Balance-Aware Grid Collage for Small Image Collections

Yu Song, Fan Tang, Weiming Dong, *Member, IEEE*, Feiyue Huang, Tong-Yee Lee, *Senior Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

Abstract—Grid collages (GClg) of small image collections are popular and useful in many applications, such as personal album management, online photo posting, and graphic design. In this article, we focus on how visual effects influence individual preferences through various arrangements of multiple images under such scenarios. A novel balance-aware metric is proposed to bridge the gap between multi-image joint presentation and visual pleasure. The metric merges psychological achievements into the field of grid collage. To capture user preference, a bonus mechanism related to a user-specified special location in the grid and uniqueness values of the subimages is integrated into the metric. An end-to-end reinforcement learning mechanism empowers the model without tedious manual annotations. Experiments demonstrate that our metric can evaluate the GClg visual balance in line with human subjective perception, and the model can generate visually pleasant GClg results, which is comparable to manual designs.

Index Terms—Grid collage, visual balance, reinforcement learning

1 Introduction

The collages of small image collections in grid view play an important role in various applications, ranging from online photo posting and personal album management to graphic design [1], [2], [3], [4]. For example, social networking services (SNSs), such as Facebook and WeChat, allow users to post a limited number of photos at a time (e.g., ten for Facebook and nine for WeChat) for sharing ideas or recording daily events, as shown in Fig. 1. Two patterns are generally used to present the posting: showing photos individually (users need to slide the posting region to view each photo) or collectively (photos are collaged together). Collective patterns, which collage multiple photos into a specific grid layout, are referred to as *grid collage* (GClg) in this study. By default, most SNSs arrange photos one by one in upload order into the grid through a fixed scheme (e.g., scan line order). However, the visual effects of different image arrangements can vary greatly. For nonprofessional users, predicting the collage results in advance and designing the arrangement sequence deliberately are difficult. Most people upload images in a random order for convenience; thus, in many cases, the results are not visually pleasant even after interactive adjustments. Collage methods are of considerable practical use in summarizing and exploring large image collections [5], [6], [7], [8], [9], [10]. The position and rotation of each image must be orchestrated. This kind of method focuses on information gathering and dissemination to help users analyze the images. Usually, layouts are automatically generated, guided by the content information of the images; however, the visual pleasantness of the overall picture is ignored. Such consideration contributes to large-scale image collection, which collages dozens of images or more. However, the goal of GClg, of which the collection volume is small scale, is different from general image collage. GClg only involves a small number of images and focuses on arranging them in a specific layout, aiming to achieve a pleasant visual appearance and draw attention from other SNS users. Moreover, SNSs usually fix the layout for online photo sharing in accordance with the number of images uploaded at a time.

Different from previous image collage works, this paper focuses on the image arrangement of GClg. The key concept in presenting an attractive GClg is to exploit a reasonable formulation in guiding the image arrangement process. The lack of visual effect analysis in existing methods motivates us to introduce a new method that enables automatic GClg of small image collections. However, automatic GClg is hindered by two major challenges. The first is how to evaluate the visual quality of GClg in which human visual perception can be embodied. Psychologists have pointed out that when given a field of several elements, people perceive them as a whole rather than individuals [11]. They have investigated the relationship between the element characteristics of the

- *Yu Song, Weiming Dong, and Changsheng Xu are with the NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100040, China. E-mail: {songyu2017, weiming.dong, changsheng.xu}@ia.ac.cn.*
- *Fan Tang is with the School of Artificial Intelligence, Jinlin University, Changchun, Jilin 130012, China. E-mail: tangfan@jlu.edu.cn.*
- *Feiyue Huang is with the Youtu Lab, Tencent, Shanghai 200233, China. E-mail: garyhuang@tencent.com.*
- *Tong-Yee Lee is with the National Cheng Kung University, Tainan 701, Taiwan. E-mail: tonylee@mail.ncku.edu.tw.*

Manuscript received 22 Oct. 2020; revised 9 Sept. 2021; accepted 12 Sept. 2021. Date of publication 16 Sept. 2021; date of current version 30 Dec. 2022.
This work was supported by National Key R&D Program of China under Grant 2020AAA0106200, in part by National Natural Science Foundation of China under Grants 61832016, U20B2070, and 6210070958, in part by Ministry of Science and Technology under Grant 110-2221-E-006-135-MY3, Taiwan, and in part by Open Projects Program of National Laboratory of Pattern Recognition.
(Corresponding author: Weiming Dong.)
Recommended for acceptance by W. Wang.
Digital Object Identifier no. 10.1109/TVCG.2021.3113031

Fig. 1. Grid collage of photos are popular on SNSs.

image and the image liking rate. Meaningful conclusions (e.g., the visual appreciation of a picture largely depends on the perceptual balance of its elements [12], and a positive relation exists between balance and liking for a multielement picture [13]) have been drawn. To measure the character of balance, most psychological studies refer to the assessment of preference for balance (APB) [14] and deviation of the center of mass (DCM) [12]. However, these measures cannot provide sufficient guidance, such as feature selection and semantic extraction, which are important for evaluating the GClg template to achieve balance perception. The second challenge is how to design a reasonable GClg arrangement strategy. The existing online photo posting services of social networks and album management apps usually arrange images into a grid randomly or following the uploading order. Thus, users need to adjust the positions of some images to improve the appearance. However, the adjustment process is tedious and difficult for nonprofessional users. Thus, automatically creating GClg in a small image collection can lead to visually pleasant and informative photos; such automation is meaningful and valuable for many applications.

To address the first challenge, a balance-aware metric based on APB and DCM was designed. By refining the visual and semantic information of input images, the engaged metric was formulated by combining the image features of color shade, color tint, content, and object size. To address the second challenge, an end-to-end reinforcement learning method was developed, which optimizes the arrangement step by step (see Fig. 2). The position exchange of two images was defined as the action and the balance-aware metric as the environmental reward was used. Two additional penalty items are added to accelerate training, i.e., action search penalty and step penalty. The images' input order of a collection is regarded as the initial arrangement, as guided by the metric. The final GClg can become pleasant through multistep interaction with the former arrangement.

The major contributions of this work are as follows:

- The concept of GClg, which is a fresh perspective of the multi-image joint presentation issue, is presented for the first time.
- A novel psychology-based balance-aware metric is proposed to evaluate the visual quality of GClg.
- An end-to-end reinforcement learning model is built to arrange images automatically into a specific grid template to form a visually pleasant GClg.

2 Related Work

This work intersects several previous research lines: single-image presentation assessment, user preference behavior, image collage, multi-image presentation assessment, and reinforcement learning.

Single-Image Presentation Assessment. Aesthetics assessment is a typical problem in evaluating the visual effect of single-image presentation. Murray *et al.* [15] proposed an aesthetics dataset with rich annotations; the dataset contains not only aesthetic scores but also categories and photographic styles. Kong *et al.* [16] modeled the relative ranking of photo aesthetics in the loss and incorporated image information to regulate the rating result. The current methods are mainly aggregation-based methods. Lu *et al.* [17] proposed an aggregation network that uses different patches captured from source images to obtain fine-grained details of images for aesthetic evaluation. Ma *et al.* [18] proposed a novel image patch selection strategy that can deeply understand the image features and combine them with the overall features. Lu *et al.* [19] added style and semantic attributes to accelerate the process of aesthetic estimation. Wang *et al.* [20] jointly trained the search task of the image attention box and aesthetics classification task, which can benefit aesthetics tasks by the interrelationship between the two tasks. Mail *et al.* [21] aggregated original pixel information by using an adaptive spatial pooling strategy to contribute to aesthetic assessment. Sheng *et al.* [22] proposed the Gourmet Photography Dataset (GPD) and realized a nonstationary regularization method to assess the visual aesthetics of food images. However, the above methods are mainly designed for a single image [23], [24], [25], [26], which limits their application to evaluate multi-image presentation.

User Preference Behavior. The explosion of social media has led to research on user preference behavior, which investigates the relationship between visual effects and user reactions. The image liking rate is a mechanism of SNSs to express user affection, which can be used as a bridge between image visual effect assessment and user preference [27]. The higher the liking rate is, the more attractive

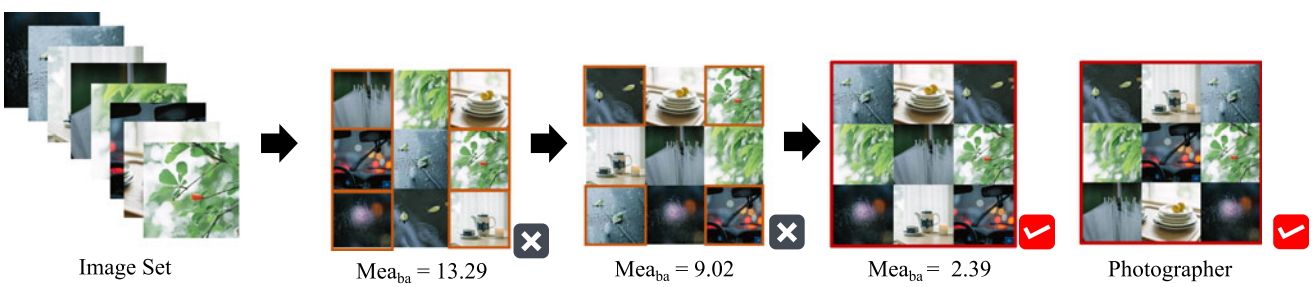


Fig. 2. System workflow. The GClg network optimizes the arrangement of the input image set step by step. The final result generated by our method is visual pleasure and comparable to arrangement by photographers.

and visually pleasant the result is. In psychology, several studies have focused on predicting the liking rate. Wilson and Chatterjee [14] introduced the APB and reported a method to derive an objective balance score. Hübner and Fillinger [12] presented DCM, which represents the center of perceptual "mass" in a picture and its deviation from the geometric center. Comparisons showed that DCM scores account better for balance ratings than APB scores, whereas the opposite is true with respect to preference [12]. Studies have also shown that the fusion of balanced features and other unique image features is effective in measuring evolution. Thömmes and Hübner [28] showed that SNS users prefer visually balanced photos and combined balance and curvature as a metric of architectural photos. Thömmes and Hübner [27] considered content (gender information) and balance in dance photos. Hübner and Fillinger [13] argued that images composed of several objects become popular with a strong sense of balance.

In this paper, we focus on the generation of GClg, which is a multi-image presentation problem and is widely used by SNSs. In terms of quality evaluation, GClg has a different scene than the aesthetics assessment. Aesthetics assessment judges the quality of single and complete images, whereas layout focuses on the element's location design. In our work, we creatively transform the custom psychology metric into a computer graphics area. Combining the basic balance concept in psychology and analysis of content and other characters, we formulate a comprehensive metric and devote a solution for GClg. Our method can generate visually pleasant GClg results, which are comparable to those of photographers.

Image Collage. Image collage is a common technique for multi-image presentation. The traditional goal is to highlight the salient information of an image and simplify the browsing process. Tan *et al.* [7] used a graph-based algorithm to build a global arrangement and then applied online Voronoi tessellation to refine the final layout. Cao *et al.* [29] automatically generated a stylized comic layout from a group of input artworks with user-specified semantics. Cao *et al.* [30] further synthesized comic elements to guide the reader's attention by using a probabilistic graphical model. Yu *et al.* [31] used a circle to approximate the salient area of each image and then formulated the photo collage as a circle packing problem. Han *et al.* [32] proposed a tree-based collage approach, which can generate a collage with several interesting shapes. Liu *et al.* [8] presented an irregular shape-based canvas partition method to generate compact collage. Liang *et al.* [33] recomposed image collection on the basis of sample photos and generated a layout by using a Voronoi tree map. Wu *et al.* [34] proposed a binary tree-based page division and adjustment algorithm and preserved the image correlations. Zheng *et al.* [35] synthesized layout design on the basis of the visual and textual semantics of user inputs with a deep generative model. Pan *et al.* [10] used aesthetic and content features to summarize an image collection and then visualized the subset images via tree-based layout generation. Gan *et al.* [3] managed albums in a comic-like layout on the basis of image classification. With the popularity of mobile devices, Kong *et al.* [16] proposed a strategy to collage phone images into a centroidal Voronoi diagram. Song *et al.* [36] constructed multipage photo collage for image collection and its formulation as an issue of joint optimization.

Multi-Image Presentation Assessment. Multi-image presentation assessment focuses on the evaluation of the organization of multiple images and other visual elements. Layout presentation assessment is an important brand of multi-image presentation. Locher *et al.* [37] revealed that balance is an important factor in creating visual displays. Geigel *et al.* [38] combined several distribution characteristics, including balance, emphasis, chronology, and unity, to measure layout quality. Lok *et al.* [39] introduced the concept of WeightMap and encoded the visual weight for an entire layout to evaluate its effectiveness. Yang *et al.* [4] combined information on aesthetic principles with that on image features to design an attractive visual–textual presentation layout. However, the core task for layout research is to determine the center point and size of components.

However, all image collage and multi-image presentation assessment strategies above work on image sets of tens or even hundreds of volumes, and the optimized variables are always the images' position and rotation, as well as scaling ratio. The goal of general collage is to obtain a visual result with compact information and a clear outline. For online GClg, the image set is usually small in size, and its goal is to obtain a comfortable and attractive visual effect, which is different from general collage. In this study, we explore the online GClg formula in which human visual perception can be encoded.

Reinforcement Learning. Reinforcement learning is a distinct strategy that has emerged in recent years. In contrast to the label requirement of supervised learning and the rough label of unsupervised learning, reinforcement learning can complete learning through rewards. Therefore, it can work without labels. It has contributed to a few research fields. In Atari games that defeat human beings, reinforcement has received considerable attention [40]. Zhou *et al.* [41] rewarded the representativeness and diversity of the summary to obtain video skimming. Clegg *et al.* [42] used reinforcement learning to animate the process of dressing. Satoshi *et al.* [43] achieved image enhancement provided by unpaired images through reinforcement. In this study, we introduce reinforcement learning into a training strategy. Combining the online GClg formulation as the reward, a solution without any annotation can be achieved.

3 GClg of Small Image Collection

In contrast to the collage of large-scale image sets, the automatic collage generation of limited images needs further research. Taking online photo sharing on SNSs as an example, a collage is a typical small image collection presentation scene. Mainstream SNSs, such as Facebook and Twitter, allow users to upload and publish images online and then generate a poster to show these images. By observing how mainstream SNSs generate such posters, we can summarize the differences between the generation of small and large image collection collages as follows:

1) *Limited number.* The number of images is the direct difference between the two collage types. For a small image collection, the maximum number of images

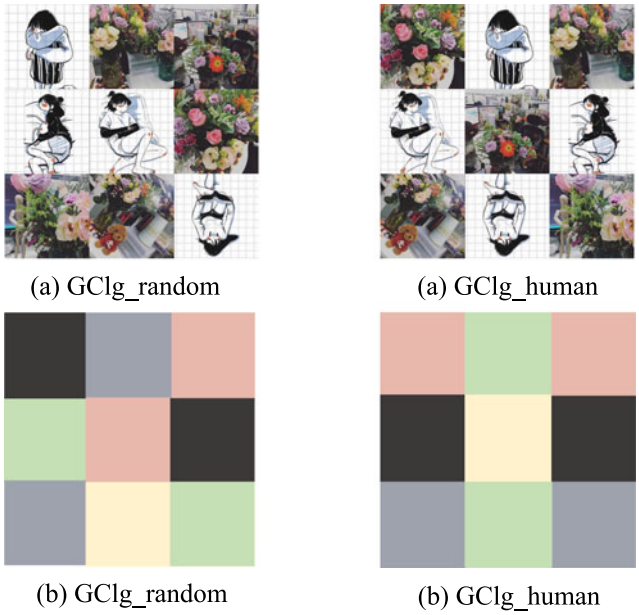


Fig. 3. Examples of GClg in user study. (a) and (b) are GGlg of two image sets, GClg_random and GClg_human are the results generated by random arrangement and human arrangement.

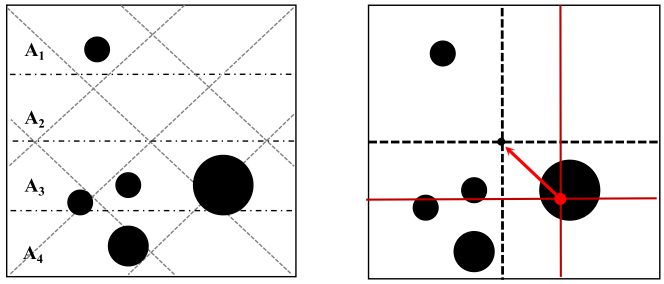
displayed together is usually limited. For Facebook, users can only upload images with no more than ten at a time, whereas only a maximum of four can be uploaded on Twitter.

2) *Fixed layout*. When generating posters, the layout is usually fixed by the SNS. Although different platforms have different layout configurations, most collage layouts are fixed in grid view to generate simple and compact collages.

In this study, a small image collection in a specific grid layout is referred as a GClg. A potential way to make the posters win more "likes" could be by rearranging images in GClg. To determine how the arrangements of the image positions in the GClg influence human behavior on SNSs, we conducted a user survey involving 181 individuals (100 females and 81 males, age ranging from 20 to 45). Approximately 70.72% of the participants used to post multiple images on SNSs, and over 80.11% of them claimed to have arranged photos to improve their visual experience.

Furthermore, a total of 22 sets of images were collected; 9 images were contained in each set. The images were arranged into a nine-grid (3×3) view per set. Participants are asked to choose which arrangement is better, or the two arrangements seem nearly. Fig. 3 shows the GClgs of two different image collections; for Fig. 3a, 73.48% of participants chose GClg_human, and for Fig. 3b, 68.51% of participants chose GClg_human. Overall, the study collected 3982 votes, of which 74.26% resulted in GClg_human and 9.97% GClg_random, and the rest showed no difference.

In summary, the survey shows that different arrangements lead to different visual effects, and a consensus on preferences occurs among audiences. Moreover, almost all the participants expressed their hope for the development of an automatic mechanism that can arrange the release of photos. From the chosen GClg, it was concluded that the users prefer the result with a "balanced" character. In this paper, a new view of the "balance" concept was proposed with novel quality evaluation metrics for our online GClg.



(a) APB calculation diagram in the horizontal direction (b) DCM calculation diagram

Fig. 4. Basic visual balance calculation diagram.

4 Balanced GClg Formulation

Researchers have found that users prefer results with "balanced" characteristics [12], [28]. In this study, we summarize research on "balance" in different fields and propose a "balance" metric that is tailored to online GClg. Online GClg is a composition of different individuals, for which new insight should be taken.

4.1 Basic Concept of Visual Balance

APB [14] and DCM [12] are the two basic balance concepts in psychology. Each of them can be calculated into a score, which can reflect the balance characteristics of an image. For *APB*, the source image needs to be evenly divided into four parts by region. As shown in Fig. 4a, the horizontal axis was used as an example. The variables, A_1, A_2, A_3, A_4 represent four parts, and $f(A_1), f(A_2), f(A_3), f(A_4)$ are the sums of gray values for each part. The sum of the entire image N can be denoted as $N = f(A_1) + f(A_2) + f(A_3) + f(A_4)$. Then, for the horizontal axis, two terms must be computed. The first term is defined as

$$h = \left(\frac{|[f(A_1) + f(A_2)] - [f(A_3) + f(A_4)]|}{N} \right) \cdot 100. \tag{1}$$

The second term is defined as

$$h_{io} = \left(\frac{|[f(A_1) + f(A_4)] - [f(A_2) + f(A_3)]|}{N} \right) \cdot 100. \tag{2}$$

All the above process steps are aimed at the horizontal axis, and three other axes (vertical and two diagonal axes) must be treated similarly. The final APB is then calculated as

$$APB = \frac{\sum_{axe_i=1}^{4} h_{axe_i} + h_{io\_axe_i}}{8}. \tag{3}$$

DCM calculates the deviation between the "mass" and the geometric center of the image. This concept introduces the physical explanation of images; that is, darker colors look heavier visually. As shown in Fig. 4b, the black circle in the lower right corner looks "heavier" than the rest in the image. The mass point of the entire image is likely to be inclined to the lower right corner. The intersection of red solid lines is the center of mass of the image, and the intersection of black dashed lines is the geometric center of the

image. The red arrow indicates the deviation between the mass points of images and the geometric center. Compared with APB, DCM focuses on the entire image instead of analyzing the subregions. The b_x and b_y are denoted as "mass" centers in horizontal and vertical dimensions respectively, and calculate them by using the following formula:

$$b_x = \frac{\sum_{i=1}^{W} h_i^x \cdot r_i^x}{\sum_{i=1}^{W} h_i^x}, b_y = \frac{\sum_{i=1}^{H} h_j^y \cdot r_j^y}{\sum_{i=1}^{H} h_j^y}, \tag{4}$$

where W is the canvas width, H is the canvas height, h_i^x is the pixel gray value in the i_{th} column, h_j^y is the pixel gray value in the j_{th} row, and r is the coordinate value. Then, the DCM of a picture can be calculated as

$$DCM = \left(\frac{\sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}}{0.5} \right) \cdot 100, \tag{5}$$

where $\mathrm{d}x = 0.5 - b_x/W$, $\mathrm{d}y = 0.5 - b_y/H$.

In summary, APB and DCM focus on the image's color shade (APB focuses on the local view, whereas DCM focuses on the global view) and thus can only act as a basic balance statement but hardly provides a comprehensive description of GClg. To solve this problem, we combine APB and DCM and add additional information to the balanced formulation of GClg.

4.2 Image Visual Balance for Online GClg

The main difference between GClg and single images is that GClg consists of multiple images, which are usually regarded as a whole. In this study, a metric called "image visual balance" was proposed to judge the balance of GClg. Four items are expanded to basic balance concepts, namely, color shade, color tint, image content, and objective size. The addition of new items introduces additional information, thus making the quality evaluation of GClg reasonable. Moreover, a bonus term is also added, giving a window to cope with the users' personal preference. Generally, five items are used to calculate image visual balance: color shade, color tint, image content, object size, and bonus.

Color Shade Balance. Color shade sets the fundamental tone of the entire GClg. As mentioned earlier, APB and DCM emphasize judging the color shade. APB can estimate the color shade's local distribution, whereas DCM focuses on the whole visual effect. A linear combination of APB and DCM was used to express the degree of color balance. The balance of color shade can be calculated as

$$bal\_col = DCM + APB. \tag{6}$$

Color Tint Balance. Tint distribution is an important feature of an image. The color histogram is used to describe its tint characteristics in RGB space as the feature of image color tint. The Bhattacharyya distance [44] is used to measure the distance between two histograms. Tint balance ba_tint between images is designed to determine whether the paired images in a symmetrical position of GClg have the same or similar color distribution. The basic form of ba_tint can be calculated as (using the x-axis direction as an example)

$$ba\_tint = \frac{1}{M} \cdot \sum_{i=1}^{M} \left| tint(I_i) - tint(I_i^*) \right|, \tag{7}$$

where M is the number of image pairs in symmetrical positions, and $tint(I_i)$ and $tint(I_i^*)$ are the tint features of the paired images.

Content Balance. Content is the core information carrier of an image. Content differences seriously affect the visual effect of GClg. The content balance ba_con between different images is used to judge whether the image content in a symmetrical position in a picture has similar characteristics. The image features were extracted by using VGG-16[45] pretrained on ImageNet. The output of the last layer is taken as the content feature of each image. Then, the content distance between two images can be calculated as

$$ba\_con = \frac{1}{M} \cdot \sum_{i=1}^{M} \left| fea(I_i) - fea(I_i^*) \right|, \tag{8}$$

where M is the number of image pairs in symmetrical positions, and $fea(I_i)$ and $fea(I_i^*)$ are the content features of paired images.

Size Balance. Salient object size is another factor that has a great influence on the collage result. That is, when the size difference of image objects in symmetrical positions is too large, it will lead to poor visual experience. Size balance ba_size among different images is used to judge the similar degree of the symmetrical position in GClg. The binarized saliency map was employed [46] to judge the object size. The more similar object sizes the images in symmetrical positions have, the more comfortable the result is. The images with similar salient object sizes are expected to be laid at symmetrical positions and to form a size balance. Thus, ba_size can be calculated as

$$ba\_size = \frac{1}{M} \cdot \sum_{i=1}^{M} \left| SA(I_i) - SA(I_i^*) \right|, \tag{9}$$

where M is the number of image pairs in symmetrical positions, and $SA(I_i)$ and $SA(I_i^*)$ are the number of black pixels of the corresponding salient maps of paired images.

Bonus. For GClg, in addition to the image information itself, user preference is another factor that cannot be ignored when evaluating the entire GClg. *Special location* and *uniqueness value* are two issues that can show one's preferences. Special location can explain the preference of grid view position, and it may occupy the middle position or large area. In our framework, the special location of a grid template is specified by the user. Uniqueness value can explain the preference for subimages, to which users can score images in their way to express the importance among all images. To define uniqueness concisely and fairly, we calculate it automatically by judging the deviation between information and other information. Information for each image includes the color shade, color tint, and content information. The calculation is as follows:

$$UQ_{dep_i} = \frac{\sum_{j \neq i} \left| dep_j - dep_i \right|}{\sum_{j=0}^{j=T} \left| dep_j \right|}, \tag{10}$$

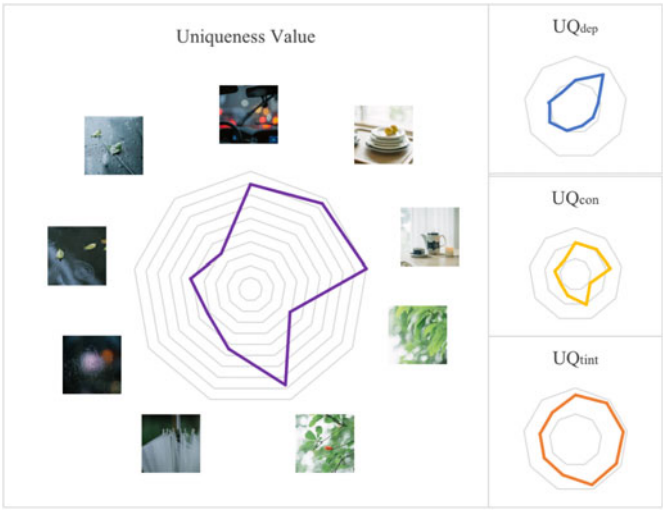


Fig. 5. Uniqueness value distribution of the image set. The left part is the distribution of the comprehensive uniqueness value, and the right part is the three subdistributions of UQ_{dep_i}, UQ_{con_i}, and UQ_{tint_i}, respectively.

$$UQ_{tint_i} = \frac{\sum_{j \neq i} |tint_j - tint_i|}{\sum_{j=0}^{j=T} |tint_j|}, \tag{11}$$

$$UQ_{con_i} = \frac{\sum_{j \neq i} |con_j - con_i|}{\sum_{j=0}^{j=T} |con_j|}, \tag{12}$$

$$UQ_i = UQ_{dep_i} + UQ_{tint_i} + UQ_{con_i}, \tag{13}$$

where T is the image set volume, dep_i, $tint_i$, and con_i are the shade, tint, and content information of image i, respectively. Fig. 5 shows an example of the uniqueness value distribution of an image set. To view from human perception, the image set can be divided into bright and dark color dominated images. The bright images can be further divided into images with obvious objects and images with unobvious objects. To view from the calculation value, the subdistributions of UQ_{dep_i}, UQ_{con_i} and UQ_{tint_i} are illustrated in the right corner. For UQ_{dep_i}, the first image in the upper right corner has the maximum value for the obvious color distribution difference, which is consistent with the real situation; i.e., the top half of the image is the blank space, but the bottom half of the image is full of shadows. For UQ_{con_i}, the second image in the upper right corner has the maximum value, which is consistent with the real situation that this image has the most obvious object; however, for UQ_{tint_i}, the value among images does not seem to vary much but the images with bright colors obtain higher values, which is consistent with the fact that images with bright color are less than images with dark color. Combining these three parts, we can obtain the final "uniqueness value", and the second image in the upper right corner is the most unique image. Thus, when an image with the maximum unique value is placed in the special position, we set $Bonus = 0.8$, otherwise, $Bonus = 1.0$.

Final Metric. In summary, all the above items are combined and a new metric to evaluate the balance property is proposed, in which dep_i and $tint_i$ are the two major items and the impact of the other items depends on the concrete circumstances of the image set. In particular, this paper normalizes the difference value of each modeling subitem. If T images $(I_1 \ldots, I_T)$ are presented in a set for GClg, then the final balance measurement and objective equation are as follows:

$$Mea_{ba} = ba\_col \cdot ba\_tint \cdot ba\_con \cdot ba\_size \cdot Bonus, \tag{14}$$

$$\arg\min Mea_{ba}(I_1, \ldots, I_T). \tag{15}$$

The visual balance formulation expands the meaning of the basic metric of visual balance, which engages additional views to describe the entire GClg.

5 Approach

5.1 Overview

This study proposes an end-to-end reinforcement network to arrange a collection of images step by step. We choose a reinforcement learning strategy for three main reasons. First, the result of image arrangement is a discrete variable, which is difficult to solve by traditional deep learning methods. Second, reinforcement learning can train the network without labels, which can reduce the requirement of data annotation. Third, reinforcement learning can train the network by multiple interactions, which can reduce the requirement of data amount. This study recasts the issue of discrete image arrangement to the issue of action space search, which can be merged into reinforcement learning.

The images in a collection are sent to the network in conjunction and randomly initialize their arrangement in the grid. The feature of a GClg is a combination of its VGG-16 [45] content information, size information, and tint information, which is 4105 dimensions in total. The overall structure of the GClg generation network is shown in Fig. 6 and can be divided into two parts: feature extraction and action prediction. The feature extraction part is made of one fully connected layer, which fuses image information itself to reduce input feature dimensions. The input images are sent to the feature extraction part and processed as a Siamese architecture [47]. The output of the feature extraction part is 9×512 dimensions and is then sent into the action prediction part. The action prediction part consists of three fully connected layers, and the top layer is equipped with the sigmoid function. The number of neurons in the last fully connected layer is the volume of action space.

5.2 Action Choice

In essence, GClg generation is a problem of image arrangement. With an initial input, we define the action as the position interchange of two images ($[I_i, I_j]$). For example, if N images are present in a group, the initial order is $I_1 \ldots, I_i \ldots, I_j \ldots, I_N$, and the action occurs on I_i and I_j, then the arrangement after this action becomes $I_1 \ldots, I_j \ldots, I_i \ldots, I_N$. Obviously, there are C_N^2 exchange actions in the action space, and we add a terminal action additionally. Therefore, for the image action space, the total volume is $C_N^2 + 1$. For image action choice, we formulate the output of action prediction as a *multinomial* distributed vector

$$a_i \sim multinomial(p_i), \tag{16}$$

where p_i is the probability sequence vector at step i; after the multinomial sampling process, it becomes an action index. Then, the new image arrangement is decided by the image exchange according to the index.

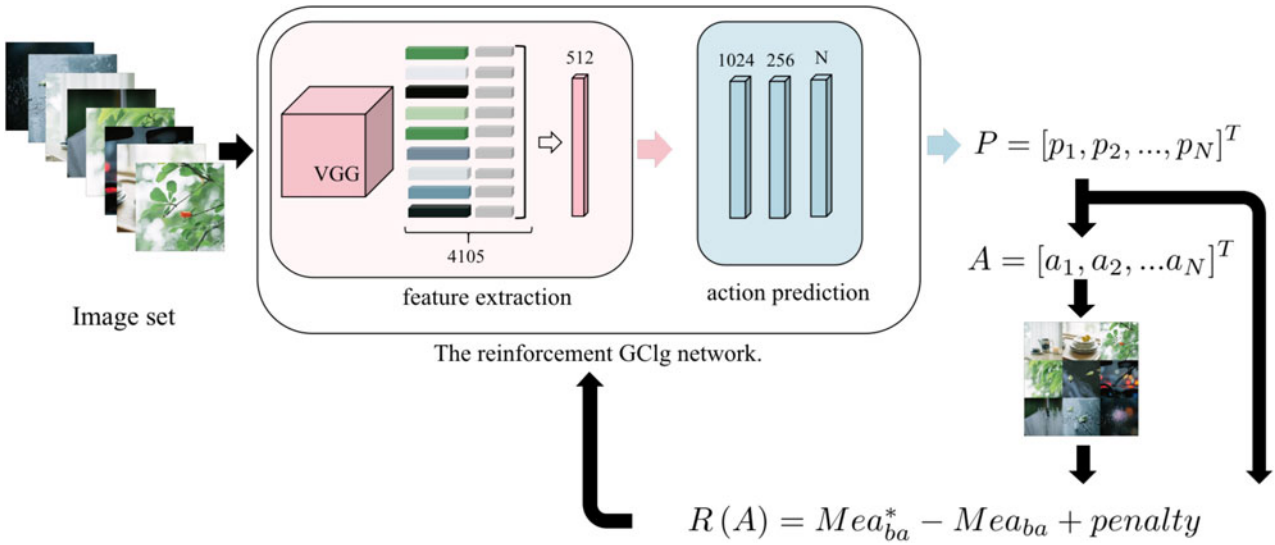


Fig. 6. Structure of the reinforcement GClg generation network. The network takes all the images as input and will output an action probability distribution P according to the image features and the current image arrangement order. An image exchange action will be taken according to multinomial distribution sampling based on P. The reward will be calculated by the variance between the balance measurement before action taken (Mea^*_{ba}) and balance measurement after action taken (Mea_{ba}). A penalty item is also added to the reward function for an effective and fast search. Then, the images in new arrangement order will be sent to the network again until the terminal action or the max step.

Ideally, the GClg generation network can provide an effective sequence of actions, i.e., under the premise of ensuring effective image rearrangement, the network can be implemented with fewer steps and fewer adjacent repeated actions (action search penalty). To address this problem, we add a *penalty term* to the reward to help the network quickly eliminate this state. Details are described in the next section.

5.3 Reward Function

According to the action mode described above, an arrangement of image collection corresponds to a GClg result. The balance metric described in Section 4 naturally has the judging attribute for GClg. For stable training, a $sign$ function is utilized to demonstrate the effectiveness of an action. The smaller the measurement is, the more balanced the GClg becomes. Thus, the action reward function can be expressed as

$$r(a) = sign(Mea^*_{ba} - Mea_{ba}), \tag{17}$$

where Mea^*_{ba} is the measurement before the action and Mea_{ba} is the measurement after the action. In addition, the final $Reward$ function can be expressed as

$$R(A) = \sum_{i=0}^{T} r(a_i), \tag{18}$$

where A is the action series in the episode, T is the total step number and $r(a_i)$ is the action reward at step i .

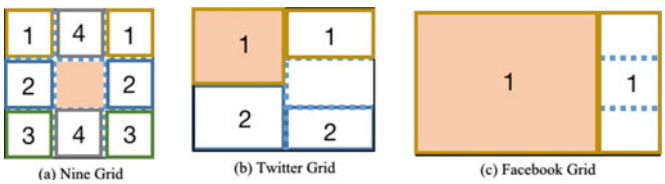


Fig. 7. Three typical GClg templates.

5.4 Penalty Term

Two kinds of penalty items (action search penalty and step penalty) are added to push the GClg generation network an efficient search. The action search penalty item (pt_{search}) is designed to address the situation in which the network takes the same action twice, which makes the current image arrangement return to the arrangement two steps ago. The step penalty aims to achieve the final arrangement in fewer steps. The item is defined as

$$pt_{search} = \begin{cases} -0.1 & \text{two consecutive identical actions,} \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

The step penalty pt_{step} is added to push the network to search for additional actions and achieve the final result

$$pt_{step} = \lambda_2 * i, \tag{20}$$

where i is the current step number in reinforcement learning. The variable is set to $\lambda_2 = -0.01$.

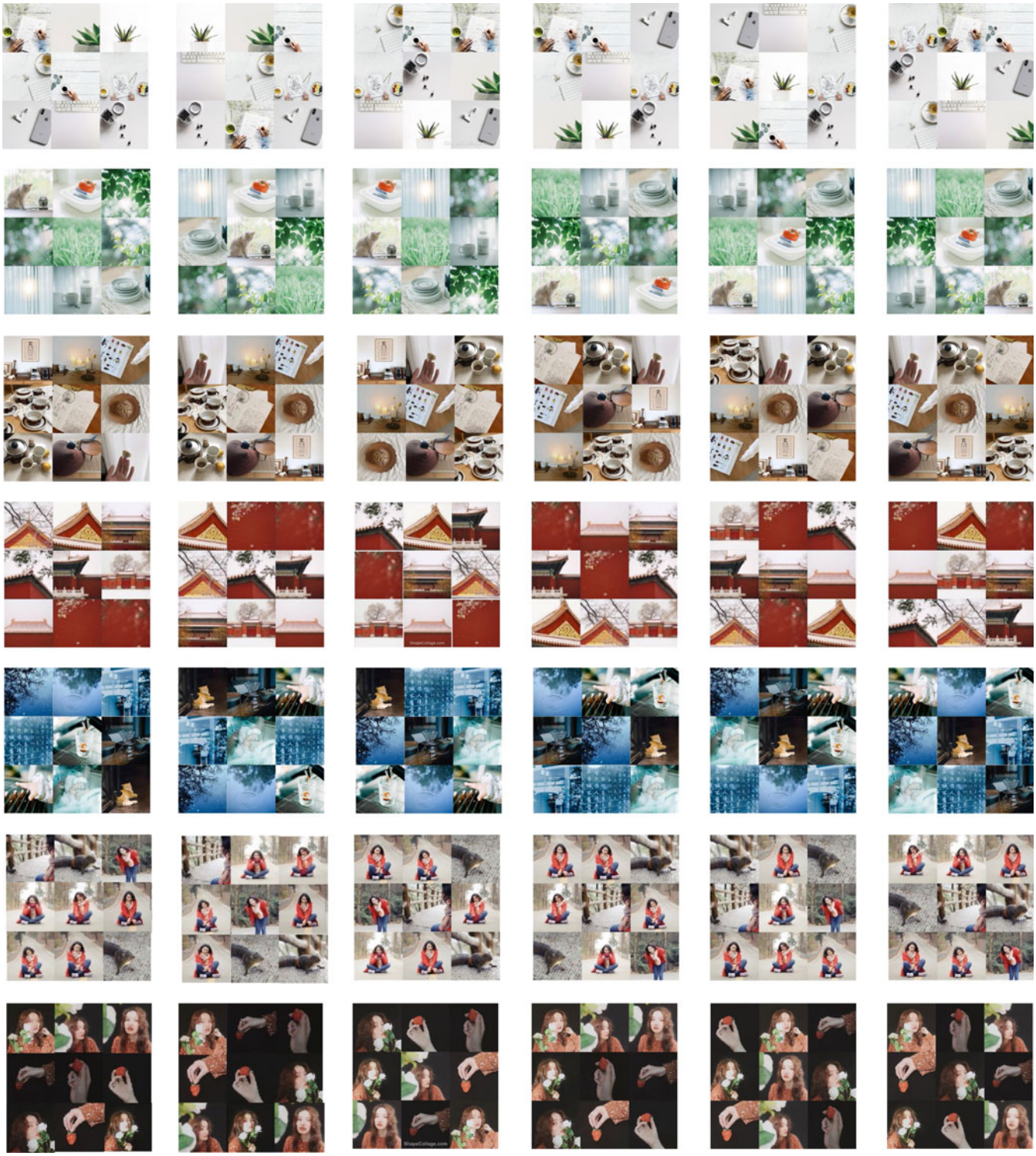
5.5 Optimization

The GClg generation network is used to determine a policy that minimizes the reward function in Section 5.3

$$J(\theta) = \mathbb{E}_{p_\theta(a_i:T)}[R(A)], \tag{21}$$

where $p_\theta(a_i : T)$ is the probability distribution of action choice. The deviation can be calculated as

$$\nabla_\theta J(\theta) = \mathbb{E}_{p_\theta(a_i:T)}[R(A) \sum_{i=0}^{T} \nabla_\theta log \pi_\theta(a_i|h_i))], \tag{22}$$



(a) Baseline 1 (b) Baseline 2 (c) ShapeCollage [46] (d) CollageIt [45] (e) Ours (f) Human

Fig. 8. Comparison of our results with those of other methods.

which can be processed for brevity as

$$\bigtriangledown_\theta J(\theta) \approx \frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{T}[(R_j - b)\bigtriangledown_\theta log\pi_\theta(a_i|h_i))], \tag{23}$$

where K is the total episode number, T is the total step number, R_n is the reward at episode j, $\pi(\theta)$ is the policy of our GClg generation network, and b is the average value of the experienced reward to make the optimization efficient. The variables $K = 5$ and $T = 10$ are set as default.

6 Implementation and Experiments

6.1 Implementation Details

Grid View Configuration. In this study, a 3×3 grid is used as the template. As shown in Fig. 7a, four symmetrical positions are set (blocks with the same color), and the central

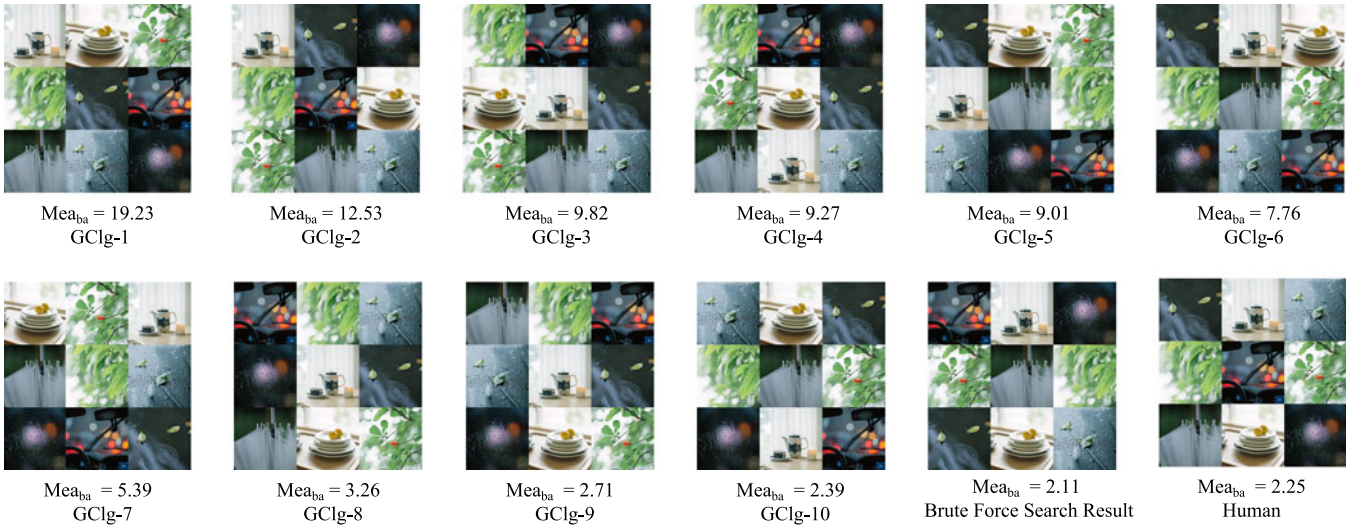


Fig. 9. Different GClg arrangements with balance values.

block (light orange) is chosen as the special location. For convenience, all images in a collection are resized to 300×300, and the GClg is led to $(300 \times 3) \times (300 \times 3)$.

Dataset. We collect 300 GClgs on the basis of the two following principles. 1) *Comprehensiveness*. Ninety percent of all GClgs are obtained from popular SNSs, such as Instagram, WeiBo[1] and TuChong;[2] the rest are composed of blocks in different pure colors. The positions of pure color blocks in GClgs are manually arranged. GClgs collected from websites involve rich image themes, whereas pure images can keep the basic consideration of color, thus reducing the negative effects of noise and ambiguities of website images. 2) *Popularity*. When collecting GClgs from SNSs, only the most popular and recommended postings were collected. Furthermore, each GClg was split into separate images.

Training Configuration. The model was trained by using 80% of the data as the training set and the rest for evaluation. Adam acts as the optimization algorithm of the back-propagation gradient, and the learning rate is 1×10^{-5}. The maximum number of steps in each episode is set as 10.

6.2 Experiments

Baselines. Two baseline methods were constructed based on information aggregation, which is often used by traditional image collage strategies. First, the images were clustered into different classes by K-means according to content features (*Baseline 1*) or color tint features (*Baseline 2*). The variable was set as $K = 3$ in all the experiments. Then, the images are arranged to the grid by class and the images of the same class are arranged in order.

Qualitative Results. The generated results were compared by the method in this study with those produced by baselines and the two commercial software programs (i.e., CollageIt [48] and Shape Collage [49]), as shown in Fig. 8. We can see that the baselines and commercial software can hardly ensure a reasonable GClg, and our method can lead to visually pleasant results, which are comparable to human manually designed results. Specifically, there is usually more than one type of "balance" for an image set, and our method may generate results in a different balance form from humans, while the final result is also appealing and visually pleasant.

Metric and Variation Analysis. The image set in Fig. 9 contains 10 different arrangements. In these results, the measured values vary. Usually, the one with a smaller measured value looks more balanced and visually pleasant, and GClgs with similar measured values have similar visual balance characteristics. A brute force search result is also shown in Fig. 9. Although this result has a smaller measurement value than the result (GClg-10), it costs several hours to compute. However, this method can achieve a good result that is comparable to the brute force search result and the human result in a short time (3–7 seconds). Table 1 and Fig. 10 report the measured value variations of different combinations of components. In Fig. 10, when the size item is removed, GClg-2 should be more visually pleasant than most of the others. However, it contradicts the truth, i.e., GClg-2 is visually unclear. Something similar happens in measurement removed tint item. When the omitted item is tint, GClg-6 is considered more unbalanced than most of the others (e.g., GClg-3, GClg-4, GClg-5). In summary, for the example illustrations in Fig. 9, when items are removed from the measurement, confusion tends to occur during GClg evaluation. Fig. 11 visualizes the

TABLE 1
Comparison of Different Metric

GClg set	Mea_{ba}	Mea_{ba} w/o tint	Mea_{ba} w/o content	Mea_{ba} w/o size	Mea_{ba} w/o bonus
GClg-1	19.23	22.42	16.61	28.39	19.23
GClg-2	12.53	14.63	10.44	18.91	12.53
GClg-3	9.82	10.93	8.35	21.98	12.29
GClg-4	9.27	10.25	8.41	22.31	9.27
GClg-5	9.01	10.66	7.37	17.16	9.02
GClg-6	7.76	11.17	6.71	15.41	7.76
GClg-7	5.39	7.31	4.77	24.40	5.39
GClg-8	3.23	4.04	2.65	11.87	4.03
GClg-9	2.71	3.29	2.31	6.11	3.39
GClg-10	2.39	4.35	2.20	5.90	2.39

1. https://www.weibo.com
2. https://tuchong.com/community

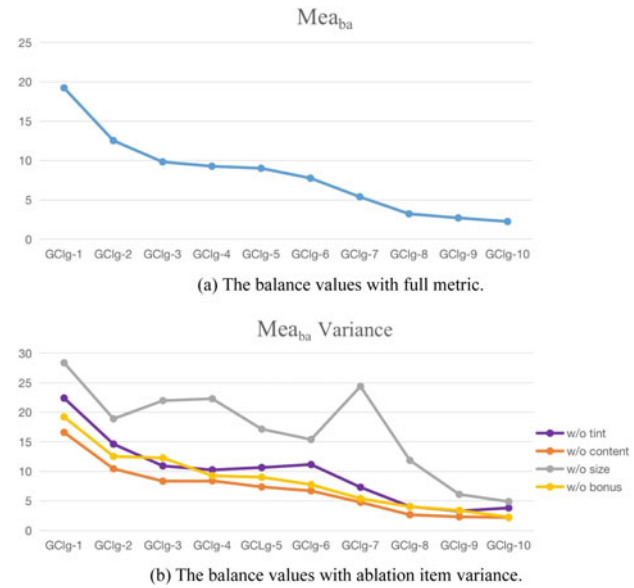


Fig. 10. Illustration of balance value variance in Fig. 9.

items involved in the measurement formulation. It is easy to see how those items work to optimize the visual balance of the GClg. For example, APB and DCM jointly represent the color shade balance. Histogram represents the color tint balance.

Ablation Study. In addition to the measurement change analysis experiment, we also conducted an ablation study. GClg is generated directly, guided by Mea_{ba}, as well as Mea_{ba} with a subitem missing. The results are shown in Fig. 12. Overall, the comprehensive Mea_{ba} has superior capability to generate fresh and balanced GClg results, and any missing item could lead to chaos not only for the item itself but also for the other items.

User Study. The work was evaluated via a two-step user study. The study involved 50 participants, 25 males and 25 females, who had different backgrounds. Thirty groups of new image sets were collected; these images were released by photographers in accordance with the designed arrangement. At the beginning, participants were shown five different GClg results for an exact group: random result, Baseline 1 result, Baseline 2 result, our result, and human result. Considering the massive amount of GClg information, we first ask them to choose the two best GClgs from each group. Then, after a break, we start the second round by only showing our result and human result each time. At this time, the participants could only choose one result.

In Table 2 (the votes in step 2 are enclosed in brackets), it can be observed that the method proposed in this study is superior to both the random method and the baseline methods and has votes close to the human.

6.3 Extension to More Templates

In this study, we mainly use 3×3 as a general template and formulate the GClg evaluation metric based on this template. However, this template can be easily extended into other templates. Twitter and Facebook templates, as the most common templates in our daily life, have also been tested. The configurations of Twitter and Facebook are shown in Figs. 7b and 7c. We set one image balance pair for Facebook template and two image balance pairs for Twitter template, and the special locations are annotated in light orange. The uniqueness value is calculated automatically as before.

Figs. 13 and 14 show different GClgs with their balance values. The balance value illustrates a GClg's ability to express visual sensation. Moreover, we show the results with different uniform and nonuniform templates in Figs. 15 and 16. We can see that our balance metric and image arrangement algorithm both have good generality to different templates.

The Twitter and Facebook templates can also be customized by changing the blocks' configuration. In Figs. 17 and 18, we manually change the special location and special image of the basic template. The special images are set as the human body image (default) and grassland image

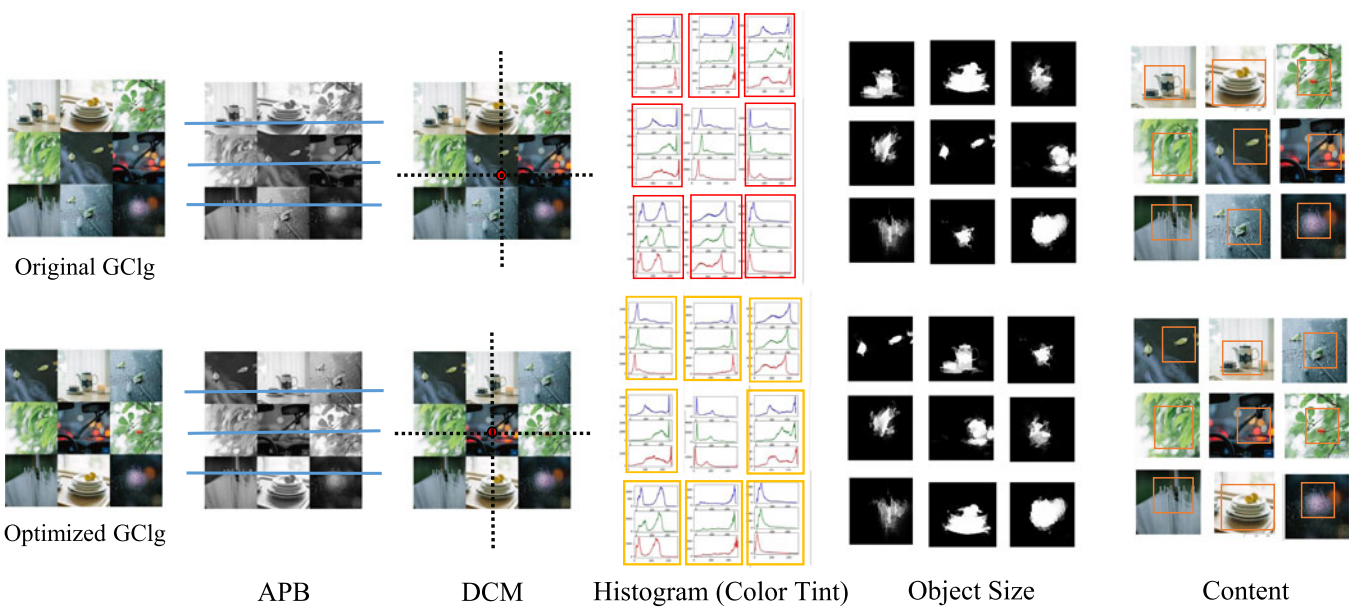


Fig. 11. Measured item visualization of different collages. The yellow box in the content column marks the main content in the image. Please zoom in the document to see the details of the diagrams.

(a) Mea_{ba} w/o tint (b) Mea_{ba} w/o content (c) Mea_{ba} w/o size (d) Mea_{ba} w/o bonus (e) Mea_{ba}

Fig. 12. Results of the ablation study.

TABLE 2
Statistics of the Votes of Different Posting Strategies

Strategy	Random	Baseline1	Baseline2	Ours	Human
Votes	4.5%	11.5%	8.8%	33.9% (47%)	41.3% (52%)

(changed manually). The special locations are annotated as light orange blocks and red boxes in the template. Mea_{ba} is the value calculated by the default configuration, and the value in brackets is calculated by the changed configuration. From this value, we can find that different configurations can lead to different bonus trigger forms, and the corresponding balance values also change.

6.4 Discussions

Nonsquare Inputs. In this paper, the discussion was mainly focused on the interactive generation of GClgs for square inputs. Although images are usually displayed in square shape, the original aspect ratios of images that are uploaded to SNSs are usually nonsquare, as shown in Fig. 19. There are two ways to change the aspect ratio of nonsquare inputs

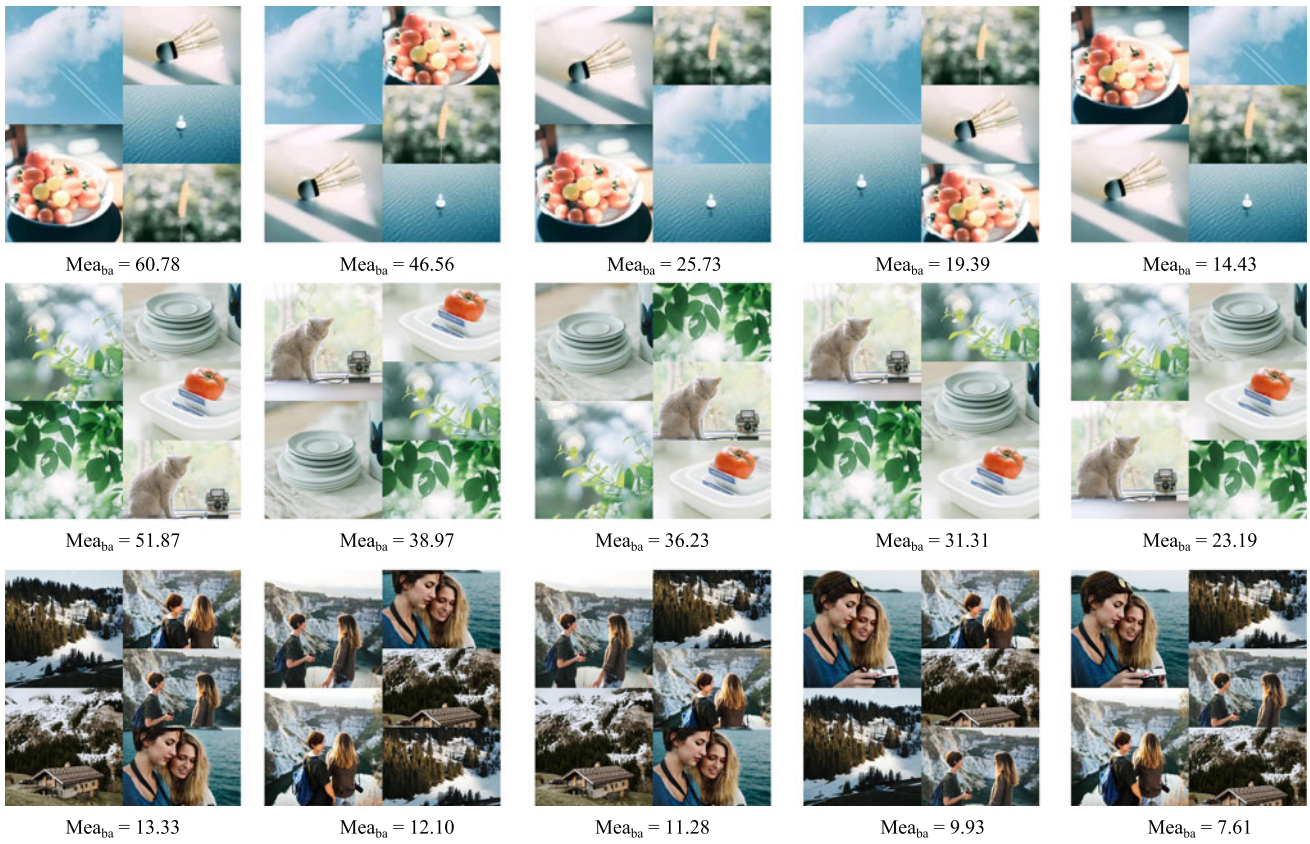


Fig. 13. Extension to Facebook template.

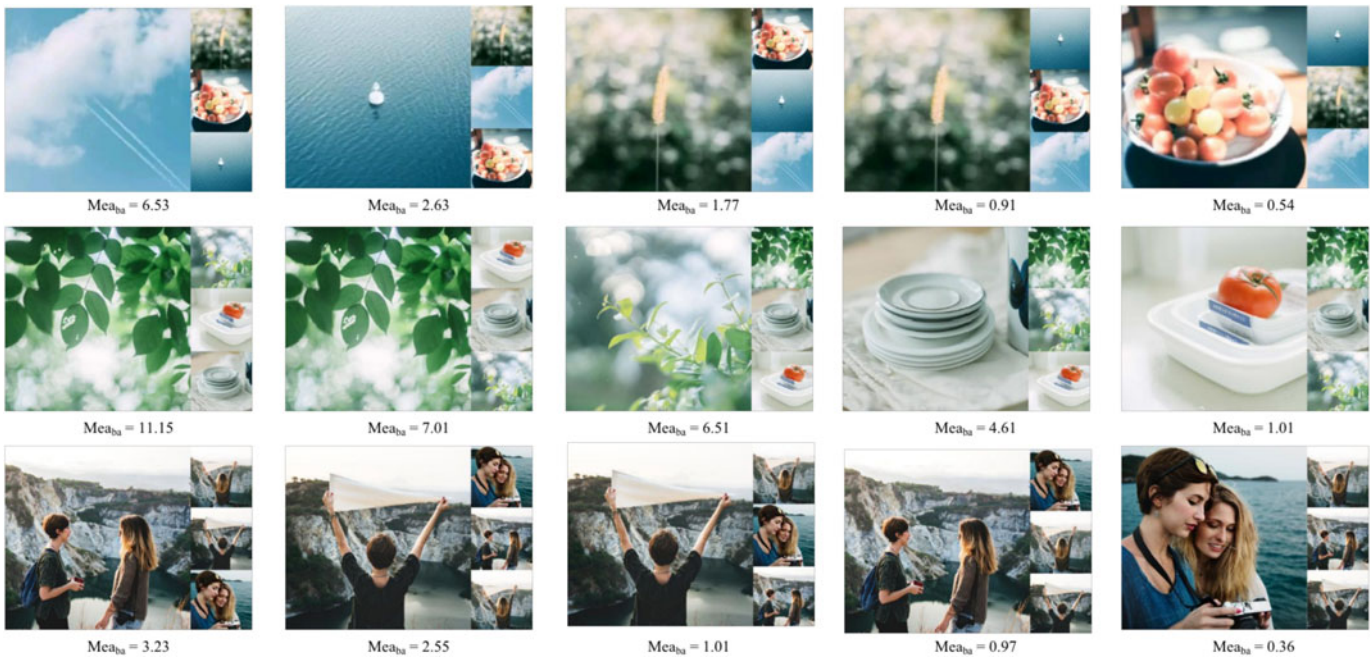


Fig. 14. Extension to Twitter template.

to squares, namely, center-cropping, which is widely used in SNSs, and content-aware image retargeting (CAIR) [50]. Fig. 19 shows the GClg results by different squarization operations. From the two results, it can be observed that using center-cropping based operations will lead to considerable content and composition loss. However, the CAIR method can benefit the original image information reservation and lead to a more informative GClg. How to jointly optimize the squarization and GClg process is an important direction for us to research in the future.

Limitations. One limitation of our method is that the reinforcement learning-based strategy may be trapped in local minima due to the nonconvex loss functions in deep learning. However, the experiments show that our method can

Fig. 15. Uniform templates and the GClg results. The light orange blocks in the templates are user-annotated special locations.

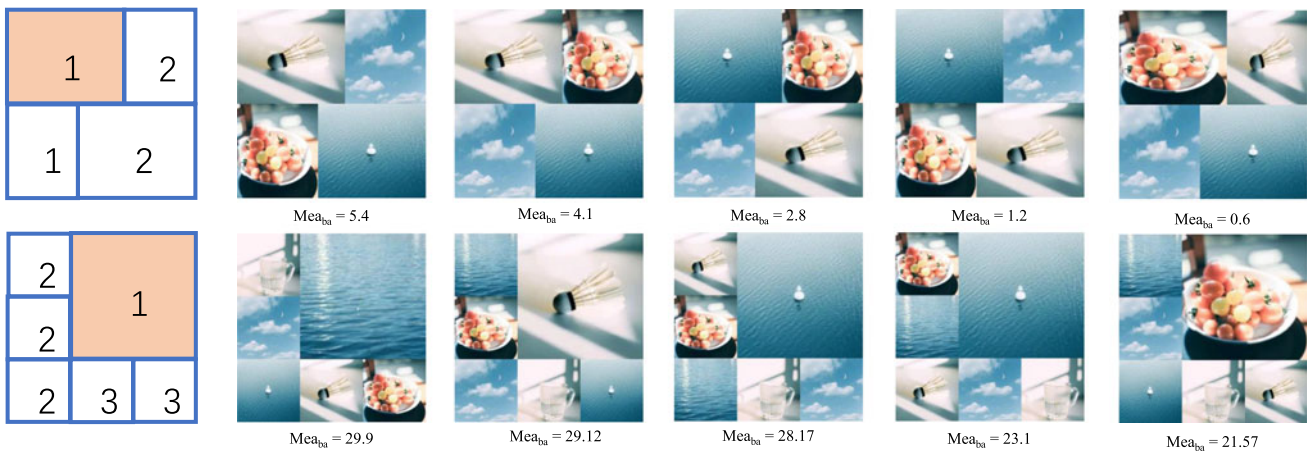


Fig. 16. Nonuniform templates and the GClg results. The light orange blocks in the templates are user-annotated special locations.

still generate good results for most examples. The quality of the results is comparable with the manually designed human results. Speed is another limitation of our work. Currently our unoptimized program usually costs 3–7 seconds to generate a GClg. Using more efficient image feature extraction methods and more efficient reinforcement learning algorithms can accelerate the optimization process.

Bad Case. Fig. 20 illustrates a bad case in our experiment. Our results are visually different from the human results. There are two possible reasons. First, some objects in the images are incomplete, which makes the recognition of semantic information difficult. At the same time, the color features of most images are similar. Thus, it is difficult for our method to accurately measure the visual balance of the GClg. The other possible reason is the local minima. However, it can be observed that the results from this study remain optimal than the random distribution.

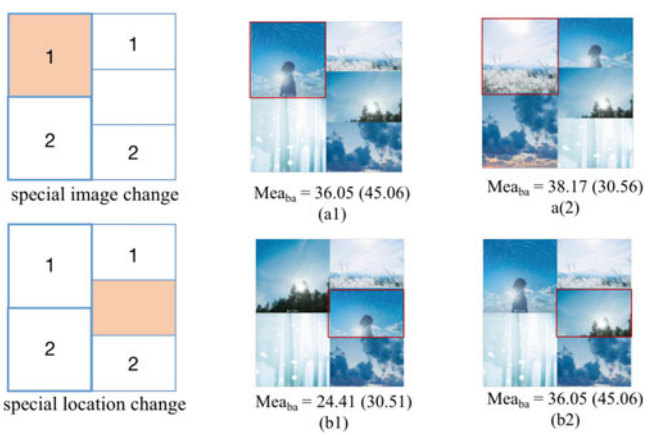


Fig. 17. Personalization on the Twitter template.

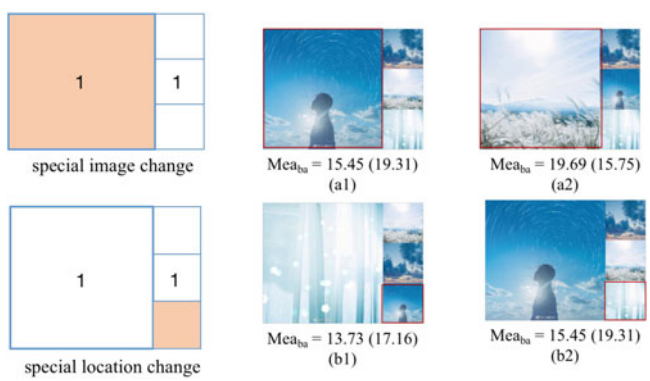


Fig. 18. Personalization on the Facebook template.

Fig. 19. GClg results by different image aspect ratios.

Fig. 20. A bad case of our method.

7 Conclusion

In this study, we focus on the issue of GClg generation for small image collections, which are commonly used in SNSs. In contrast to large-scale image collections, the collage of small-scale image collections must be analyzed in terms of user behavior. A psychology-based balance-aware metric is formulated to evaluate the collage effect for GClg; the metric considers the color shade, color tint, content and object size. Furthermore, a reinforcement GClg generation network is proposed under the guidance of this method to predict the arrangement of the image collection. Experiments show that the metric can reasonably evaluate GClg, and our method can produce visually pleasing GClg results, which are comparable to those of photographers. In the future, we will consider how to transform the general metrics to other templates in a subtle manner.

Acknowledgments

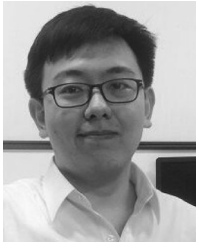
The authors would like to thank the anonymous reviewers for their valuable comments.

References

[1] J. Geigel and A. Loui, "Using genetic algorithms for album page layouts," *IEEE MultiMedia*, vol. 10, no. 4, pp. 16–27, Oct.–Dec. 2003.
[2] C. B. Atkins, "Blocked recursive image composition," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 821–824.
[3] Y. Gan, Y. Zhang, Z. Sun, and H. Zhang, "Qualitative photo collage by quartet analysis and active learning," *Comput. Graph.*, vol. 88, pp. 35–44, 2020.
[4] X. Yang, T. Mei, Y.-Q. Xu, Y. Rui, and S. Li, "Automatic generation of visual-textual presentation layout," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 2, pp. 1–22, Feb. 2016.
[5] D.-S. Ryu, W.-K. Chung, and H.-G. Cho, "Photoland: A new image layout system using spatio-temporal information in digital photos," in *Proc. ACM Symp. Appl. Comput.*, 2010, pp. 1884–1891.
[6] P. Brivio, M. Tarini, and P. Cignoni, "Browsing large image datasets through Voronoi diagrams," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1261–1270, Nov./Dec. 2010.
[7] L. Tan, Y. Song, S. Liu, and L. Xie, "ImageHive: Interactive content-aware image summarization," *IEEE Comput. Graph. Appl.*, vol. 32, no. 1, pp. 46–55, Jan./Feb. 2012.
[8] L. Liu, H. Zhang, G. Jing, Y. Guo, Z. Chen, and W. Wang, "Correlation-preserving photo collage," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 6, pp. 1956–1968, Jun. 2018.
[9] X. Xie, X. Cai, J. Zhou, N. Cao, and Y. Wu, "A semantic-based method for visualizing large image collections," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 7, pp. 2362–2377, Jul. 2019.
[10] X. Pan *et al.*, "Content-based visual summarization for image collections," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 4, pp. 2298–2312, Apr. 2021.
[11] J. Wagemans *et al.*, "A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization," *Psychol. Bulletin*, vol. 138, no. 6, pp. 11–72, 2012.
[12] R. Hübner and M. G. Fillinger, "Comparison of objective measures for predicting perceptual balance and visual aesthetic preference," *Front. Psychol.*, vol. 7, pp. 335:1–335:15, 2016.
[13] R. Hübner and M. G. Fillinger, "Perceptual balance, stability, and aesthetic appreciation: Their relations depend on the picture type," *i-Perception*, vol. 10, no. 3, pp. 1–17, 2019.
[14] A. Wilson and A. Chatterjee, "The assessment of preference for balance: Introducing a new test," *Empirical Stud. Arts*, vol. 23, no. 2, pp. 165–180, 2005.
[15] N. Murray, L. Marchesotti, and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2408–2415.
[16] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 662–679.
[17] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, "Deep multi-patch aggregation network for image style, aesthetics, and quality estimation," in *Proc. IEEE Int. Conf. Comput. Vis*, Dec. 2015, pp. 990–998.
[18] S. Ma, J. Liu, and C. Wen 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. 4535–4544.
[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] W. Wenguan and S. Jianbing, "Deep cropping via attention box prediction and aesthetic assessment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2186–2194.
[21] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 497–506.
[22] K. Sheng *et al.*, "Learning to assess visual aesthetics of food images," *Comput. Vis. Media*, vol. 7, no. 1, pp. 139–152, 2021.
[23] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 879–886.
[24] C. Cui, H. Liu, T. Lian, L. Nie, L. Zhu, and Y. Yin, "Distribution-oriented aesthetics assessment with semantic-aware hybrid network," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1209–1220, May 2019.
[25] L. Li, H. Zhu, S. Zhao, G. Ding, and W. Lin, "Personality-assisted multi-task learning for generic and personalized image aesthetics assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 3898–3910, Jan. 2020.
[26] J. Zhou, Q. Zhang, J.-H. Fan, W. Sun, and W.-S. Zheng, "Joint regression and learning from pairwise rankings for personalized image aesthetic assessment," *Comput. Vis. Media*, vol. 7, no. 2, pp. 241–252, 2021.
[27] K. Thömmes and R. Hübner, "Investigating online liking behaviour for dance portraits on instagram," in *Proc. Vis. Sci. Art Conf.*, 2018.
[28] K. Thömmes and R. Hübner, "Instagram likes for architectural photos can be predicted by quantitative balance measures and curvature," *Front. Psychol.*, vol. 9, pp. 1050:1–1050:17, 2018.
[29] Y. Cao, A. B. Chan, and R. W. H. Lau, "Automatic stylistic manga layout," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–10, Nov. 2012.
[30] Y. Cao, R. W. H. Lau, and A. B. Chan, "Look over here: Attention-directing composition of manga elements," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–11, Jul. 2014.
[31] Z. Yu, L. Lu, Y. Guo, R. Fan, M. Liu, and W. Wang, "Content-aware photo collage using circle packing," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 2, pp. 182–195, Feb. 2014.

[32] X. Han, C. Zhang, W. Lin, M. Xu, B. Sheng, and T. Mei, "Tree-based visualization and optimization for image collection," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1286–1300, Jun. 2016.
[33] Y. Liang, X. Wang, S. Zhang, S. Hu, and S. Liu, "Photorecomposer: Interactive photo recomposition by cropping," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 10, pp. 2728–2742, Oct. 2018.
[34] Z. Wu and K. Aizawa, "Very fast generation of content-preserved photo collage under canvas size constraint," *Multimedia Tools Appl.*, vol. 75, no. 4, pp. 1813–1841, Feb. 2016.
[35] X. Zheng, X. Qiao, Y. Cao, and R. W. H. Lau, "Content-aware generative modeling of graphic design layouts," *ACM Trans. Graph.*, vol. 38, no. 4, pp. 1–15, Jul. 2019.
[36] Y. Song, F. Tang, W. Dong, and C. Xu, "Non-dominated sorting based multi-page photo collage," *Comput. Vis. Media*, 2021.
[37] P. J. Locher, P. J. Stappers, and K. Overbeeke, "The role of balance as an organizing design principle underlying adults' compositional strategies for creating visual displays," *Acta Psychol.*, vol. 99, no. 2, pp. 141–161, 1998.
[38] J. Geigel and A. C. P. Loui, "Automatic page layout using genetic algorithms for electronic albuming," in *Proc. Internet Imag. II*, 2000, pp. 79–90.
[39] S. Lok, S. Feiner, and G. Ngai, "Evaluation of visual balance for automated layout," in *Proc. Int. Conf. Intell. User Interfaces*, 2004, pp. 101–108.
[40] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," in *Proc. NIPS Deep Learn. Workshop*, 2013.
[41] K. Zhou, Y. Qiao, and T. Xiang, "Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7582–7589.
[42] A. Clegg, W. Yu, J. Tan, C. K. Liu, and G. Turk, "Learning to dress: Synthesizing human dressing motion via deep reinforcement learning," *ACM Trans. Graph.*, vol. 37, pp. 179:1–179:10, 2018.
[43] T. Y. Satoshi Kosugi, "Unpaired image enhancement featuring reinforcement-learning-controlled image editing software," in *Proc. Amer. Assoc. Artif. Intell.*, 2020, pp. 11296–11303.
[44] E. Choi and C. Lee, "Feature extraction based on the bhattacharyya distance," *Pattern Recognit.*, vol. 36, no. 8, pp. 1703–1709, 2003.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, May 2015.
[46] G. Lee, Y. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 660–668.
[47] J. Bromley, I. Guyon, Y. Lecun, E. Säckinger, and R. Shah, "Signature verification using a siamese time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1993, pp. 669–688.
[48] "Collageit," Accessed: 2019. [Online]. Available: http://www.collageitfree.com/
[49] V. Cheung, "Shape collage," Accessed: 2013. [Online]. Available: https://shapecollage.com/
[50] Y. Song, F. Tang, W. Dong, X. Zhang, O. Deussen, and T.-Y. Lee, "Photo squarization by deep multi-operator retargeting," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 1047–1055.

Yu Song received the MSc degree in automation from Tianjin University in 2017. She is currently working toward the PhD degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her research interests include image collage, image retargeting, and machine learning.



Fan Tang received the PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, in 2019 and the BSc degree in computer science from North China Electric Power University in 2013. He is currently an assistant professor with the School of Artificial Intelligence, Jilin University. His research interests include image synthesis and image recognition.

Weiming Dong (Member, IEEE) received the BSc and MSc degrees in computer science from Tsinghua University, China, in 2001 and 2004, respectively, and the PhD degree in computer science from the University of Lorraine, France, in 2007. He is currently a professor with the Sino-European Lab in Computer Science, Automation and Applied Mathematics and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include visual media synthesis and image recognition. He is a member of the ACM.

Feiyue Huang received the BSc and PhD degrees in computer science from Tsinghua University, China, in 2001 and 2008, respectively. He is currently the director of Youtu Lab, Tencent. His research interests include image understanding and face recognition.

Tong-Yee Lee (Senior Member, IEEE) received the PhD degree in computer engineering from Washington State University, Pullman, 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.

Changsheng Xu (Fellow, IEEE) is currently a professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and the Executive Director of the China-Singapore Institute of Digital Media. His research interests include multimedia content analysis or indexing or retrieval, pattern recognition, and computer vision. He has hold 30 granted or pending patents and has authored or coauthored more than 200 refereed research papers in these areas. He is a fellow of IAPR and an ACM Distinguished Scientist. He was a program chair of ACM multimedia 2009. He was an associate editor, a guest editor, a general chair, a program chair, an area or track chair, a special session organizer, a session chair, and a TPC member for more than 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops. He is an associate editor for the *ACM Transactions on Multimedia Computing, Communications and Applications* and Editor-in-Chief of *ACM/Springer Multimedia Systems Journal*. He is IAPR fellow, and ACM Distinguished scientist. He was the recipient of the Best Associate Editor Award of *ACM Transactions on Multimedia Computing, Communications and Applications* in 2012 and the Best Editorial Member Award of *ACM/Springer Multimedia Systems Journal* in 2008.