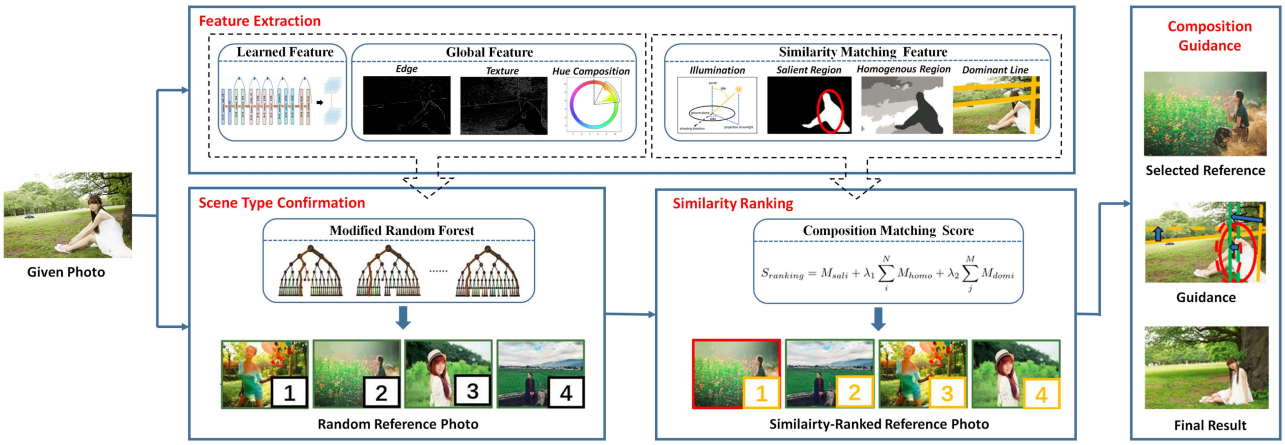


Fig. 1. The flow path of our method. We first extract learned features, global features and similarity matching features of the given photo. With learned and global features as input, the modified random forest predicts a certain scene type to the photo. Next, the composition matching score of each reference photo is calculated by the similarity between itself and the given photo with their similarity matching features. According to these scores, we get the ranking list of the reference photos, which will be shown to the user for selection in this order. Suppose the user chooses the photograph in red rectangle as reference, our method will automatically show the visualized composition guidance on the screen to help the user take a better picture.

scene in a certain aspect, minor adjustment for weight parameters of the corresponding feature can be carried out to meet this demand. The analysis of the corresponding evaluation and comparison experiments shows that our recommended photographs are similar to the current scene and users can produce better portrait photos by following our guidance. In general, our work makes the following key contributions:

1) *Comprehensive Feature Fusion:* We utilize empirical rules, traditional classification algorithms, deep learning, and statistical methods to comprehensively fuse various kinds of features, including color, illumination, and spatial composition features. In addition to some commonly used features, we also raise new features in our method, such as homogeneous regions and dominant lines, which are proven effective. With this feature-fused system, we construct a model for analyzing, classifying and retrieving photographs.
2) *Interactive Photographing Guidance:* By integrating the result of modified random forest and composition matching scores, we provide users with the most similar photographs to the current scene in our dataset. Once the user selects a photograph as a reference, our approach would show the corresponding visualized spatial composition guidance on the screen, which photographers can follow to produce a better photograph conveniently.

II. Related Work

A. Photograph Aesthetic Assessment

Although it has been studied for a long time, the evaluation of photograph aesthetics is still difficult to summarize as a common model. The existing knowledge of photography suggests there should be some general standards for photo evaluation underneath the current theory [6]. Most traditional studies on image assessment attempt to assess photos' aesthetic degrees based on various kinds of features. In previous studies, visual features, including color, texture and composition, were introduced. Color features include light, colorfulness, brightness, saturation and hue of the photo. In addition, color features are always different in different regions of a single photograph; thus, the distribution of these features in the whole picture is also considered a feature [7]. Texture features are often described with edge information, blur distributions, wavelet features and the histogram of oriented gradients. Composition features are also important in aesthetic evaluation. Some studies have segmented the image into regions and computed the importance of each region to extract salient regions and determine their relative distances [8]. Other studies have tried to use ovals and lines to represent the composition of the photograph [3]. In conclusion, most traditional image features are defined by empirical rules [9], and after the extraction of these features, the quality of the photograph can be evaluated by combining all the features, such as [10].

In recent years, the utilization of traditional features has been gradually replaced by learned features extracted from learning-based models in the computer vision field [12]. First, machine learning approaches for aesthetic evaluation have been extensively studied in the past ten years [13]–[15]. By constructing predictive models from large-scale labeled training data, machine learning methods can be more robust than traditional algorithms. Recently, deep neural networks have achieved remarkable results in many image processing tasks [11], [16], [17], including several methods to investigate image aesthetics prediction [18]. Lu *et al.* [19] adopted a deep neural network approach with unified feature learning and classifier training to estimate image aesthetics, which learned aesthetic-related features automatically from a large-scale image dataset. Campbell *et al.* [20] trained a 10-layer deep belief network to discover features of evolved abstract art images. Luo and Tang [14] presented a work to extract the subject region from a photo to evaluate its quality. Other researchers have tried to use a content-aware approach to remodel the scene [21]. However, the common disadvantage of all of the above studies is that they do not adequately explain what has been learned from the network; thus, it is almost impossible to understand the implicit aesthetic standard from

the corresponding evaluation. In our method, we take these previous studies as a reference to learn and improve their feature extraction methods. At the same time, we still use traditional aesthetic rules to choose and collect professional photos for our database. This makes the aesthetic value of our reference photo more analyzable and makes our feature extraction method more well directed for the following photo classification step.

B. Photographing Assistant Tools

The most direct idea to obtain a high-quality photograph is postprocessing. However, it may change or even damage the details of the original photograph, and it is often difficult for common people to use the corresponding professional tools. Thus, some methods instead provide auto-postprocessing techniques. Most of these studies focus on searching for a better subfigure from a given photograph. For example, Liang *et al.* [3] applied an example-based recomposition to crop the given photograph to a smaller frame that has a better layout than the original one. There are also many other kinds of recomposition models, such as the probabilistic model [22], the maximally aesthetic model [23], and the energy model [24]. However, cropping will decrease the resolution of the photograph, and some important regions may be dropped out during this process. Furthermore, if the original given photo is of low quality, there may be no acceptable result even with an exhaustive search. Thus, the demand for real-time photographing assistance tools has come into being.

Many kinds of modern real-time shooting techniques have already been developed and are intended to help users increase the color quality of their photo. Some methods also try to help users in different aspects, such as providing pose and composition suggestions [25], [26]. For PCs, consumer-level depth cameras such as Microsoft Kinect and Intel RealSense have been developed to collect the depth information of the objects. For mobile platforms, Yin *et al.* [27] proposed a socialized mobile photographing system to assist mobile users in capturing high-quality photos using both the rich context available from mobile devices and crowd-sourced social media. Wang and Cohen [28] proposed an algorithm that integrates matting and compositing into a single optimization process for implanting foreground elements into a new background. Cheng *et al.* [13] presented a method to learn the object spatial correlation distribution and applied this method to guide the composition arrangement of professional photos.

Based on these previous studies, we find that regarding photo enhancement, there has already been much work done for the purpose of postprocessing, including the development of many mature image-processing software programs. These methods may have good effects on the improvement of the color aspect of the photos, but at the same time, they almost have no use for spatial composition enhancement. In turn, there is little research focusing on the preprocessing or shooting guidance of photographing. In addition, many studies on the spatial composition of photos are often targeted at image assessment rather than image generation. Focusing on these problems and the research objective, we first combine deep neural networks and traditional algorithms to extract different kinds of features from the photo, and then use these features to quantify and visualize the difference between a given photo and its similar professional photo to achieve the goal of spatial composition guidance before photographing.

III. Model Construction

A. Data Preparation

Our method mainly contains two parts: professional photo retrieval and photographing guidance generation. First, we restrict the photo type in our dataset to portrait photos and construct our professional photo dataset, which includes a total of 4122 photos. All photos are collected from professional photography websites/electronic magazines and stored in our local server. To classify these photos, we manually label these photos by the environment where the photos were taken, such as grasslands and highways, and finally summarize all photos into 15 classes. During the whole working period, this reference photo collection is used to be both the training set for modified random forest and the data source for photo retrieval.

B. Feature Extraction

In our method, all features used are divided into three types: learned, global and similarity matching features. Learned features are used to train a random forest. Global features are used to improve the classification accuracy of some leaf nodes in the forest. Similarity matching features are used to quantify the similarity between the given photo and the reference photo. During feature extraction, in addition to using several traditional methods, we also developed some new aesthetic features and their extraction means.

1) Learned Features: Scene type, or in other words, the environmental background, is always one of the most important topics when implementing photo retrieval. However, scene type is not an object with some special characteristics but a macroscopic concept that is difficult for humans to quantify. Thus, we use a deep neural network to extract these abstract, incomprehensible scene types from the given photos as their learned features. In our method, a pretrained deep residual network [11] that has been well trained on the Places365 dataset [29] is employed to finish this work. In detail, we first input the photo into the pretrained ResNet and then remove the final classification probability layer of the network to obtain the raw 365-dimensional feature vector from the last FC. We then take this feature as the learned feature of the given photo. This process is shown in Fig. 2. On the one hand, photos of different scene types will produce different value distributions on some dimensions of their learned features (as in the example shown in Fig. 3(a)). On the other hand, photos of the same scene type will become easier to cluster on the t-SNE embedding plot by their learned features than the traditional GIST features [30] (as in the example shown in Fig. 3(b)). These experiments indicate that learned features indeed have some superiority over other image features when being used to implement photo classification to some extent.

2) Global Features: When learned features are extracted, global features of the given photo can also be calculated and prepared for the modification of the random forest in the next

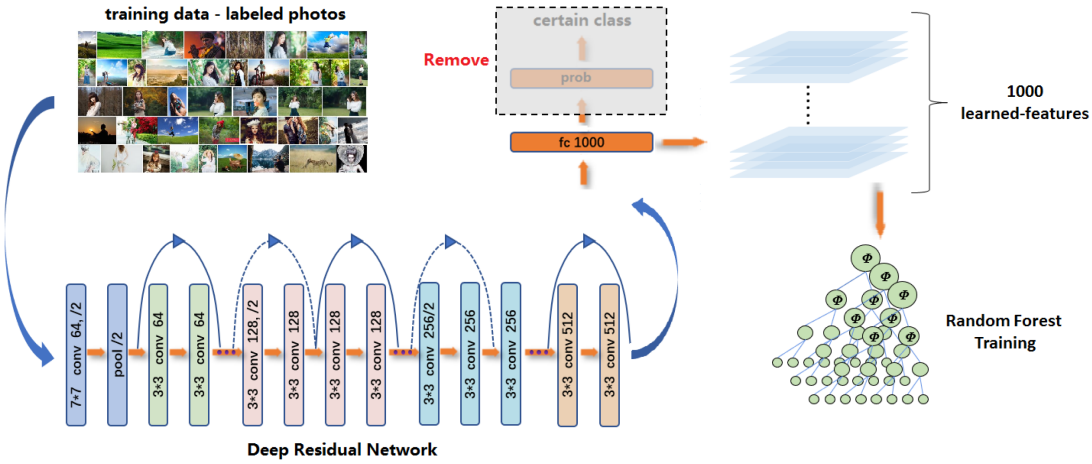


Fig. 2. The extraction of learned features. We use a pretrained ResNet [11] to get learned features. After removing the last probability layer, we take the output of the last FC layer as our learned features to train the random forest in the next step.

step. The global features used in this model include three parts: edge, texture and hue composition.

a) Edge: The edge of an image is defined as the set of pixels in which the grayscale of the surrounding pixels changes dramatically. In our model, taking the HOG (histogram of oriented gradient) feature extraction algorithm as a reference, we directly calculate the gradient of a certain pixel in RGB color space as Gp and its direction as Dp. After that, we implement a noise reduction process to eliminate the disturbance of the pixels with high Gp but not on the edge of the objects. Then, with the remaining edge pixels, we divide all of them into 6 classes from 0 to 180 degrees and take the quantity of pixels in each class as the corresponding edge length. Finally, based on these data, we construct the length-direction histogram E_v of the given photo and take it as its edge feature.

b) Texture: As a feature that does not depend on the change of color or brightness to reflect the homogeneity of an image, texture characterizes the distribution of the grayscale in a certain pixel domain. Taking the LBP feature extraction algorithm as a reference, we obtain the texture information of the target pixel by observing the grayscale distribution of its eight surrounding pixels. Based on the shape formed by pixels of similar grayscale in the 3×3 domain, we divide the texture pattern of all pixels into nine classes, including eight classes of different texture growth directions and one class of no texture. According to this classification, we take the texture direction Dt and the proportion of texture area At (the proportion of pixel amounts in one class) to quantify the texture of the image. Similar to edge feature extraction, an area-direction histogram T_v is constructed, and we take this vector as the texture feature of the given photo.

c) Hue composition: Luo *et al.* [21] proposed that, for most professional photographs, the hue composition can be summarized into several kinds of patterns. To describe these patterns as quantifiable features, we first construct the hue histogram of the image and search the dominant color intervals. The dominant color interval is defined as a continuous color-concentrated region in the histogram. In our model, we defined the interval covering more than 60% of pixels as the major interval and another (if it existed) that covers over 20% of pixels as the subinterval. Based on these intervals, we used two parameters to show the hue distribution modes: the hue focus of the dominant color interval (represented as α) and the spans of the interval (represented as ω). As shown in Fig. 4, two patterns of hue composition can be concluded.

3) Similarity Matching Features: Similarity matching features are used to calculate the composition matching score of each reference photo with the same scene type as the given photo. The similarity matching features used in this model include four parts: illumination conditions, salient regions, homogeneous regions and dominant lines.

a) Illumination condition: The impression of a scene is determined to a great extent by the prevailing illumination conditions [31], and intensity and direction are often regarded as two significant features of the illumination condition. Directly calculating the intensity of the current scene is difficult; therefore, we use brightness instead. The mean brightness I_m can be calculated by the following equation:

$$I_m = \frac{\sum_{i=1}^{N} \max(i_r, i_g, i_b)}{N} \tag{1}$$

where, i_r, i_g, i_b means the RGB value of the pixels. And N is the the amount of all pixels in the photograph. Furthermore, we can also calculate the contrast C of the photo using the following equation:

$$C = \frac{I_h}{I_l}$$
$$\text{s.t. } I_h = \frac{1}{N_h}\sum_{i=1}^{N_h} I_p, \quad I_l = \frac{1}{N_l}\sum_{i=1}^{N_l} I'_p \tag{2}$$

where, I_p is the mean brightness of the pixels higher than I_m while the I'_p represents the lower one. To obtain the illumination direction, we need to know the light source location. Lalonde *et al.* [31] proposed a method to estimate the sun visibility and location with zenith angle and azimuth angle, which we show in Fig. 5(a). Inspired by their work, we modify their algorithm to a simplified version, as shown in Fig. 5(b). We define I_d as the symbol of the light direction feature and divide the ground plane into four regions to summarize lights from all directions into three classes: front-light, side-light and back-light. We then

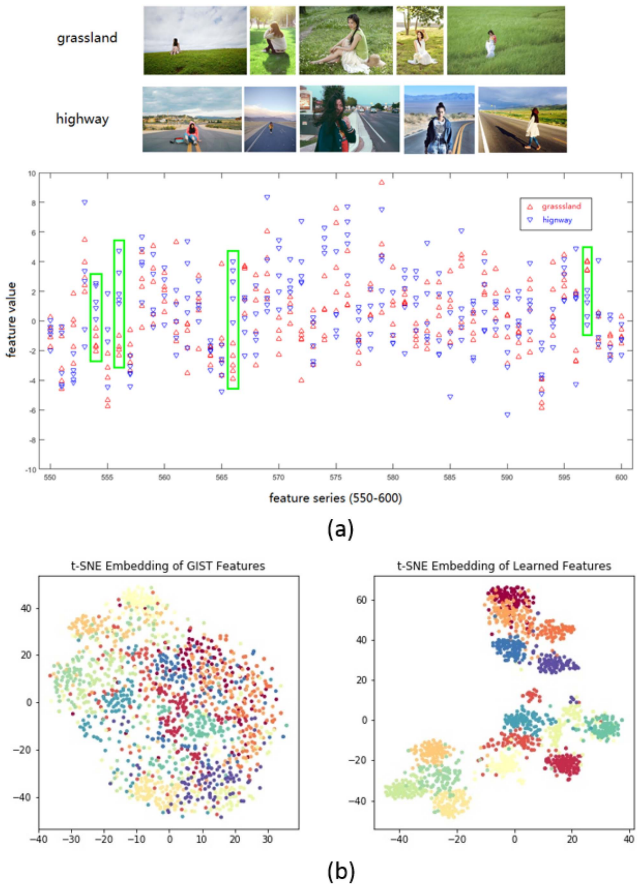


Fig. 3. The distribution of learned features' value. In (a) we show the grassland and highway photos. Part of their learned features' value are shown in the scatter plot below. We use green rectangles to show that two kinds of photos have significant value distribution differences at some certain feature dimensions. In (b) we show the t-SNE embedding plot for the learned features and traditional GIST features respectively, indicating the advantage of learned features in this classification task.

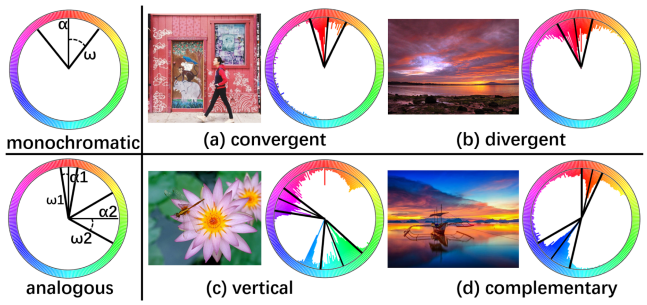


Fig. 4. The hue composition of monochromatic and analogous modes. The monochromatic mode contains one dominant color region with convergent or divergent hue distribution as in (a) and (b). The analogous mode contains two dominant color region with vertical and complementary hue distribution as in (c) and (d).

take these three labels as the feature values of I_d instead of those two angles. If there is no obvious light source in the surrounding environment, we take this as the front-light situation where the light zenith angle is set as $45°$.

b) Salient region: For a given photograph, a salient region is defined as the region occupied by the focused object to which the photographer wants viewers to pay attention. In our method,

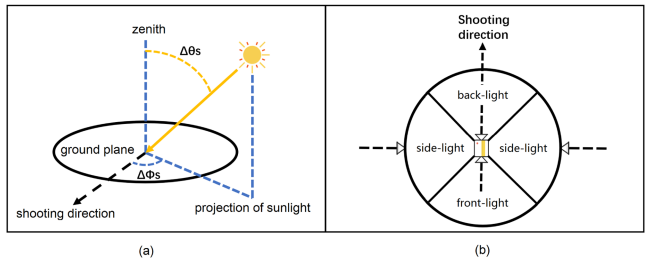


Fig. 5. Light direction features. In (a), the zenith angle is the angle between the zenith and the sunlight direction. The azimuth angle is the angle between the shooting direction and the projection of sunlight. In (b), we divide ground plane into 4 regions and conclude lights from all directions into 3 classes: front-light, side-light and back-light.

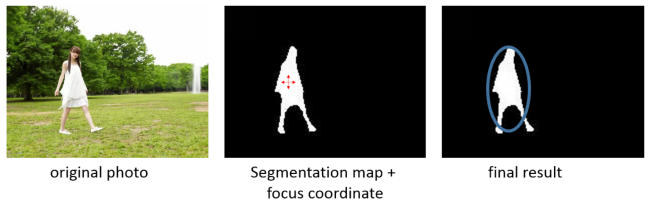


Fig. 6. Salient region extraction. The left shows the original photo, the middle shows the combination of segmentation map and focus coordinate. The right shows the final result.

we use both deep learning and focus point coordinates to extract the salient region of the photo. First, Pose2Seg [32] is used to obtain the corresponding segmentation map of the given photograph. Then, after the user touches the screen to specify the focus point, we record this touching coordinate and determine the semantically segmented region in which the touching coordinate falls to define it as the salient region. (We also manually define the area where the person stands as its salient region in our reference photo dataset.) Finally, to represent the salient region, we use ovals to circle the salient region and record the oval with central coordinates (Sr_x, Sr_y), major axis length O_a and the angle O_{angle} between the major axis and y axis as the salient region feature. We show the whole process in Fig. 6.

c) Visual homogenous regions: In a photo, the homogenous region is defined as the area where each of its subregions shares similar hue, texture, intensity, etc., usually covering integral, semantically similar objects. In our method, we first take a graph-based image segmentation method [33] to directly divide the hue map of the photo into several superpixels. Then, we use a graph $G(V, E)$ to record the segmented regions and their relationship. (The salient region obtained before will not be processed here.) Node V of the graph is the representation of a homogeneous region. For each node, we save at most three main hue values $(Hr_{h1}, Hr_{h2}, Hr_{h3})$ of the region, their proportions $(Hr_{r1}, Hr_{r2}, Hr_{r3})$, the central coordinates of the whole region (Hr_x, Hr_y) and the size of the region Hr_s. Edge E of the graph records the connected region pairs. With all these elemental variables defined, we start to construct more meaningful functions.

Suppose R_1 and R_2 are the region graph nodes discussed before. We define $SR(R_1, R_2)$ as the size ratio of two regions, which is used to encourage a large region to absorb a smaller

region instead of another large region.

$$SR(R_1, R_2) = \frac{\max(R_1.Hr_s, R_2.Hr_s)}{\min(R_1.Hr_s, R_2.Hr_s)} \tag{3}$$

We define $D(R_1, R_2)$ as the distance between two regions, which is the length of the shortest visited path between them. $D(R_1, R_2)$ is used to encourage the merging of adjacent regions instead of the far-separated pairs.

We also define $H(R_1, R_2)$ as the hue similarity of two regions. We use the distribution of hue in the corresponding region to calculate it. According to our observation, we assume that the overall hue distribution can be described as a sum of several Gaussian functions. Thus, we take the hue distribution HD as:

$$HD(x|R) = \sum_{i=1}^{3} \frac{R.Hr_{ri}}{\sqrt{2\pi}} exp(-\frac{1}{2}(x - R.Hr_{hi})^2)$$

$$\text{s.t.} \sum_{i=1}^{3} R.Hr_{ri} = 1 \tag{4}$$

Therefore, the hue similarity can be calculated as:

$$H(R_1, R_2) = \sum_{i=0}^{255} \min(HD(i|R_1), HD(i|R_2)) \tag{5}$$

Finally, we can define the similarity score S_{simi} as the similarity measurement between two regions:

$$S_{simi}(R_1, R_2) = \frac{SR(R_1, R_2) \cdot exp(\alpha + H(R_1, R_2))}{\sqrt{D(R_1, R_2)}} \tag{6}$$

where, the coefficient α of Eq. (6) is used to control the weight of hue similarity. In our experiments, we set $\alpha = 3$. During the merging process, we merge one region pair that has the highest similarity score in this calculating circle and then update the attribute of this merged area as a new region in the next iteration until only five regions remain.

Our homogeneous region results are shown in Fig. 7. We can see that the regions close to each other with similar semantic meaning have been merged into one part. Although the precision of this segmentation is not perfect, by matching the homogeneous regions, we can still ensure the rough composition similarity of two photographs.

d) Dominant lines: A dominant line is defined as the boundaries between different homogenous regions that are long enough to influence the composition of the photograph, such as the skyline. Most of the composition rules are related to the location of dominant lines, such as the Rule of Thirds and the Diagonal Rule. To extract the dominant lines, we first perform a Randomized Hough Transform [34] on the intermediate product of the region merging process discussed before. Then, we use the following formula to identify and merge the short lines to construct the final theoretical dominant line:

$$L_{simi}(L_1, L_2) = \frac{w_1}{exp(A(L_1, L_2))} + \frac{w_2}{exp(D(L_1, L_2))} \tag{7}$$

where, L_1, L_2 are two lines extracted by Hough Transform. $A(L_1, L_2)$ is the angle between line pair L_1, L_2 and $D(L_1, L_2)$ is the distance between two lines' centers. w_1, w_2 are two weight parameters to balance these two factors which we set $w_1 = 0.15$ and $w_2 = 0.1$ respectively.

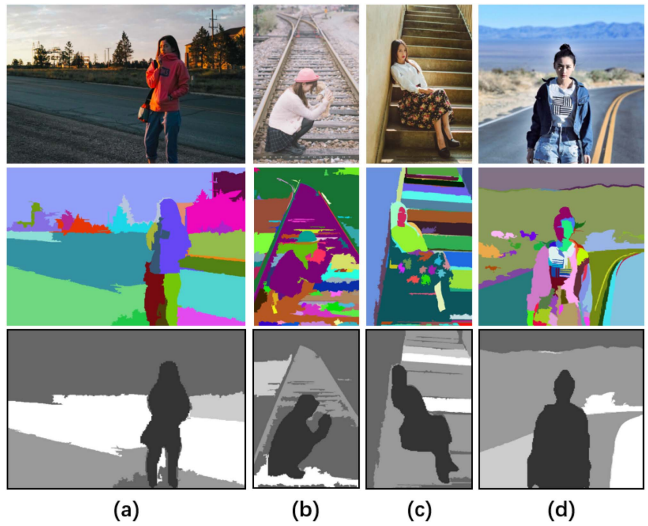


Fig. 7. Homogenous region merging results. The first row shows the original photos. The second row shows the direct segmentations. And the last row shows our merging results.

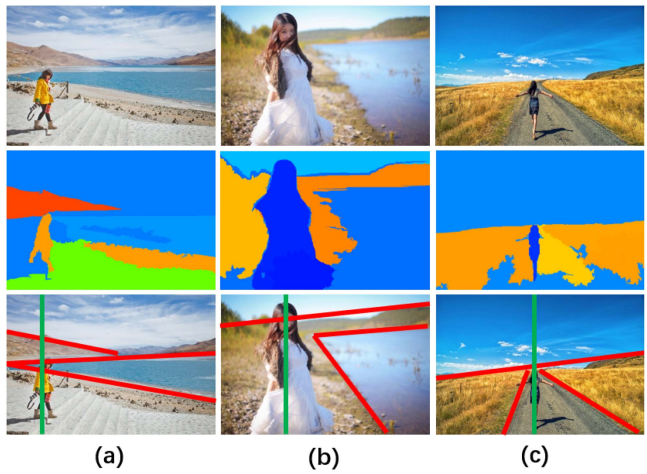


Fig. 8. Dominant lines examples. We show three different dominant line results in (a), (b), and (c). Red lines are dominant lines and green ones are the long axis of the salient oval. From top to bottom: the original images, the intermediate region merging results and dominant line results.

In this experiment, we merge only the line pair of which the L_{simi} value exceeds the threshold set as 1.2. After this process, most short, close lines are merged into several long straight lines. Finally, we select at most three longest lines from the remaining ones as dominant lines. We show part of the results in Fig. 8.

C. Photograph Recommendation

Photograph Recommendation includes two parts: professional photos batch retrieval (based on a modified random forest) and similarity matching ranking (based on the composition matching score).

1) Modified Random Forest: After feature extraction, 4122 labeled photos with learned and global features are used to train a modified random forest with 300 CARTs. Bootstrap sampling

is used to build up the training set for each tree, and the attribute space is composed of 365 features (all 365 dimensions of the learned feature vector) extracted from ResNet. For each node on each tree, we randomly select 30 features as candidates and take one of them with the best Gini value for splitting. Trained in this way, every tree in the forest grows freely without pruning until each leaf node contains only photos of the same class or there are no features to be split. The minimum examples per end node are manually set to 20 to restrict the depth of the tree and improve the efficiency of the whole forest. The learned features extracted from the network are all local features that are sensitive to the component parts of an object. However, scene type is an abstract macroscopic concept, so global features also play an important role in this task. Thus, we verified each tree to pick out the leaf nodes in which the proportion of major photo class accounts for less than 90 percent and further split these nodes using the 3 global features discussed above. Finally, after modifying the whole forest, we define the class of each leaf node as the class where most photos of this node fall.

2) Composition Matching Score: In this section, we combine all four similarity matching features to compute the composition matching score for each reference photo. To be more specific, we first calculate three subscores according to each corresponding feature and then add them up with different weights to make the final score.

a) Illumination condition: To filter photos by this feature, we first eliminate the photos with different light directions and then further eliminate the photos in which the brightness or contrast differs from the given photo by over 20%. After this process, we finally obtain the primary photo subset for the matching score calculation.

b) Salient region: To measure the salient region similarity, we first resize two photographs into the same size and correspondingly transform the salient oval. Then, the salient matching score can be calculated by the following equation:

$$M_{sali} = m_{sali}(S_{r1}, S_{r2}) = \frac{\lambda_1}{d} + \lambda_2 cos(\theta) \tag{8}$$

where, d is the distance between two elliptical centers and θ is the included angle formed by long axis of two ovals. λ_1 and λ_2 are adjustable parameters to balance the importance of distance and direction. In this model, we set λ_1 and λ_2 to 0.05 and 10.

c) Visual homogenous regions: Similar to the process of salient region matching, we resize two photographs and correspondingly transform the homogenous regions. The calculation of each region pair's matching score is very similar to the computation of the region merging process. The equation is shown below:

$$m_{homo}(R_1, R_2) = \frac{JC(R_1, R_2) \cdot exp(\alpha + H(R_1, R_2))}{\sqrt{D(R_1, R_2)}}$$
$$\text{s.t. } JC(R_1, R_2) = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|} \tag{9}$$

where, R_1, R_2 are two homogeneous regions of different photographs. $H(R_1, R_2)$ and $D(R_1, R_2)$ are hue similarity and central distance that have already been defined in region merging section. To be noticed, the $H(R_1, R_2)$ and $D(R_1, R_2)$ here are calculated with two regions from different photos rather than from the same picture as in the region-merging process. (The reasonability of this calculation is because we have resized the photos and thus can put them into the same coordinate system.) $JC(R_1, R_2)$ is the Jaccard coefficient, which is used to calculate the overlapping degree of two sets. For each homogenous region on the A photo, we compared it with every region on the B photo to calculate the matching scores and take the one with the highest score as its matched region until all regions of the A photo have been paired. Specifically, if one larger region is matched to a smaller one, the overlapped region will be subtracted from the larger region. Then, the remaining part will be regarded as a new homogenous region to be matched in the next iteration. Finally, the homogenous region matching score can be calculated as:

$$M_{homo} = \frac{\sum_{i=1}^{N} m_{homo}(R_i, R_i')}{N} \tag{10}$$

where, N is the number of matched region pairs, (R_i, R_i') is the best matched pair of R_i.

d) Dominant lines: The matching score of each dominant line pair can be calculated by the following equation:

$$m_{domi}(L_1, L_2) = \frac{1}{exp(\frac{s}{S})} \tag{11}$$

where, L_1, L_2 are two dominant lines of different photographs. s is the amount of pixels in the quadrangle formed by linking four endpoints of these two lines and S is the total pixel number of the photograph. Similar to the process of homogenous region matching, we determined every best matched line pair until no pair could be matched between two photos and finally calculated its average value as the dominant line matching score:

$$M_{domi} = \frac{\sum_{i=1}^{N} m_{domi}(L_i, L_i')}{N} \tag{12}$$

where, N is the number of matched line pairs, (L_i, L_i') is the best matched pair of L_i.

e) Composition matching score: To sum all submatching scores up with different weight parameters, we finally obtained the composition matching score $S_{ranking}$:

$$S_{ranking} = M_{sali} + \lambda_1 \sum_{i}^{N} M_{homo} + \lambda_2 \sum_{j}^{M} M_{domi} \tag{13}$$

where, N is the number of homogeneous regions matched pairs and M is the number of dominant line matched pairs. λ_1 and λ_2 are adjustable parameters to balance the proportion of three sub matching scores. In our model, we empirically assumed that the dominant line is the most important feature contributing to the similarity, followed by homogenous regions and finally salient regions. Therefore, λ_1 and λ_2 are set to 1.5 and 3, respectively. For comparison, we display a different retrieval result when the two parameters are reversely set to 3 and 1.5 to show the influence of these parameters in Fig. 9.

Now, with $S_{ranking}$ of all recommended photos in hand, we can rank photographs according to their similarity with respect to the current scene.

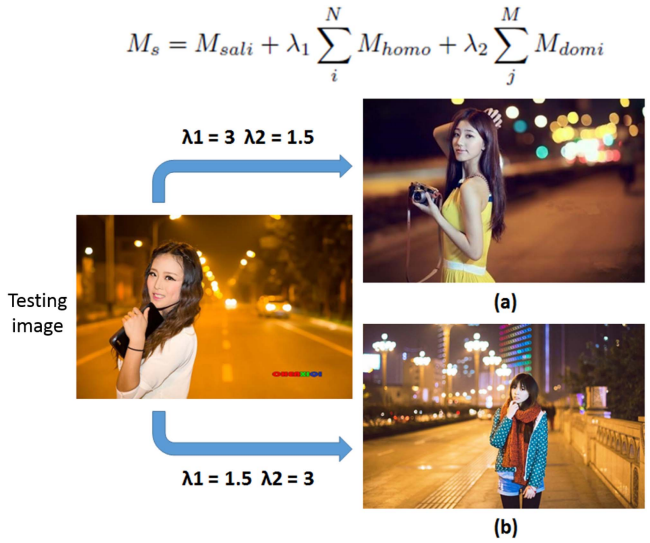


Fig. 9. The influence of different parameter values to the retrieval. In (a), we put more weights on homogeneous regions while in (b), we put more weights on dominant lines.

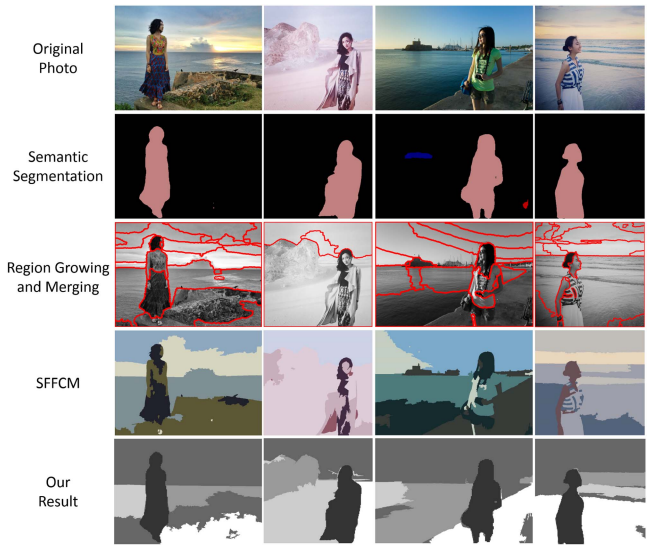


Fig. 10. Comparison with other region segmentation results. For rows from top to bottom: the original photographs; semantic segmentation [35] results; region growing and merging [36] results; superpixel based FCM clustering [37] results and our results.

D. Photographing Guidance

The color composition of the photo is easier to adjust using other postprocessing applications, such as Photoshop and Photo-Paint. Therefore, we concentrated more on spatial composition guidance instead. Notably, the distribution of the homogenous regions is a natural element of the current scene, which is very difficult to change unless the overall background is replaced. Therefore, only the salient region and dominant lines are considered in our composition guidance. We use full lines to mark these two features of the current scene and dotted lines for those of the reference photo. Given these auxiliary lines, users can change the shooting direction or focal length to imitate the spatial composition of the reference photo.

IV. Evaluation Tests of the Model

In this section, we design different evaluation tests targeting to different parts of our model. For our newly developed features, we directly show the results of different feature extraction methods to point out the suitability and superiority of our algorithms. For the modified random forest, we perform quantitative comparison experiments to compare it with other existing classification strategies on two different datasets. Finally, for photographing guidance, we conduct a user study and explain the effectiveness of our model based on the results analysis.

A. Evaluation of New Developed Features

In this part, we will prove the availability and suitability of our two newly developed features: homogeneous regions and dominant lines. First, we compare homogeneous regions with 3 existing image segmentation models [35]–[37] in Fig. 10. As defined in the previous section, a homogeneous region is the area where its subregions share similar hue, texture, intensity, etc. with no concern about the component objects. In this way, there is neither necessity nor appropriateness to use pixel-level semantic segmentation for our model. As shown in Row 2 of the figure, the algorithm [35] only segments persons and boats (blue region in Column 3) in the picture with no regard to the other part of the background. This is due to the limitation of its training samples and is far from suitable for our model. In addition, we also show 2 other segmentation results by the region growing and merging method [36] (Row 3) and the superpixel-based FCM clustering method [37] (Row 4). From the perspective of segmentation strategy, these 2 methods are more suitable for our model than semantic segmentation, with even some similar concepts to our algorithm. However, both of these methods have some problems of oversegmentation in the background or in the objects. In contrast, our results provide more correct but simpler segmentations in all 4 examples shown.

For dominant lines, we also compare our algorithm with 3 popular line detection models in Fig. 11: LSD [38] (in Column 2), Markov chain marginal line segment detector [39] (in Column 3) and CannyLines [40] (in Column 4). According to the definition of the dominant line, we only choose the 3 longest lines detected in each picture by different algorithms to be their dominant lines. As shown below, for the straight lines in the original picture, all 4 algorithms extract them correctly with little difference in length. However, compared with our results, other algorithms may provide some unimportant lines that are neglectable in the original picture (such as in Row 1 Column 3 and Row 2 Column 4) or may fail to observe the abstract lines formed by some short lines with little differences in the direction (such as the line formed by the mountain ridge in Row 2). In contrast, our algorithm gives all important segmentation lines and shows the spatial composition of the picture most clearly in all 4 examples.

B. Evaluation of Classification Performance

For the modified random forest, we also compare it with 4 different classification strategies to evaluate the performance of our method. Since this part of the model is targeted for photo

TABLE I
CLASSIFICATION REPORT OF DIFFERENT MODELS

	Scene-15 Dataset				Our Dataset				Our Dataset(gray)			
Models	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
ResNet(Places365)	0.88	0.85	0.85	0.86	0.90	0.86	0.86	0.84	0.84	0.79	0.82	0.81
ResNet(ImageNet)	0.82	0.81	0.85	0.82	0.87	0.85	0.88	0.87	0.80	0.76	0.79	0.77
InceptionV3(ImageNet)	0.78	0.82	0.78	0.77	0.84	0.86	0.84	0.84	0.77	0.68	0.71	0.70
GF+RF	0.66	0.60	0.62	0.62	0.71	0.62	0.69	0.68	0.62	0.59	0.63	0.62
LF+RF	0.93	0.87	0.87	0.87	0.92	0.86	0.89	0.88	0.88	0.83	0.85	0.84

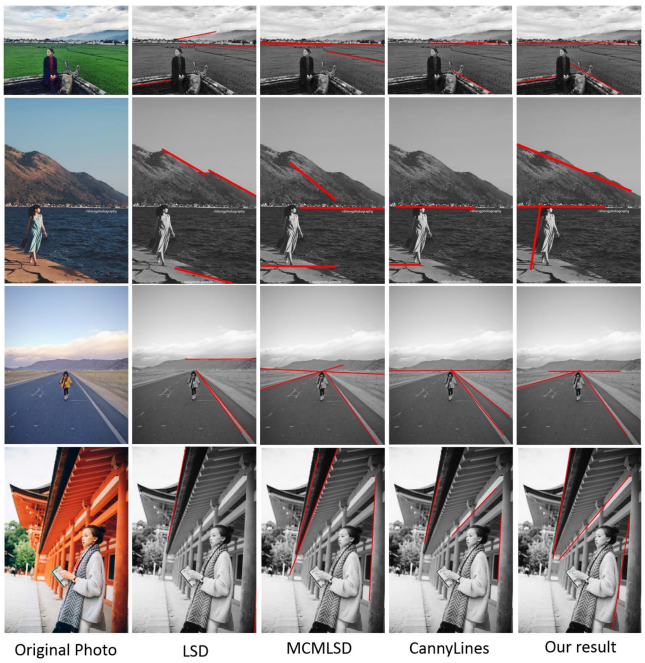


Fig. 11. Comparison with other line detection results. For columns from left to right: the original photographs; LSD [38] results; MCMLSD [39] results; CannyLines [40] results and our results.

classification, except for our own reference photo dataset, another public scene type classification dataset, Scene-15 [41], is also used in this evaluation test. A detailed description of these 2 datasets is as follows:

*1) Scene-15 Dataset:*Fifteen scene classes, including bedrooms, houses, factories, and supermarkets, are included in this dataset. Each class has approximately 200-300 photos with no people in every picture. We randomly select 150 photos in each class as training samples and use the other 50 photos as testing samples; thus, in all, we have 2250 photos in the training set and 550 photos in the testing set.

2) Our Dataset: Due to the insufficient background scenes in our reference photos, we summarize 15 representative classes of scene types for the overall photo collection, including highways, grasslands, snowscapes, and streetscapes. Each class has approximately 200-300 photos with a person in the foreground in every picture. We randomly select 150 photos in each class as training samples and 60 other photos as testing samples; thus, in all, we have 2250 photos in the training set and 900 photos in the testing set.

In this experiment, we carefully design four other classification strategies to evaluate the effectiveness of our model from different aspects and show the results in Table I and Fig. 12. In detail, our model is composed of a ResNet pretrained on Places365 and a modified random forest (represented as LF+RF). The other models are described as follows: an individual ResNet pretrained on Places365 (represented as ResNet(Places365)), an individual ResNet pretrained on ImageNet (represented as ResNet(ImageNet)), an individual InceptionV3 pretrained on ImageNet (represented as InceptionV3(ImageNet)), and a GIST feature extraction model with a modified random forest (represented as GF+RF).

As shown in Table I and Fig. 12, first, the 4 models with deep features all perform much better than the one with traditional GIST features, which shows the power and effectiveness of the learned features. Then, for these models based on deep learning, ResNet(ImageNet) gives a better result than InceptionV3(ImageNet) and also performs a lower prediction accuracy than ResNet(Places365). This indicates that, on the one hand, the ResNet structure is more powerful than InceptionV3 for this classification task; on the other hand, the network pretrained on Places365 can produce better results since it is used for a similar scene type classification task, which is consistent with the theory of transfer learning to some extent. Finally, our model outperforms the other four models on both Scene-15 and our own reference photo dataset, showing the effectiveness and generalization of our method.

Most of the photos in the Scene-15 dataset are of the grayscale type, while those in our dataset are of the RGB type. To explore the effect of this hue difference on the models, we transform all photos in our dataset into grayscale type and use the transformed photos as a new dataset to retrain the models. As expected, since the scene background has few areas and often includes humans in the portrait picture, color plays a significant role while the models try to classify the photos of our dataset. Thus, after removing the hue information by gray processing, the testing results of all 5 models decrease (especially when distinguishing 'snowscape' between 'forest,' or 'grassland' between 'wheatland'). However, under this situation, our model still performs better than the other four methods. This may indicate that our model has less dependency on hue information when implementing classification, which shows a stronger robustness than the other models to some extent.

C. Evaluation of Photo Matching

As the main purpose of our model, since it is used to provide some spatial composition guidance for novices, the professional photos recommended by our method are supposed to show more composition similarity to the testing photo rather

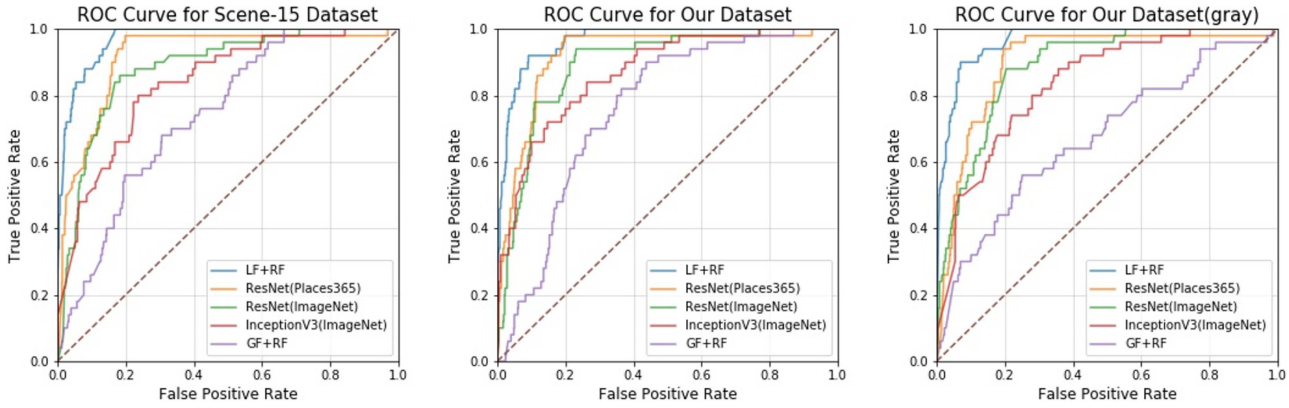


Fig. 12. Classification performance of different models. We show ROC curves of different models performing on 3 datasets respectively. On every dataset, our model achieves the highest AUC score.

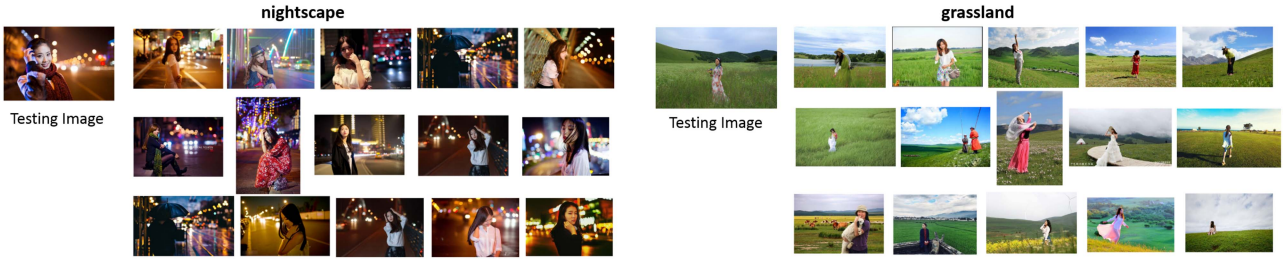


Fig. 13. Comparison to other similarity metric. We show two scene examples including 'nightscape' and 'grassland'. For each testing image, we show the top 5 best-matched images ranked by 2 different image similarity metrics of which each row corresponds to one. From top to bottom: ranked by our method, by SSIM and by cosine similarity.

than its color aspects. To validate these results, we use another 2 popular image similarity metrics, SSIM and cosine similarity, to rank the same group of recommended photos in different orders and compare their best matched photos with each other. As shown in Fig. 13, photos ranked by our method obtain more similar elements, such as spatial perspective and region composition, than those ranked by the other 2 methods. This indicates that the composition matching score developed by our method can indeed maintain more composition similarity between photos than the common metrics, which lays a solid foundation for the following composition guidance of our model.

D. User Study of Aesthetic Quality

We conduct a user study to evaluate the aesthetic quality of a photo that is taken after the guidance given by our application. The goal of this experiment is to determine whether our method is helpful to people lacking photography skills and to gather suggestions from users to improve our system.

1) Participants: Thirty volunteers without professional skills participate in the experiment. All of them are undergraduates. Before they start the experiments, we help them become familiar with our application to avoid interruptions.

2) Tasks: We design a comparison experiment with five different shooting modes: No-Guidance mode, Only-Photo mode, Only-SR mode, Only-DL mode and Full-Guidance mode. In No-Guidance mode, users take photos without any help from our application. In Only-Photo mode, users are only provided with recommended photos with no guidance on them. In the Only-SR mode and Only-DL mode, users can obtain recommended photos with only salient region suggestions or dominant line suggestions. Finally, users can obtain reference photos with all suggestions in Full-Guidance mode. Grasslands and highways are chosen to be shooting locations because photos taken from those places always have simple spatial compositions, which makes it easier and more precise to grade these photos from aesthetic degrees by our judges. Each volunteer is asked to take photos of 3 different scenes (randomly chosen by themselves) at one place, and for each scene, the volunteer should take 5 pictures in the 5 different shooting modes mentioned above. To eliminate the influence of the suggestions on volunteers, the sequence of shooting in each mode is No-Guidance mode, Only-Photo mode, Only-SR mode, Only-DL mode and Full-Guidance mode.

3) Scoring: We invite other 30 online and 30 offline judges to grade the photographs taken by the volunteers. All 450 photos are divided into 90 groups, and each group includes 5 photos taken in 5 different modes at the same scene by the same volunteer. Judges are provided with photos of one group after another, and each photo is graded according to their aesthetic degree from 0 to 10. Notably, sequences of groups given and photos in each group are both randomly shuffled to eliminate preference bias.

4) Scoring Results: With all grading results in hand, we summarize all photos taken in the same shooting modes to calculate the average score and standard deviation of each mode. As we can see in Table II and Fig. 14, Full-Guidance gets the highest score, while No-guidance gets the lowest score,

TABLE II
RATING RESULTS OF 5 PHOTOGRAPHING MODES

	Grassland		Highway	
	average score	standard deviation	average score	standard deviation
No-Guidance	6.2	2.59	6.8	2.62
Only-Photo	8.3	1.37	7.9	1.32
Only-SR	8.2	2.01	8.1	1.95
Only-DL	8.7	1.23	8.8	1.58
Full-Guidance	9.2	0.92	9.5	1.17

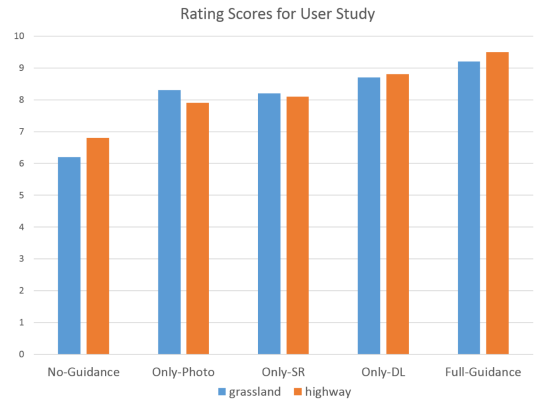


Fig. 14. Histogram of average rating scores.

demonstrating that our method can indeed improve users' photographing skills, such as view-choosing or shooting. Only-Photo and Only-SR receive almost the same scores, obviously higher than No-Guidance and slightly lower than Only-DL. This situation may imply two points. First, after being given reference photos, photographers imitate their composition naturally, which makes their new photographs better than their original ones. Second, during imitation, the salient region (the location of the person) can be copied simply, while the dominant lines in the picture are relatively harder to imitate or even notice, which explains why Only-Photo performed almost the same as Only-SR, but both ranked lower than Only-DL.

For the standard deviation, we find that shooting with more effective suggestions may produce lower standard deviations, which implies that without guidance, photographers choose view and composition randomly, neglecting aesthetic rules. These kinds of photos may meet the demands of certain judges but generally would be regarded as mediocre, which leads to high standard deviations with low average scores. In contrast, photos taken with guidance usually accord with some artistic regulations, thus pleasing the criteria of the general public, resulting in low standard deviations with relatively high average scores. In conclusion, our system can indeed help users create appealing photos of which the aesthetic quality is significantly higher than average level of the general public.

E. Subjective Evaluation of Aesthetic Quality

In addition to subjective user study, we further implement an objective assessment for the photos improved by our photographing guidance. The experimental details are as follows: First, we use two different photo optimization algorithms (RAG [22] and CAW [24]) on the 90 photos to produce their corresponding better versions by composition rearrangement (each photo is taken by the volunteers from the user study without our photographing guidance). Then, together with photos taken by volunteers after our photographing guidance, for each original photo, we obtain three improved versions. After that, we use two other image assessment models (DMA-Net [42] and A-Lamp [43]) to grade all photos (from 1 to 10) and calculate the average increased scores for the photos improved by the different methods. Thus, together with the voting from the user study, we ultimately obtain 9 average increased rating scores by 3 different photo improving methods with 3 different assessment standards. The results are shown in Table III. As seen from the results, whether based on subjective user voting or objective assessment model grading, photos taken with our photographing guidance always show significant progress in aesthetic quality compared with the original photos, which soundly proves the practical value of our model. Moreover, although other traditional optimization models do have some positive effects on the photo composition, our model still obtains the highest scores among all the methods, thus proving the superiority of our model over other existing works.

TABLE III
INCREASE OF RATING SCORES BY 3 MODELS

	DMA-Net	A-Lamp	User Voting
RAG	1.01	0.57	0.20
CAW	0.83	0.33	0.21
Our Model	1.69	1.89	2.77

TABLE IV
TIME PERFORMANCE OF EACH PROCESSING STEP

Feature Extraction	Learned Feature	Learned Features	0.1623s	3.9958s
	Global Feature	Edge	3.7061s	
		Texture	1.2257s	
		Hue Composition	0.0563s	
	Composition-Matching Feature	Illumination	3.9958s	
		Salient Region	1.0527s	
		Homo Regions	3.9465s	
		Dominant Lines	0.2916s	
Photograph Recommend				0.7661s
Guidance Generate				0.0097s
Total				4.7716s

F. Evaluation of Time Consumption

Our experiments are mainly implemented in a MATLAB platform and run on a single Intel(R) Core(TM) i7-6700 K CPU without GPU acceleration. As shown in Table IV, the extraction of different features is parallel. Thus, we take the time consumed in the visual illumination section, the longest one, as the overall time spent in feature extraction. We can see that the total time is approximately 5 s, which is slightly long for users to wait for guidance when photographing. Thus, we leave the optimization of our method and deploying the program on a GPU as our future work. We also record the time consumption of different

Fig. 15. Two examples of our method. For each row, we show a whole process of photographing guidance. For the columns, we show (a) original photo, (b) selected reference photo, (c) Unselected reference photo, (d) composition guidance, and (e) photo after guidance respectively.

TABLE V
AVERAGE USER SPENDING TIME

	Full-Guidance Mode	No-Guidance Mode
Before Focus	2.69s	5.73s
Wait Photo	6.33s	-
Select Ref	1.79s	-
Wait Guidance	0.12s	-
Manually Adjust	3.30s	-
After Focus	1.77s	3.80 -
Total	14.23s	9.53s

steps in different modes and show the results in Table V. We divide the time spent taking photographs into six sections: 'Before Focus,' the period from turning on the camera to specifying the focus region; 'Wait Photo,' the period from specifying the focus to getting the recommended photos; 'Select Ref,' the time spent on selecting reference photograph; 'Wait Guidance,' the time spent on waiting composition guidance; 'Manually Adjust,' the time spent on adjusting the current scene according to the guidance; and 'After Focus,' the period from focusing (again) to taking a new picture. We can see that 'Wait Photo' and 'Manually Adjust' in Full-Guidance mode account for nearly 67% of the whole time, approximately 6 s and 3 s, respectively. This emphasizes our priorities for future work: first, optimizing the photo recommendation algorithm to reduce the waiting time; and second, making the guidance simpler and clearer to understand and learning to reduce the adjustment time.

G. Resulting Examples of the Overall Model

Our final results are shown in Fig. 15 (only the salient oval with its long axis and dominant lines are shown on the real screen; the arrows are used here to show the composition adjustment more clearly). In Fig. 15, we can see that the resulting photos (taken after guidance) are much more similar to the selected reference photos than to the original photos. We can feel that the resulting photos indeed look better, indicating the practicality and effectiveness of our method. Notably, the details of the resulting photographs may be different from those of the reference photographs, such as the subject's pose or some dominant lines. This is due to the different photography preferences of different users. We do not consider these differences to be an important issue for our results.

V. CONCLUSION

In this paper, we present a one-stop portrait photographing assistant to help novices improve their photographing skills. It is designed in a feature-based manner and is mainly composed of two steps: reference photo retrieval and photographing guidance generation. In our model, we not only utilize existing feature extraction methods but also develop new features and their extraction means. We develop a modified random forest and composition matching score to generate similar photograph batches and practical photographing guidance. Thus, users can conveniently follow the given guidance to significantly improve the aesthetic quality of their photos. However, our work still has some limitations, such as restrictions on portrait photos and the long time requirement for composition. In the future, we will continue to optimize our algorithms and test our model on high-performance GPUs.

REFERENCES

[1] Z. Zhu, H.-Z. Huang, Z.-P. Tan, K. Xu, and S.-M. Hu, "Faithful completion of images of scenic landmarks using internet images," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 8, pp. 1945–1958, Aug. 2016.

[2] S.-M. Hu, F.-L. Zhang, M. Wang, R. R. Martin, and J. Wang, "PatchNet: A patch-based image representation for interactive library-driven image editing," *ACM Trans. Graph.*, vol. 32, no. 6, 2013, Art. no. 196.

[3] Y. Liang, X. Wang, S.-H. Zhang, S.-M. Hu, and S. Liu, "PhotoRecomposer: Interactive photo recomposition by cropping," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 10, pp. 2728–2742, Oct. 2018.

[4] F.-L. Zhang, *et al.* "PlenoPatch: Patch-based plenoptic image manipulation," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 5, pp. 1561–1573, May 2017.

[5] H.-H. Su, T.-W. Chen, C.-C. Kao, W. H. Hsu, and S.-Y. Chien, "Preference-aware view recommendation system for scenic photos based on bag-of-aesthetics-preserving features," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 833–843, Jun. 2012.

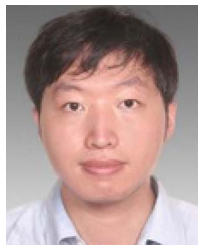
[6] J. Collier and M. Collier, *Visual Anthropology: Photography as a Research Method*. Albuquerque, New Mexico, USA: University of New Mexico, 1986.

[7] J. Park, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon, "Modeling photo composition and its application to photo re-arrangement," in *Proc. IEEE Int. Conf. Image Process.*, 2012, pp. 2741–2744.

[8] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch, "Automatic image retargeting," in *Proc. Int. Conf. Mobile Ubiquitous Multimedia*, 2005, pp. 59–68.

[9] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2006, pp. 288–301.
[10] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1930–1943, Dec. 2013.
[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[12] A. Jahanian, S. V. N. Vishwanathan, and J. P. Allebach, "Learning visual balance from large-scale datasets of aesthetically highly rated images," in *Proc. Int. Soc. Opt. Eng.*, vol. 9394, 2015.
[13] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *ACM Multimedia*, 2010, pp. 291–300.
[14] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subject," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2008, pp. 386–399.
[15] X. Fu, Y. Wang, H. Dong, W. Cui, and H. Zhang, "Visualization assessment: A machine learning approach," in *Proc. 30th IEEE Visual. Conf.*, 2019, pp. 126–130.
[16] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Proc. Int. Conf. Learn. Representations Workshop*, 2014, pp. 1–8.
[17] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 1988–1996.
[18] Y. Chen, Y. Hu, L. Zhang, P. Li, and C. Zhang, "Engineering deep representations for modeling aesthetic perception," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3092–3104, Nov. 2018.
[19] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rating image aesthetics using deep learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2021–2034, Nov. 2015.
[20] A. Campbell, V. Ciesielksi, and A. K. Qin, "Feature discovery by deep learning for aesthetic analysis of evolved abstract images," in *Proc. Int. Conf. Evol. Biologically Inspired Music, Sound, Art Des.*, 2015, pp. 27–38.
[21] W. Luo, X. Wang, and X. Tang, "Content-based photo quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2011, pp. 2206–2213.
[22] L. Zhang, M. Song, Q. Zhao, X. Liu, J. Bu, and C. Chen, "Probabilistic graphlet transfer for photo cropping," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 802–815, Feb. 2013.
[23] L. Liu, R. Chen, L. Wolf, and D. Cohen-Or, "Optimizing photo composition," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 469–478, 2010.
[24] Y. Jin, Q. Wu, and L. Liu, "Aesthetic photo composition by optimal crop-and-warp," *Comput. Graph.*, vol. 36, no. 8, pp. 955–965, 2012.
[25] F. A. Kondori, S. Yousefi, H. Li, S. Sonning, and S. Sonning, "3D head pose estimation using the kinect," in *Proc. Int. Conf. Wireless Commun. Signal Process.*, 2011, pp. 1–4.
[26] Š. Obdržálek *et al.*, "Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 1188–1193.
[27] W. Yin, T. Mei, C. W. Chen, and S. Li, "Socialized mobile photography: Learning to photograph with social context via mobile devices," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 184–200, Jan. 2014.
[28] J. Wang and M. F. Cohen, "Simultaneous matting and compositing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.
[29] B. Zhou, À. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[30] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[31] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan, "Estimating the natural illumination conditions from a single outdoor image," *Int. J. Comput. Vis.*, vol. 98, no. 2, pp. 123–145, 2012.
[32] S.-H. Zhang *et al.*, "Pose2Seg: Detection free human instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 889–898.
[33] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.
[34] L. Xu, E. Oja, and P. Kultanen, "A new curve detection method: Randomized hough transform (RHT)," *Pattern Recognit. Lett.*, vol. 11, no. 5, pp. 331–338, 1990.
[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[36] G. P. Balasubramanian, E. Saber, V. Misic, E. Peskin, and M. Shaw, "Unsupervised color image segmentation using a dynamic color gradient thresholding algorithm," in *Proc. Int. Soc. Opt. Eng.*, 2008, vol. 6806, pp. 536–544.
[37] T. Lei, X. Jia, Y. Zhang, S. Liu, H. Meng, and A. K. Nandi, "Superpixel-based fast fuzzy c-means clustering for color image segmentation," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 9, pp. 1753–1766, Sep. 2019.
[38] R. Grompone von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: A line segment detector," *Image Process. On Line*, vol. 2, pp. 35–55, 2012.
[39] E. J. Almazàn, R. Tal, Y. Qian, and J. H. Elder, "MCMLSD: A dynamic programming approach to line segment detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5854–5862.
[40] X. Lu, J. Yao, K. Li, and L. Li, "CannyLines: A parameter-free line segment detector," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 507–511.
[41] Q. Qiu and G. Sapiro, "Learning transformations for classification forests," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–9.
[42] X. Lu, Z. Lin, X. Shen, R. Měch, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 990–998.
[43] S. Ma, J. Liu, and C. W. Chen, "A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 722–731.

Nan Jiang (Member, IEEE) received the B.Sc. degree in material science from Fudan University, Shanghai, China, in 2016, and the M.Sc. degree in applied mathematics and data science from the Macau University of Science and Technology, Macau, China, in 2021. He is currently working toward the Ph.D. degree in computer science with Shanghai Jiao Tong University, Shanghai, China. His current research interests include photographing assistant, computer vision, machine learning, and pattern recognition.



Bin Sheng (Member, IEEE) received the B.A. degree in english and the B.Eng. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2004, and the M.Sc. degree in software engineering from the University of Macau, Macau, China, in 2007, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2011. He is currently a Full Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include virtual reality and computer graphics. He is an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently a Research Assistant Professor with The Hong Kong Polytechnic University, Hong Kong. He has authored or coauthored many top-tier scholarly research papers and has excellent research project reported worldwide by ACM TechNews. His current research interests include image or video stylization, colorization, artistic rendering and synthesis, and creative media.

Tong-Yee Lee (Senior Member, IEEE) received the Ph.D. degree in computer engineering from Washington State University, Pullman, WA, USA, in May 1995. He is currently a Chair Professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan. He leads the Computer Graphics Group, Visual System Laboratory, National Cheng-Kung University. His current research interests include computer graphics, non-photorealistic rendering, medical visualization, virtual reality, and media resizing. He is a member of the ACM.