Content-Based Visual Summarization for Image Collections

Xingjia Pan, Fan Tang, Weiming Dong, *Member, IEEE*, Chongyang Ma, Yiping Meng, Feiyue Huang, Tong-Yee Lee, *Senior Member, IEEE*, and Changsheng Xu, *Fellow, IEEE*

Abstract—With the surge of images in the information era, people demand an effective and accurate way to access meaningful visual information. Accordingly, effective and accurate communication of information has become indispensable. In this article, we propose a content-based approach that automatically generates a clear and informative visual summarization based on design principles and cognitive psychology to represent image collections. We first introduce a novel method to make representative and nonredundant summarizations of image collections, thereby ensuring data cleanliness and emphasizing important information. Then, we propose a tree-based algorithm with a two-step optimization strategy to generate the final layout that operates as follows: (1) an initial layout is created by constructing a tree randomly based on the grouping results of the input image set; (2) the layout is refined through a coarse adjustment in a greedy manner, followed by gradient back propagation drawing on the training procedure of neural networks. We demonstrate the usefulness and effectiveness of our method via extensive experimental results and user studies. Our visual summarization algorithm can precisely and efficiently capture the main content of image collections better than alternative methods or commercial tools.

Index Terms—Visual summarization, photo collection, collage layout, tree-based algorithm, gradient back propagation

1 Introduction

With the rapid growth of the Internet and the explosion of big visual data, effectively organizing large-scale image collections and enabling efficient and engaging viewing experiences have become increasingly important. Thus, visual summarization has become an important practice in the automation of these processes [1].

Automatic visual summarization involves the following two major steps: (1) choosing an appropriate number of images from the collection as the summary and (2) collaging them into a visually pleasing layout based on the image content. The summarization of an image collection aims to select a brief yet informative subset of images (summary) to accurately represent the visual information of a large set [2], which is difficult for three reasons. First, the objects, scenes, and visual quality of different images often vary considerably. Second, repetitive or similar images may be present in the collection, and sometimes, the number of images in the summary also needs to be flexibly controlled. Third, computing *content-based* summarization, which aims to select a summary to cover the event of an image collection, remains an open problem. Previous studies have shown that when users choose images from a collection, the results tend to be a random assortment of their favorite photos rather than a group of images that tells a story and effectively works as an overview [3]. Observers perceive certain elements of the visual field as a whole more than the sum of parts [4]. The field of layout generation also faces several challenges. First, images exhibit diverse aspect ratios and sizes, thereby making it difficult to generate a compact layout. Second, informative layouts are usually content aware, indicating that they should also preserve the correlation among the input images. Lastly, an image layout should meet certain high-level aesthetics principles, such as perception balance, to be visually pleasing.

Many researchers have developed automatic or interactive summarization systems by using handcrafted (color, texture, and edge) or deep features to generate a satisfactory summary of an image collection [5], [6], [7]. Automatic image collection summarization typically uses a content analysis scheme, such as latent topic analysis [8], which may not always provide a desirable summary of photos that satisfy the information integrity and visual aesthetics. Optimal visualization is seldom considered in current image collection summarization techniques. The summary is usually displayed in a simple layout, thereby making it difficult for the viewers to obtain information efficiently.

- *X. Pan, W. Dong, and C. Xu are with the NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100864, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: {panxingjia2015, weiming.dong, changsheng.xu}@ia.ac.cn.*
- *F. Tang is with the Fosafer, Beijing 100142, China. E-mail: tangfan@fosafer.com.*
- *C. Ma is with the Kuaishou Technology, Beijing 100085, China. E-mail: chongyangm@gmail.com.*
- *Y. Meng is with the Didi Chuxing, Beijing 100000, China. E-mail: mengyipingkitty@didichuxing.com.*
- *F. Huang is with the YouTu Lab, Tencent, Shanghai 200233, China. E-mail: garyhuang@tencent.com.*
- *T.-Y. Lee is with the National Cheng Kung University34912, Tainan 701, Taiwan. E-mail: tonylee@mail.ncku.edu.tw.*

Manuscript received 10 Jan. 2019; revised 15 Oct. 2019; accepted 16 Oct. 2019. Date of publication 23 Oct. 2019; date of current version 25 Feb. 2021. (Corresponding author: Weiming Dong.)
Recommended for acceptance by X. M Tricoche.
Digital Object Identifier no. 10.1109/TVCG.2019.2948611

Although manual or interactive generation and layout of summary images can achieve desirable visual summarization, these processes depend on the user expertise and can be time consuming.

Several researchers have developed several automatic image layout generation methods to visualize image summarization effectively [9], [10], [11], [12], [13], [14]. These approaches have achieved a certain level of success in improving the visual perception of image summarization. However, these image layout methods either ignore the aspect ratio [1], [12], [13], [14] or fail to illustrate the entire content structure of images clearly [9], [11]. Without an aesthetic evaluation as guidance, they cannot ensure the visual attractiveness of the collage results, which is a key factor of visual summarization [15], [16]. To solve this issue, researchers have proposed interactive methods, which conduct subjective experiments to capture and learn aesthetic appeal from user preferences in terms of the placement of multiple photos on a canvas [17] or the composition of visual elements [18]. However, these methods either require laborious data annotation or ignore the utility of visual summarization in information presentation. Both drawbacks limit the practical application of these methods.

Our main idea for the development of a visual summarization method mainly comes from two key observations. First, we should improve the *representativeness* of an image set that needs to accurately measure the *diversity*, *redundancy*, and *aesthetics* of an image set for the generation of an informative summary. Second, we should integrate visual aesthetics into the layout generation process by considering the content structure of the images to display the summary clearly and to facilitate the information acquisition and visual search. In this work, we propose a content-based approach to generate visual summarization automatically, which integrates the image collection summarization and layout generation processes. We present a content-based summarization method, which selects images by considering the content diversity, content conciseness, and visual aesthetics simultaneously to create an image summary of the image collection. Thereafter, we propose a novel tree-based layout generation method to produce a visually pleasing, informative, and compact image layout. Specifically, we partition a canvas into cells with various sizes and aspect ratios by using a binary tree whose leaf nodes are associated with the appropriate images. The tree optimization method consists of two steps, namely, coarse adjustment and refinement via gradient back propagation based on the training process of neural networks.

Our work makes the following technical contributions:

- *a hidden topic-based diversity analytical method* to measure the representativeness of an image collection summary effectively;
- *a visual aesthetics-embedding mechanism* that formulates high-level aesthetics principles, such as Gestalt theory and perceptual balance, into a computational model to guide the visual summarization of image collections;
- *a tree-based image layout framework* that integrates a two-step optimization strategy. To the best of our knowledge, we are the first to apply back propagation to the optimization of a tree-based structure for image layout generation.

2 Related Work

Photo Summarization. A high-quality summary helps people to effectively and efficiently capture the key information of the original dataset. Kennedy et al. [19] generated a summary of frequently photographed views by mining the photos shared among many individuals. Additional information, such as geo-locations or landmark tags, is required in the aforementioned authors' approach. Sinha et al. [20] defined three properties, namely, quality, diversity, and coverage, to select an informative summary of portrait photos by using the metadata information of images given a size constraint. For image collages, Tan et al. [1] and Liu et al. [11] considered several low-level features and used the k-means clustering algorithm to group images according to photo correlation, content similarity, or color distribution. Kim et al. [21] extracted key frames from image and video collections for storyline reconstruction by considering diversity and coverage with the help of text annotations. By considering only diversity and coverage (nonredundancy) [19], [20], [21], the summarizations usually contain certain low-quality images because noise usually differentiates these images from the others. Lidon et al. [22] first removed less informative images via a CNN-based filter and then fused three results ranked on the basis of saliency, objectness, and face number to select a diverse and nonredundant image set. Cong et al. [7] designed a dictionary selection model to select a subset of bases. They need to select 10 photos manually to initialize the creation of a photo album; this process is inconvenient for the summarization of large image sets. In contrast to these previous studies, we first explore the characteristics of the feature space to excavate hidden topics that assists in the creation of a representative image summary. We then add a criterion named *visual aesthetics* to remove low-quality images.

Aesthetics Analysis. Visual aesthetics is an important factor for image layouts. Locher et al. [23] reported that visual balance is one of the important elements of image design and composition. Ngo et al. [24] introduced an aesthetics model to measure the balance, equilibrium, symmetry, and sequence of a user interface and computer screen. Geigal et al. [25] measured the layout quality from several factors, including balance, emphasis, chronology, and unity. Lok et al. [26] used WeightMap to enable an automated system to evaluate layout effectiveness. Yang et al. [27] provided predefined topic-dependent templates featuring aesthetics principles and considered the visual weight of each element to maintain symmetrical balance and the golden ratio. Hubner et al. [28] indicated that the aesthetics appreciation of a picture largely depends on the perceptual balance of its element. In our work, we formulate high-level aesthetics principles to a computational model and embed them into our layout generation procedure.

Image Collage. From the generation strategy view, image layout methods can be roughly divided into two categories, namely, bottom-up and top-down layout generation. Bottom-up methods usually use well-defined objective functions and

their parameters contain the position, orientation, scale, and layer index of each image on the canvas [17], [29], [30], [31], [32], [33], [34]. These techniques usually have limitations regarding avoiding overlaps and precisely controlling the correlations among images. There are also many projection-based algorithms that can effectively preserve image correlations. These methods first calculate the correlations among images in a feature space and then project the images into a visualization space with the help of dimensionality reduction techniques [10], [12], [35]. Han et al. [10] proposed a tree-based visualization approach, which can generate a collage with several interesting shapes. However, such an approach is not seamless and needs to specify the target projection area of each image. Furthermore, VPSC [36], PRISM [37], and RWordle-C [38] proposed techniques to remove overlaps. Gomez-Nieto et al. [12], [13], [14] generated efficiently structured layouts that not only preserve the semantic relationship among objects but also remove overlaps. The use of these methods to generate an image collage has a large drawback. They did not consider preserving the aspect ratios of the images. Recently, Liu et al. [39] estimated the pixel-wise importance of one image to extract semantic components and then integrated these components by using a fusion network to generate a semantic collage. They only processed one image and mainly focused on image retargeting.

Top-down methods recast collage generation as a region partitioning problem [1], [11], [40], [41], [42]. Geigel et al. [25] used genetic algorithms for automatic page layout with fitness based on the graphic design preferences supplied by the user. Tan et al. [1] used a graph-based algorithm to establish a global placement and then applied online Voronoi tessellation to refine the final layout. Yu et al. [40] used a circle to approximate the salient region of each image and then formulated a photo collage as a circle packing problem. Liu et al. [11] presented an irregular shape-based canvas partition method to generate a compact collage. Nguyen et al.'s [35] approach and almost all the above-mentioned methods, which are based on the maximization of the image salience region, suffer from information loss due to cropping or cutting. Liang et al. [18] recomposed image collection based on example photos and generated a layout by using a Voronoi tree map that is complicatedly optimized. Wu et al. [41] proposed a binary tree-based page division and adjustment algorithm and preserved the image correlations. However, the use of an adjustment scheme alone is insufficient because the images still suffer from severe ratio distortion. Furthermore, they predefined several templates to assist in the highlighting of specific images, thereby limiting the collage scalability.

3 Overview

Our method consists of two major modules, namely, content-based summarization (Section 4) and tree-based layout generation (Section 5), to create a visual summarization for an image collection, as illustrated in Fig. 2.

Content-Based Summarization. In the content-based summarization process, we select a representative and non-repetitive subset of the input collection based on image content analysis. To this end, we propose a hidden topic-based model to measure the collection diversity. We also consider the redundancy and visual aesthetics of the input images when generating the summarization.

Tree-Based Layout Generation. After a subset of the input image collection has been selected, we organize the selected images into a collage by using a tree-based layout generation method. First, we randomly initialize the tree structure and assign a selected image to each leaf node by considering the fitness of size, ratio, content, etc. Subsequently, we refine the layout using a two-step optimization strategy. In the first step, we adjust the splitting direction and division ratio of each nonleaf node of the tree in a top-down manner. In the second step, we formulate the task of refining the tree-based layout as an optimization problem with several energy terms representing different criteria. We apply gradient back propagation algorithm to solve the optimization problem and obtain a refined tree structure as the final layout.

4 Content-Based Summarization

Summarization offers an intuitively clear method to represent the semantic topics and visual effects of a large-scale image collection. Given an image collection, a satisfactory summarization should meet the following three criteria:

- *Content diversity.* The contents of the images in the summarization should be highly diverse.
- *Content conciseness.* The images in the summarization should be nonredundant.
- *Visual aesthetics.* The images in the summarization should be visually appealing.

We define three metrics, namely, the diversity score E_{div}, concise ratio E_{con}, and aesthetics score E_{aes}, to measure the quality of a summarization $\mathcal{S}$ objectively. The optimal summarization $\mathcal{S}^*$ can then be calculated as follows:

$$\mathcal{S}^* = \arg\max_{\mathcal{S}^* \in \mathcal{S}} \mathcal{F}(E_{div}(\mathcal{S}), E_{con}(S), E_{aes}(\mathcal{S})), \tag{1}$$

where $\mathcal{F} : E_{div}(\mathcal{S}) \times E_{con}(\mathcal{S}) \times E_{aes}(\mathcal{S}) \rightarrow \mathbb{R}$. In the rest of this section, we start from the calculation of $E_{div}(\mathcal{S})$, $E_{con}(\mathcal{S})$, and $E_{aes}(\mathcal{S})$ (Sections 4.1 and 4.2) and then explain the summarization generation based on a greedy selection procedure (Section 4.3).

4.1 Hidden Topic Based Diversity Analysis

We define the diversity of an image set as the difference in the images in terms of the visual content and semantic topics. Previous studies often use metadata, tags of web images, or human-defined labels, such as whether the face is contained and the object class, as the content labels to measure diversity [19], [20]. However, human-defined labels may sometimes be incomplete; thus, an advanced semantic topic basis is required for the image representation. Based on network dissection [43], we observe that human-interpretable concepts sometimes emerge as individual latent variables in large deep neural networks. Thus, the second to last fully connected layer in most CNN networks can be treated as the linear combination of the variable semantic concepts. The linear combination serves as the global representation of an image, of which each

dimension encodes individual latent concepts, namely, the hidden semantic topics of the image.

Generally, the three distinct diversity aspects are *variety*, *balance*, and *disparity* [44]. Variety is the number of categories into which the system elements are apportioned. Balance is a function of the pattern of apportionment of the elements across categories. Disparity refers to the manner and degree by and at which the elements may be distinguished.

Given an image set $\mathcal{I}$ with N images and a summarization $\mathcal{S}$ with k images, we denote $f \in \mathbb{R}^M$ as the hidden topic feature vector of an image I and then $F_{\mathcal{I}} \in \mathbb{R}^{N\times M}$ and $F_{\mathcal{S}} \in \mathbb{R}^{K\times M}$ as the hidden topic feature matrices of $\mathcal{I}$ and $\mathcal{S}$, respectively. The diversity score of an image set is defined as follows:

$$E_{div}(\mathcal{S}) = \sum_{d=1}^{M} T(d)W(d) + \mu \sum_{d_1=1}^{M}\sum_{d_2\neq d_1}^{M} T(d_1)\cdot T(d_2)\cdot \delta(d_1,d_2), \tag{2}$$

where $T(d) = \max_i F_{\mathcal{S}}(i,d)$ indicates the probability that the image set contains the corresponding topic, and μ is a weighting parameter controlling the balance of the two items and is set to 0.001 to make their values roughly the same magnitude. Variable E_{div} can also measure the diversity of one image, indicating the presence of only one image in a certain set. $\delta(d_1,d_2)$ measures the *disparity* and indicates the distance between the d_1 and d_2 topics. $W \in \mathbb{R}^M$ represents the normalized weights of the hidden topics, which are defined as follows:

$$W(d) = \frac{e^{\sum_{d_2}\delta(d,d_2)}}{e^{\sum_d\sum_{d_2}\delta(d,d_2)}}. \tag{3}$$

The *balance* component is inexplicitly defined in Equation (2) because it is implicitly encoded in the latter item by the multiplication operation.

4.2 Conciseness and Aesthetics Evaluation

A satisfactory summarization should be content concise and visually pleasing. We start from the measurement of the degree of conciseness of an image I_i to an image set $\mathcal{I}$, as follows:

$$R(I_i,\mathcal{I}) = \min_{I_j\in\mathcal{I}\setminus I_i} Dis(I_i,I_j), \tag{4}$$

where $Dis(I_i,I_j)$ is the euclidean distance between the features of images I_i and I_j, and $\setminus$ means the difference between two sets. A large value of $R(I_i,\mathcal{I})$ means that image I_i is close to image set $\mathcal{I}$. Thus, the conciseness of an image summarization $\mathcal{S}$ can be calculated as follows:

$$E_{con}(\mathcal{S},\mathcal{I}) = \frac{1}{|S|}\sum_{I_i\in\mathcal{S}} R(I_i,\mathcal{I}). \tag{5}$$

We define $A(I)$ as the aesthetics score of an image to increase the visual quality of the summary and obtain the value by using an attention-based approach [45]. Subsequently, we calculate the aesthetics score of an image summarization as follows:

$$E_{aes}(\mathcal{S}) = \frac{1}{|S|}\sum_{I_i\in\mathcal{S}} A(I_i). \tag{6}$$

4.3 Summarization Generation

We adopt a greedy selection strategy to observe a suboptimal solution because subset selection is an NPC problem. Our summarization algorithm aims to analyze the input set $\mathcal{I} = \{I_1,\ldots,I_N\}$ and iteratively construct the summarization set containing K images $S_K^* = \{s_1^*,\ldots,s_K^*\}$, where $K << N$. We sort the images in $\mathcal{I}$ as $\mathcal{I}^r = \{I_1^r,\ldots,I_N^r\}$ based on the aesthetics score. Our algorithm begins by selecting the top ranked image in $\mathcal{I}^r$ as the first element of S^*, namely $S_1^* = \{I_1^r\}$. The optimal candidate image I_i^* is then added to the summary at iteration $i = 2,3,\ldots K$ as follows:

$$I_i^* = \arg\max_{I_j^r\in\mathcal{I}^r\setminus S_i^*} Score(I_j^r, S_i^*),$$

where $Score(\cdot)$ considers the conciseness, diversity gain ratio, and aesthetics as follows:

$$Score(I_j^r,S_i^*) = r_{con}(I_j^r,S_i^*) + r_{div}(I_j^r,S_i^*) + r_{aes}(I_j^r),$$

where $r_{\{\cdot\}}$ indicates the normalized scores depending on the ranking position; and $r_{\{\cdot\}} = \frac{m-P_{\{\cdot\}}(I_j^r)}{m-1}$, where $P_{\{\cdot\}}(I_j^r) = 1,2,\ldots,m$ is the ranking position of image I_j^r according to the conciseness, aesthetics and diversity gain. m is the number of image set $\mathcal{I}^r \setminus S_i^*$. Specifically, $P_{\{con\}}(I_j^r)$ and $P_{\{aes\}}(I_j^r)$ are obtained in accordance with Equations (4), (5), and (6), that is, $R(I_j^r,S_i^*)$ and $A(I_j^r)$), respectively. We rank the images in $\mathcal{I}^r \setminus S_i^*$ according to the diversity gain $E_{div}(I_j^r \cup S_i^*)/E_{div}(S_i^*)$ to calculate $P_{\{div\}}(I_j^r)$. Thus, the greedy selection procedure iteratively chooses the next image in the summarization as follows:

$$I_{i+1}^* = \arg\max_{I_j^r\in\mathcal{I}^r\setminus S_i^*} Score(I_j^r,S_i^*),$$
$$S_{i+1}^* = S_i^* \cup I_{i+1}^*,$$

and it stops according to the constraints on the summarization volume or the threshold of the diversity gain ratio.

As shown in Fig. 1, we produce summarizations with eight images generated by random selection, k-means clustering, and our content-based summarization, respectively. The original gallery consists of 86 images of a wedding scene. The images show the photos of the bride, the groom, and the groom's friends. The results of the top two rows mainly cover the scene containing the bride and the groom. However, the groom's friends are neglected. With sufficient information on the entire event, our method can obtain a more visually pleasing summarization of images.

5 Tree-Based Layout Generation

We use a photo collage to visualize the summarization. Intuitively, one photo collage consists of multiple microcells corresponding to each image. Thus, we present a novel tree-based layout method to generate a collage. This task is initiated to create an informative, compact, and visually

(a) Random selection

(b) k-means clustering

(c) Our method

Fig. 1. Three example summarizations obtained by random selection, k-means clustering, and our method. The original gallery consists of 86 images showing a wedding scene. The images mainly show a bride, a groom, and the groom's friends.

pleasing collage. As shown in Fig. 2, our layout generation process consists of a two-stage initial layout construction (Fig. 2c) and a two-step optimization strategy (Fig. 2d). The construction and optimization operations are based on a tree structure, which embeds the spatial partition of the canvas.

In the rest of this section, we first formulate the layout generation process as an optimization problem and propose a visual aesthetics-based object function according to high-level aesthetic principles (Section 5.1). We then introduce the two-stage tree construction for the initial layout generation by recursively splitting the canvas into a grid with microcells (Section 5.2). Lastly, we describe our two-step optimization method, which contains coarse adjustment and refined optimization (Section 5.3).

5.1 Problem Formulation

The collage of an image set is generated by placing each one into a canvas at a certain position, size and length-width ratio. We have formulated the layout representation $\mathcal{L} = \{(Lc_i, Sz_i, Sh_i)|0 < i \leq K\}$, where K is the size of the summarization. Variables Lc_i, Sz_i, and sh_i describe the location, size, and aspect ratio of the i_{th} image in the collage, respectively. Our layout goal is defined by incorporating the perceptual balance of high-level aesthetic principles into the computational model. Three criteria, namely, the ratio preservation, size diversification, and content-aware and aesthetics-embedded location, are considered. To this end, we have formulated the visualization as an optimization problem that jointly considers these criteria. The optimization of $\mathcal{L}$ is formulated as follows:

$$\mathcal{L}^{opt} = \arg \min_{Lc,Sz,Sh} (C_{sh} + C_{sz} + C_{lc}), \tag{7}$$

where $\mathcal{L}^{opt} = \{Lc^{opt}, Sz^{opt}, Sh^{opt}\}$ is the optimal location, ratio, and size of the images in the layout. $C_{sh}(\cdot)$, $C_{sz}(\cdot)$, and $C_{lc}(\cdot)$ are the ratio preservation, size diversification, and content-aware and aesthetics-embedded location costs, respectively. Satisfactory layouts can be achieved by optimizing these three criteria.

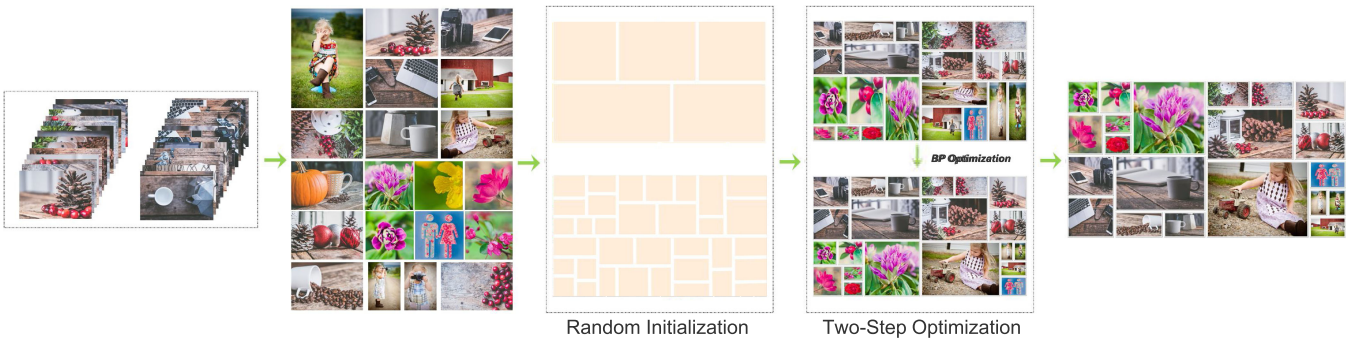


Fig. 2. Workflow of our visual summarization pipeline. Given an input image collection. (a) we first make a representative summary. (b) The output collage is randomly initialized and then refined by optimization. (c) We show the final visual summary in (d).

Ratio Preservation. We introduce a ratio preservation cost to preserve the aspect ratio of images during layout generation as follows:

$$C_{sh}(\mathcal{S},\mathcal{L}) = \frac{1}{K}\sum_{k=0}^{K} c_{sh,k}(\mathcal{S},\mathcal{L}) + \max_k c_{sh,k}(\mathcal{S},\mathcal{L}), \tag{8}$$

where $c_{sh,k}(\cdot)$ is the ratio preservation cost of the kth image and is calculated as the ratio between the image aspect ratio and that of the corresponding cell in the layout.

Size Diversification. Instead of uniformly treating each image, we place the representative ones into large cells to increase the semantic contrast between groups of images with different contents in the summarization. This process is consistent with the "similarity" principle of a Gestalt-based design method [4]. With the constraint of ratio preservation, we only consider the widths of the cells in the size preservation constraint. For instance, the width and height of one cell are w and h, and then the aspect ratio of the cell is $\alpha = w/h$. When we optimize the width of the cell, its height will automatically be optimized in accordance with $h = w/\alpha$ to preserve the aspect ratio. We denote C_{sz} as the size diversification cost of the layout and formulate the corresponding cost of the kth image as follows:

$$c_{sz,k}(\mathcal{S},\mathcal{L}) = \max\left(\frac{w_k^L}{w_k\cdot\gamma_{sz}^{0.5}}, \frac{w_k\cdot\gamma_{sz}^{0.5}}{w_k^L}\right), \tag{9}$$

where w_k and w_k^L are the width of an image and its corresponding cell in the layout, respectively. We denote $\gamma_{sz} = Area_L / \sum_{k=1}^{K} Area_k$ as the area ratio between the layout and the original images, where $Area_L$ is the layout area, and $\sum_{k=1}^{K} Area_k$ is the area sum of all images.

Content-Aware and Aesthetics-Embedded Image Location. We embed content correlation and aesthetic principles into our layout generation algorithm to create a more informative and visually pleasing layout. We formulate the image location cost as follows:

$$C_{lc}(\mathcal{S},\mathcal{L}) = \frac{1}{K}\sum_{k=0}^{K} c_{sm,k}(\mathcal{S},\mathcal{L}) + c_{cp} + c_{aes}, \tag{10}$$

where $c_{sm,k}$ is the position-sensitive cost of the kth image on the layout, which places the representative images in salient positions. c_{cp} and c_{aes} are the content preserving and aesthetics terms of the entire layout, respectively.

$\boldsymbol{c_{sm}}$ controls the image positions as follows:

$$c_{sm,k}(\mathcal{S},\mathcal{L}) = |\frac{x_k}{W_L} - \delta_k\cdot 0.5| + |\frac{y_k}{H_L} - \delta_k\cdot 0.5|, \tag{11}$$

where (x_k, y_k) is the center point coordinate of image I_k in the layout. W_L and H_L are the width and height of the canvas, respectively. δ_k is an indicator function, as follows:

$$\delta_k = \begin{cases} 1, & I_k \in \mathcal{S}_{sub}, \\ 0, & else, \end{cases} \tag{12}$$

where $\mathcal{S}_{sub}$ is a subset of summarization $\mathcal{S}$, which contains the most information. Thus, representative images will be placed as close as possible to the layout center because the center of a frame is usually the most attractive [28].

$\boldsymbol{c_{cc}}$ is the content preservation cost to place content-correlated images together, as follows:

$$c_{cp}(\mathcal{S},\mathcal{L}) = \sum_{i\in S}\sum_{j\in\int_i} Lk(i,j) + Cp(i,j), \tag{13}$$

where $\int_i = \{s_k | s_k \text{ is the content correlated to } s_i\}$. $Lk(i,j)$ and $Cp(i,j)$ are the functions that measure the connectivity and distance between images s_i and s_j, respectively. Our two-stage tree construction strategy can effectively serve this term and it ensures that the contents of similar images are closely arranged and compact.

$\boldsymbol{c_{aes}}$ is the term of the aesthetics property, which largely depends on perceptual balance [26], [28]. We use c_{aes} to control the balance of the color distribution and layout structure, as follows:

$$c_{aes}(\mathcal{S},\mathcal{L}) = \sum_{i\neq j} < H_i, H_j > + \sqrt{\frac{\sum \hat{w}_i r_x}{\sum \hat{w}_i} + \frac{\sum \hat{h}_i r_y}{\sum \hat{h}_i}}, \tag{14}$$

where H_i and H_j are color histograms of images I_i and I_j defined over a hue channel of HSV color space according to [46]. The first term of Equation (14) controls the balance of the color distribution by minimizing the euclidean distance of the color histograms between each adjacent image in the collage. $\hat{w}_i, \hat{h}_i$ are the normalized width and height of image I_i and those divided by the collage width and length, respectively. r_x, r_y are the coordinates of the center of image I_i and are also normalized. Based on the measurement of the "Deviation of the Center of Mass" (DCM) [28], we use the image width $\hat{w}$ and height $\hat{h}$ as the "mass" and distance between the center of the image and the upper left corner of the layout as "r" to calculate the balance term.

5.2 Layout Initialization

We initialize the layout recursively based on the tree structure described in detail below.

Two-Stage Tree Construction. Given the image collection summarization $\mathcal{S}$, we first select representative images by using our method described in Section 4 as anchors, S_a, by setting the threshold of the diversity gain ratio to 0.05. We then group the summarization into $|S_a|$ units through k-means clustering. Correspondingly, we recursively construct a binary tree through binary subdivision on the nodes until the tree has $|S_a|$ leaf nodes. We call each leaf node an anchor node, which is associated with one canvas subregion. Then, each anchor node is binarily divided further until each leaf node contains one image. The construction processes of both stages are the same, whereas each state has different leaf nodes. The construction in a two-stage manner can group the images with similar content together. Thus, a particular layout structure can be described by a tree, where each node represents a region formed by a series of spatial divisions. Fig. 3 shows the construction process. In contrast to [41], which splits each region represented by each binary tree node into two equal parts by *horizontal* or *vertical cut*, we adopt an adaptive strategy to split the whole canvas, thus avoiding image aspect ratio distortion.

Image Assignment. With an initial layout, we assign the parent and leaf nodes one unit and one image, respectively, considering the fitness of size, aspect ratio, content, and

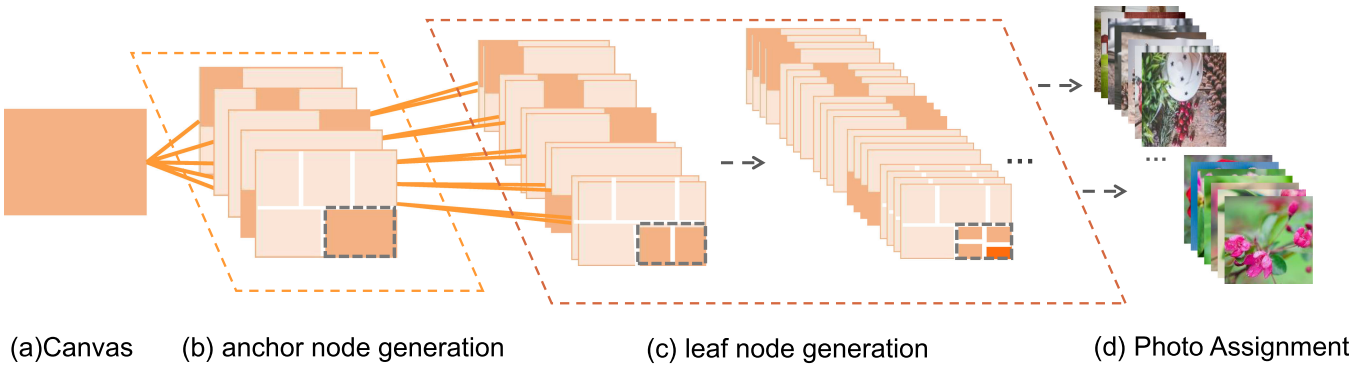


Fig. 3. Illustration of our layout initialization method. The root node (a) is divided into $|S_a|$ anchors nodes based on the anchor images in the first stage (b) of tree construction. In the second stage (c), each anchor node is recursively split by binary partition until (d) each leaf node is only associated with one image.

aesthetics globally. We model the above-mentioned process as an assignment problem and obtain the optimal result through the Hungarian algorithm [47]. Given that K is small in our problem, the solution can be efficiently found. Fig. 2c shows a random initial layout. Such a layout suffers from various problems, such as severe distortion and failure to fit inside the image collection summarization, and it is visually peculiar. Thus, we propose a two-step coarse-to-fine scheme to optimize the layout further.

5.3 Layout Optimization

The optimization of Equation (7) is difficult to solve due to the nonlinear nature. Therefore, we propose a two-step process to optimize the problem. First, we coarsely adjust each tree node by adjusting the division way, splitting ratio, and number of images to contain, etc. We apply a back propagation algorithm to our tree to optimize the result further based on the training process of the neural network. The proposed coarse-to-fine scheme formulates various constraints, such as the image size, aspect ratio, and position, as a series of costs and utilizes a two-step process to achieve an optimized collage.

5.3.1 Coarse Adjustment

The visual effect of the initial layout is far from our expectation because it is randomly generated. There are two steps to generate a collage, i.e., generating a layout and assigning an image to each cell. We propose an alternating adjustment algorithm. Specifically, the division types (horizontal,'H' or vertical,'V'), division ratio (α), number of images contained in subregions (N_{exp}), and expected aspect ratio (ar_{exp}) are adjusted on the basis of a previous collage. Thereafter, an image is reassigned to each cell of the layout by utilizing the Hungarian algorithm [47], thereby producing a new collage.

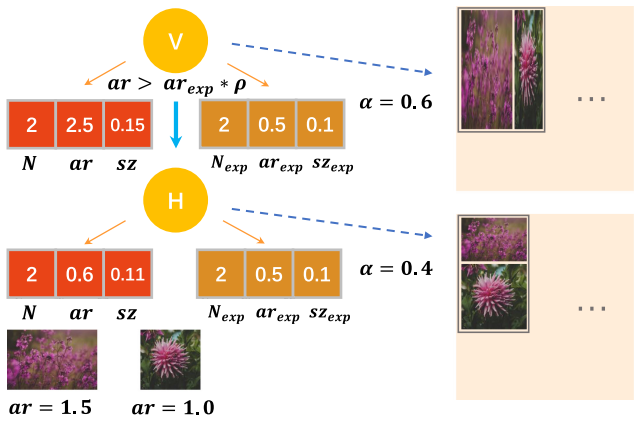


Fig. 4. Illustration of the coarse adjustment of one node. Given that the condition ($ar > ar_{exp} \cdot \rho$) is met, we change the division type from $Vertical$ to $Horizontal$ and adjust α according to Equation. (15) The node area is accordingly adjusted when the entire tree updates.

After an initial collage is generated, each binary tree node contains some attributes, such as $N_{exp}, ar_{exp}, sz_{exp}$ and N, ar, sz. $(\cdot)_{exp}$ indicates the expected properties, whereas those without exp represent the real ones. We first need to adjust the splitting direction to reduce the gap between the expected values and the real ones. In our work, we adopt a method similar to that of [41], that is, if the $ar > ar_{exp} \cdot \rho$, then we change the splitting direction of this node to H. Otherwise, we change to V. ρ is a threshold parameter.

We adjust the division ratio α based on the splitting method, as follows:

$$\alpha = \alpha_{ar} + \alpha_{sz}, \tag{15}$$

where α_{ar} and α_{sz} are the adjusted values based on the child nodes' real aspect ratios and sizes. With the adjusted splitting method and division ratio, the expected sz_{exp} and ar_{exp} of the child nodes can be updated. Considering that the number of images associated with a node greatly influences the node size, we adjust the N_{exp} based on only α_{sz}, which is different from the adjustment strategy of sz_{exp} and ar_{exp}. We truncate the subtree in which the N_{exp} and N of the node are different and then regenerate it as expected according to $N_{exp}, ar_{exp}, sz_{exp}$. Fig. 4 illustrates the coarse adjustment of one node. Given that the condition ($ar > ar_{exp} \cdot \rho$) is met, we change the division type from V to H and adjust α according to Equation (15). Accordingly, the area of this node is adjusted when the entire tree updates.

After one adjustment through a tree, we reassign units to each anchor node and images to each leaf node using the Hungarian algorithm and then perform another iteration until the termination criteria are met. Corresponding to the two-stage constructed tree, we adjust the anchor nodes and leaf nodes alternately by adjusting one and fixing the other.

5.3.2 Refinement Through Optimization

Fig. 2 shows the following three issues of the adjusted collages: 1) each image may suffer from ratio distortion; 2)

certain images may have size bias; and 3) the entire layout may look visually unbalanced. Thus, we further present a refinement step via optimization. In contrast to previous methods, we propose the use of the back propagation algorithm to optimize a tree inspired from the learning procedure of the deep neural network and, thus, achieve the final satisfactory collage.

Given the aspect ratio and width of a collage, we can calculate the width and aspect ratio of each cell associated with one image according to the splitting method S and the splitting ratio α of each node. We must delicately optimize the α of each node to optimize the collage further. Meanwhile, the cell height can be uniquely determined in accordance with cell width w and α. The issue of placing images in a content-aware and aesthetic manner is roughly achieved using the Hungarian algorithm. Our refined optimization operation mainly focuses on image ratio preservation and size diversification. We must calculate the partial derivative of each node by using the chain rule to utilize the back propagation algorithm. Given that the $\max$ item in Equation (8) is not differentiable, we replace it with the Soft-Max operator. The partial differential values of C_{sh} and C_{sz} with respect to the ar_{exp} of the node in depth i are calculated as follows:

$$\begin{aligned}\frac{\partial C_{sh}}{\partial ar^i_{exp}} &= \frac{\partial C_{sh}}{\partial ar_{exp}} \cdot \frac{\partial F_n}{\partial F_{n-1}} \cdot \frac{\partial F_{n-1}}{\partial F_{n-2}} \cdots \frac{\partial F_{i+1}}{\partial ar^i_{exp}},\\ \frac{\partial C_{sz}}{\partial ar^i_{exp}} &= \frac{\partial C_{sz}}{\partial ar_{exp}} \cdot \frac{\partial H_n}{\partial H_{n-1}} \cdot \frac{\partial H_{n-1}}{\partial H_{n-2}} \cdots \frac{\partial H_{i+1}}{\partial ar^i_{exp}},\end{aligned} \tag{16}$$

where ar^i_{exp} is the ar_{exp} of the node in depth i; and F_i and H_i are the expected aspect ratio and width of the node in the ith layer, respectively. The arbitrary partial derivative of cost functions C_{sh} and C_{sz} with respect to α in depth i can be obtained as follows:

$$\begin{aligned}\frac{\partial (C_{sh}+C_{sz})}{\partial \alpha^i} &= \frac{\partial C_{sh}}{\partial ar_{exp}} \cdot \frac{\partial F_n}{\partial F_{n-1}} \cdot \frac{\partial F_{n-1}}{\partial F_{n-2}} \cdots \frac{\partial F_{i+1}}{\partial \alpha^i}\\ &\quad + \frac{\partial C_{sz}}{\partial ar_{exp}} \cdot \frac{\partial H_n}{\partial H_{n-1}} \cdot \frac{\partial H_{n-1}}{\partial H_{n-2}} \cdots \frac{\partial H_{i+1}}{\partial \alpha^i}\\ &= \frac{\partial C_{sh}}{\partial ar^{i+1}_{exp}} \cdot \frac{\partial F_{i+1}}{\partial \alpha^i} + \frac{\partial C_{sz}}{\partial ar^{i+1}_{exp}} \cdot \frac{\partial H_{i+1}}{\partial \alpha^i}.\end{aligned} \tag{17}$$

Based on the gradients of α of each node, we use the "momentum" method by referring to deep learning to update α. Fig. 2d shows an optimized collage. The detailed derivations in this section are presented in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2019.2948611, due to the page limit.

6 Experiments and Evaluations

We collect more than 3500 images from 22 users' photo galleries in Pexels[1] to validate our method. The topics of these galleries involve landscape, food, animals, and selfies, etc. All images are represented by features extracted from the penultimate layer of VGG-16, which has been pretrained on ImageNet [48] and Places365 [49]. The algorithms independently run on each gallery with a desktop PC with 4.2 GHZ Intel Core i7 and 16 GB RAM. The optimization time depends on the number of images and the parameter settings. The optimization time in our experiments ranges from 1 s to 30 s for approximately 150 images. The dataset is available at http://ivc.ia.ac.cn/VisualSum/.

6.1 Image Summarization

Baseline. Based on our summarization method, we compare the summarization results of our method with the following three baseline methods: *random* selection; k-means clustering, which selects the closest k images to the cluster centers; and k-means-D, which selects the most diverse image as the representative one from each cluster.

Metrics. Four metrics are adopted to evaluate the summarization quality, as follows:

- D_{DR}, diversity ratio. Based on the definition in Section 4.1, D_{div} measures the diversity of the summarization by considering the variance, balance, and disparity. This metric is defined as follows:

$$D_{DR}(\mathcal{I}, S) = \frac{E_{div}(S)}{E_{div}(\mathcal{I})}. \tag{18}$$

- D_{JS}, Jensen Shannon divergence. This approach is a popular mechanism used to compare the information conveyed by two probability distributions and is defined as follows:

$$D_{JS}(X||Y) = \frac{1}{2}(D_{KL}(X||M) + D_{KL}(Y||M)), \tag{19}$$

  where X and Y are discrete p.d.f, $M = \frac{1}{2}(X+Y)$ is the mean distribution, and $D_{KL}(X||Y) = \sum_i X(i) log(\frac{X(i)}{Y(i)})$ is the KL divergence between X and Y. We model the original image collection X and a candidate summarization Y as probability distributions over a multidimensional topic space.
- D_{Re}, reconstruction error. A good summarization is expected to include the typical patterns of the collections. Thus, we define D_{Re} as follows:

$$D_{Re}(\mathcal{I}, S) = \frac{1}{N}\sum_i R(I_i, S), \tag{20}$$

  where $\mathcal{I}$ and S are the image collections with N images and corresponding summarization. $R(I_i, S)$ is similar to that in Equation (4).
- D_{Rep}, representativeness. Based on the formulation in Section 4, we evaluate the summarization considering all three criteria and expose the selection process with an increasing size. Thus, we define representativeness as the weighted summation of three criteria by the following equation:

$$D_{Rep}(S, \mathcal{I}) = \frac{1}{3}(E_{con}(S, \mathcal{I}) + \frac{E_{div}(S)}{E_{div}(\mathcal{I})} + E_{aes}(S)). \tag{21}$$

For content-based summarization, hidden topic analysis of the image object and scene information is necessary. The following two datasets are involved for pretraining:

1. https://www.pexels.com

TABLE 1
Evaluation Results of the Summarization when we Extract Features by Using CNN Models Pre-Trained on Different Image Sets by Selecting 30 Percent of the Images From Each Gallery

ImageNet	*Places365*	D_{DR}	D_{JS}	D_{Re}	D_{Rep}
✓		0.636	0.031	0.793	0.703
	✓	0.401	0.101	0.993	0.503
✓	✓	**0.831**	**0.012**	**0.495**	**0.851**

ImageNet [48] is used for extracting the image object information and Places365 [49] is used for extracting the scene information. We compare the influence of considering different types of information. Table 1 shows the performance of using different datasets; it also shows that the object information is usually more important than the scene, and the result is optimal when both data are considered.

In this work, we use the diversity score E_{div}, concise ratio E_{con}, and aesthetics score E_{aes} as criteria for the gallery summarization. Table 2 shows quantitative comparisons regarding the use of different criteria when selecting 30 percent of the images from each gallery. Among these three criteria, E_{div} plays the most important role, whereas E_{aes} performs the least. However, only considering E_{div} and E_{con} without E_{aes} (as in [19], [20], [21]) will lead to low D_{Rep}. E_{aes} helps remove low-quality images, which usually contain less information and much noise. Such a score can dramatically improve the representativeness.

Fig. 5 shows a qualitative comparison of the results. Our result in Fig. 5d contains images with a great variety of species, scenes, and compositions, thereby indicating more information compared with that of Figs. 5a, 5b, and 5c. With the lowest diversity, Fig. 5a contains two portraits and the third and fourth images have similar content. The summarization using k-means-D selects the most representative image from each cluster center compared with the summarization by k-means. Most of the images in Fig. 5c are visually more informative than those in Fig. 5b. Our result selects the most representative subset from the gallery one by one.

Fig. 6 shows the quantitative comparison results. The figure plots the averaged results across all galleries. The x-axis represents the summarization size, and the y-axis represents the diversity ratio (as a fraction of the original collection), JS divergence, reconstruction error, and representativeness score. With the increase in the number of images in the summarization, the JS divergence and reconstruction error decrease, whereas the diversity increases. D_{Rep} indicates the representativeness of a summarization

TABLE 2
Evaluation Results When we Considered Different Criteria by Selecting 30 Percent of the Images From Each Gallery

E_{div}	E_{con}	E_{aes}	D_{DR}	D_{JS}	D_{RE}	D_{Rep}
✓			0.732	0.023	0.621	0.579
	✓		0.625	0.089	0.776	0.567
		✓	0.601	0.099	0.821	0.571
✓	✓		0.815	0.016	0.571	0.627
✓		✓	0.799	0.020	0.601	0.634
	✓	✓	0.655	0.074	0.718	0.631
✓	✓	✓	**0.831**	**0.012**	**0.495**	**0.851**

(a) Random selection. D_{DR}: 0.328; D_{JS}: 0.195; D_{Re}: 1.31; D_{Rep}: 0.436

(b) k-means. D_{DR}: 0.467; D_{JS}: 0.119; D_{Re}: 1.01; D_{Rep}: 0.512

(c) k-means-D. D_{DR}: 0.513; D_{JS}: 0.092; D_{Re}: 0.949; D_{Rep}: 0.539

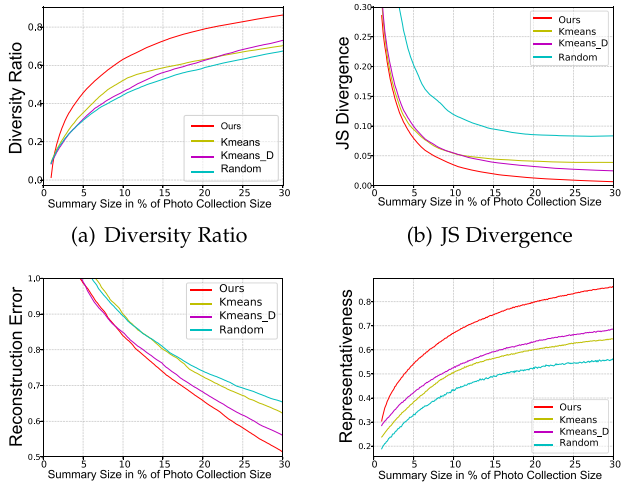
(d) Our result. D_{DR}: 0.651; D_{JS}: 0.071; D_{Re}: 0.881; D_{Rep}: 0.721

Fig. 5. Qualitative comparison of our summarization method and the three baseline methods on Jane He's gallery, which contains 68 photos.

and increases as additional images are selected. A considerable increase/decrease also occurs when the image size changes from zero to ten, whereas the rate of change slows down after 10. In all the cases, our result outperforms the baselines. In addition, k-means-D outperforms the conventional k-means clustering, indicating that the use of *diversity* improves the representativeness of the summarization. Although k-means clustering implicitly minimizes the reconstruction error in the mechanism, our method also performs better because the performance of k-means clustering strongly depends on the initiation.

6.2 Tree-Based Layout Generation

Baseline. We select three widely used photo collage methods as the baseline. These methods are commercial software packages ShpCollage [50], PicCollage [30], and Photo Recomposing collage (PRCollage) [18]. Fig. 7 shows the results generated by our method and those of the baselines. We deliberately set the same white space between images to be consistent with ShpCollage. See the collages with no gap between images in the supplementary materials, available online.



(a) Diversity Ratio (b) JS Divergence

(c) Reconstruction Error (d) Representativeness

Fig. 6. Quantitative comparison results of our summarization method and the three baseline methods by using the four metrics.

(a) ShpCollge [50] (b) PicCollage [30] (c) PRCollage [18] (d) Ours

Fig. 7. Comparison of the layout results generated by the different methods. Our result is visually pleasing and can effectively convey the information.

Metrics. We analyze the results according to four metrics that are commonly considered in state-of-the-art photo collage methods as follows:

- *Compactness*. A compact and space-efficient layout is assumed to be the optimal one. This metric is measured by the empty space ratio $M_c = S_w/S$, where S_w and S are the white space and the total collage sizes, respectively.
- *Ratio preservation*. Given the target shape and size of the canvas, we should retain the shape of the images. Specifically, the images should be shown in their original aspect ratios. The distortion of the aspect ratio is measured by $M_r = \frac{1}{K}\sum_k^K \max(r_k/r_k^L, r_k^L/r_k)$, where r_k is the aspect ratio of the original image and r_k^L is the image in the collage.
- *Correlation preservation*. Images with similar content should be placed together to facilitate informativeness. This metric is measured by $M_n = \sum_c^C \sum_i (P_i - P_c)^2$, where P_i indicates the position of the image in the collage, and P_c is the position of the cluster of the cth group, which contains images with similar content. The positions range from [0,1] by dividing the width or height of the collage.
- *Nonoverlapping constraint*. Image overlapping often decreases the representativeness and aesthetics of a

TABLE 3
Comparison of our Layout and the Three Baselines

	M_c	M_r	M_n	M_o
ShpCollage [50]	**0**	1.281	0.892	**0**
PicCollage [30]	**0**	-	0.810	0.2107
PRCollage [18]	0.089	1.231	0.693	**0**
Ours	**0**	**1.182**	**0.3132**	**0**

M_c, M_r, M_n, and M_o measure the layout quality with respect to the compactness, ratio preservation, correlation preservation and overlapping. Small values indicate high layout quality.

collage. Overlapping is avoided by minimizing $M_o = S_o/S$, where S_o is the sum of the intersection areas of any two images.

Table 3 shows the comparison of our method and the baseline ones by using the four metrics described above. The white space in ShpCollage and our layout are deliberately set to increase the visual aesthetics. Given the target shape and size of the canvas, we effectively preserve the image shapes and ratios. By contrast, PicCollage dramatically changes the shape of the original images due to the image overlap. Compared with our method, PicCollage suffers from severe image overlapping, thereby resulting in serious information loss. Although ShpCollage has no occlusion between images, great image cropping also occurs in the layout, thereby discarding meaningful information. PRCollage generates a structural layout by using a Voronoi tree map, but the performance in terms of aspect ratio preserving is worse than ours. We calculate the ratio preservation according to the structural layout instead of the images for PRCollage. With regard to the correlation preservation, we group the images with similar content together and ensure the balance of visual aesthetics, color distribution, and layout structure. ShpCollage and PicCollage consider no correlation preservation, thereby resulting in the mixing of images with different themes together. PRCollage generates a collage, which places images with a similar composition close. In our opinion, such an approach cannot considerably improve the informativeness of the collage described in Section 6.3.

6.3 User Study

We conduct three user studies to evaluate the effectiveness of our algorithm. In the first experiment, we ask participants to compare different summarizations in terms of representativeness. In the second one, the subjects are required to rate different collages according to their aesthetics and informativeness. In the last one, we design an effective and interactive online website to excavate the collage's ability to convey information accurately.

Summarization Representativeness. For each image collection (containing N images) in the testing set, we use our method and k-means clustering to generate two summarizations separately with sizes of $K = 50$ and $K = 30\% \cdot N$. Fifteen participants evaluate the representativeness of summarizations. For each pair of summarization results, the participants are asked to browse the entire collection for more than 30 s and then select the one that they think is representative. The participants are also asked to judge whether the summarization they select is sufficiently representative of the entire collection.

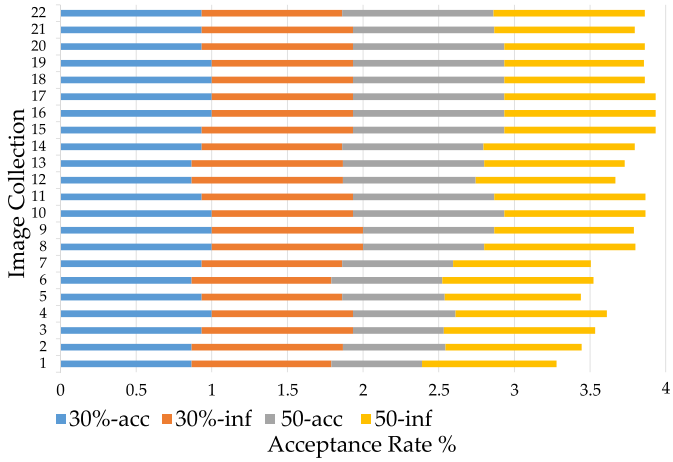


Fig. 8. Results of the user study about the visual summarization of the image collection. The galleries are ordered in accordance with their sizes.

The statistics show that 95.13 percent of the subjects prefer our summarization, among which, 98.01 percent think that our summarizations are sufficiently representative. For summarizations with $K = 30\% \cdot N$, in 96.58 percent of the cases, the participants accept our summarizations as the optimal ones, and 97.88 percent think that the summarizations are sufficiently representative. For summarizations with $K = 50$, the acceptance rates of our method are slightly low in galleries with few images. With the increase in N, the summarizations of $K = 50$ gradually approach that with $K = 30\% \cdot N$. As shown in Fig. 8, the y-axis represents the 22 collections sorted by the dataset size, and the acceptance ratio of the summarization with $K = 30\% \cdot N$, called 30%-acc, is relatively stable, and all are above 85 percent. The summarizations with $K = 50$, namely, 50-acc, gradually rises with an increase in N. This occurrence is attributed to the first few collections, which contain a small number of images (slightly more than 50), resulting in the summarizations with both constraints possessing almost the same images, thus confusing the participants. The rates of the cases in which people think that our summarizations are sufficiently representative (30%-rep and 50-rep) are all above 90 percent. This finding indicates that our summarizations can effectively cover the main information of the original image collection. In conclusion, our method can generally yield good and acceptable results in our user study.

Collage Aesthetics and Informativeness. For each image collection, we select 30 percent of the images using our summarization method and generate four collages using our layout method and the three baseline methods. Fifteen participants are first asked to rate each collage from one to nine in terms of the visual aesthetics. Thereafter, participants assess the extent to which a collage presents the amount of information in the summarization and evaluate the informativeness by ranking them from one to nine. The participants are asked to watch the collage for at least 20 s. Fig. 9 shows the results. Our method outperforms the comparative methods in terms of aesthetics and informativeness. The score distributions of the four collages for aesthetics are relatively concentrated, with standard deviations of approximately 0.4. This finding indicates that the results are convincing to a certain degree. By contrast, the distributions for informativeness are more

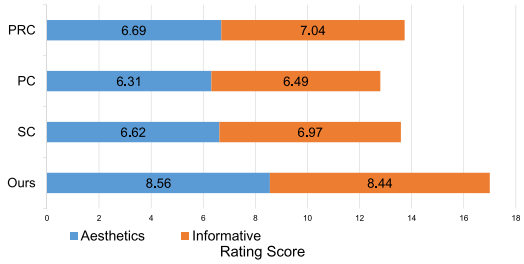


Fig. 9. Results of the user study about visual aesthetics and informativeness.

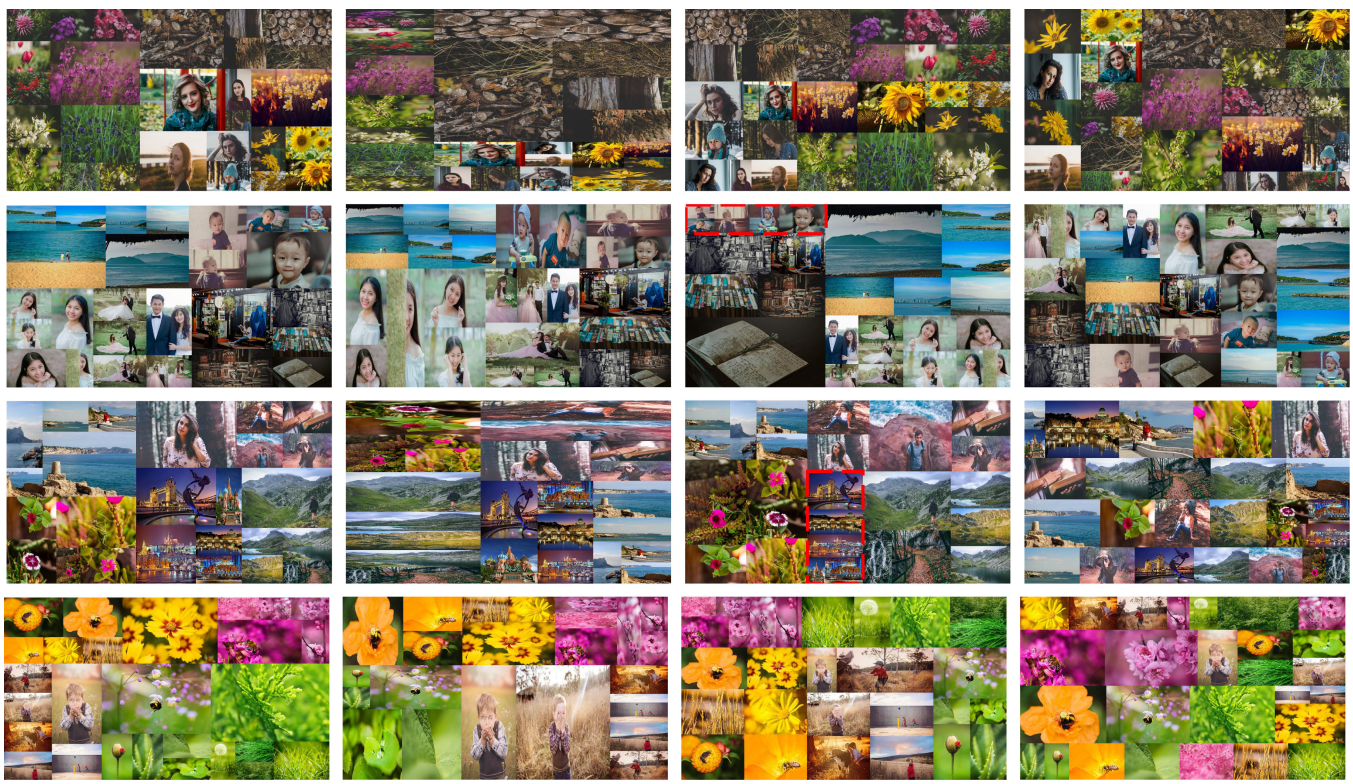
scattered. PicCollage obtains the lowest standard deviation of approximately 0.6, whereas ShpCollage and PRCollage obtain slightly higher values (i.e., 1.2 and 1.3, respectively). This outcome is attributed to the ability to transmit collage information, which is hardly generalized through a general indicator, and confuses participants. Thus, we conducted another test to accurately evaluate the informativeness.

Information Conveying. Twenty participants are asked to perform a complex test for evaluating the informativeness and ability to convey the information of the collages from a specific and deep perspective. We show participants a collage for 20 s and then ask them to select the images that they remember from a testing image set containing 10 images. We only involve the same user in investigating one collage from each collection to avoid inconsistency in the time the same picture is viewed by the same participant.

TABLE 4
Evaluation Results of the Informativeness of the Four Layouts

Evaluation	Ours	SC [50]	PC [30]	PRC [18]
Recall	**0.946**	0.793	0.749	0.801
Precision	**0.863**	0.692	0.662	0.690
Accuracy	**0.732**	0.560	0.538	0.552
F1-Score	**0.903**	0.739	0.708	0.738

Each testing image set shown to participants is randomly selected from the summarization (seven images) and the remaining images of the original collection (three images). The four participants corresponding to the four collages are assigned to the same candidate image collection. We count four indicators, namely, *Recall*, *Precision*, *Accuracy* and *F1-Score*, to evaluate the informativeness of the collages. Table 4 shows the results. Our collage outperforms the other baselines in all four indicators. PicCollage is the worst, indicating that overlapping seriously reduces the information delivery ability. Similarly, ShpCollage and PRCollage show that the ratio preserving and correlation preservation are beneficial to the delivery of information compared with our methods. The ratio of images selected by participants is approximately 3:1, indicating that the participants are inclined to choose more images as remembered, leading to a higher *Recall* than *Precision*. *Accuracy* is a rigorous measurement in our experiment because it is difficult to



(a) Results with all constraints (b) w/o ratio preservation (c) w/o size diversification (d) w/o location and aesthetics

Fig. 10. Visual comparisons between the generated collages with/without different types of constraints. Without the ratio preservation constraint, the images in the layout (b) suffer from severe ratio distortion. Without the size diversification constraint, the sizes of images in the layout (c) are mainly based on its original size. Some images (highlighted by red dashed boxes) are extremely small and attracting attention to these images without a conspicuous mark is difficult. Without the content-aware and aesthetics-embedded location constraint, the images in the layout (d) basically retain the original ratio and the representative ones are enlarged to some extent. However, the collages look cluttered because the content, color distribution, and the perception balance are neglected. Please refer to Section 6.4 for a detailed discussion.

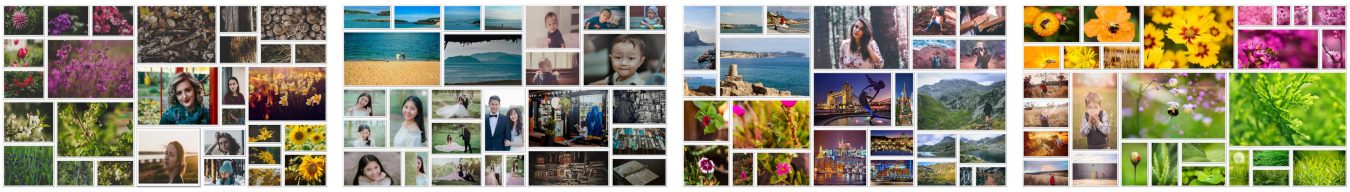


Fig. 11. Our image collage results with white space.

effectively determine whether it is the same as previously seen because several similar pictures exist in the candidate image set. We introduce the *f1-score* to obtain an accurate evaluation, thereby showing that our method is much better than the others. As discussed in [51], the visual human perception system works by first recognizing the "gist" of the image almost instantaneously, from a single glance (200 ms). Thus, we ignore the difference in the difficulty to remember among the images.

6.4 Discussions

Our proposed approach for tree-based visualization contains three items, namely, the image ratio preservation, image size diversification, and content-aware and aesthetics-embedded image location, as defined in Equation (7). We neglect certain constraints and show the results in Fig. 10. As shown in Fig. 10b, the results suffer from severe ratio distortion, but similar images are still intact, and the representative images are set to be displayed larger than the other images. The ratios of images in these four collages have only been optimized by a coarse adjustment step described in Section 5. As shown in Fig. 10c, without the size diversification constraint, the sizes of images in the collages are mainly the original size, thereby resulting in information loss. As shown in the red dashed boxes, the image sizes are small. Attracting attention without a conspicuous mark is difficult. In addition, quickly and clearly capturing the main content with a basically consistent size is more difficult than the methods shown in Fig. 10a. As shown in Fig. 10d, without the content-aware and aesthetics-embedded location constraint, the images basically retain the original ratio and the representative ones are enlarged to some extent. However, the collages look cluttered because the content, color distribution, and the perception balance are neglected. Consequently, information conveying becomes difficult and ineffective. Fig. 11 shows the visual summarization results with white space added between the images.

Developing a general method to produce an informative and beautiful visual summarization for an image collection is often complicated because information dissemination and visual aesthetics should be simultaneously integrated. In this work, our proposed method can efficiently compute results and achieve attractive visual summarizations. However, our approach is subject to several limitations. First, given that each node is divided into two parts and the shape of each region is a rectangle, the division of the canvas may look insufficiently symmetrical in certain cases, especially in that it rarely reaches rotational symmetry. Additionally, our method can only work for rectangular canvases currently. Second, without considering the story attributes of the entire collection, our hidden topic-based summarization method is sometimes ineffective at storytelling. Our result may be unable to capture the storyline, which is difficult to formulate. In consideration of our image layout being seamless, several images may sometimes be distorted in the layout. This problem can be simply solved by using cropping or content-aware image retargeting methods [52].

7 Conclusion

In this work, we develop a content-based approach to generate a visual summarization for image collections automatically. Our key contributions include a hidden topic-based image collection summarization algorithm, a tree-based image layout generation method, and a layout optimization algorithm using gradient back propagation. Our experimental results, ablation study, and user study demonstrate the efficiency and effectiveness of our approach.

Currently, the proposed approach mainly focuses on the visual summarization in a rectangle form by dividing the canvas into multiple sub-rectangles. In the future, we will explore more types of layouts (such as the Voronoi-tree layout) on different shapes of canvases. By considering more image features (such as the color and structure distribution of the images), a data-driven approach will also be explored instead of a rule-based page partition to achieve flexible and beautiful visual summarization results. Specifically, we will learn the hidden knowledge from the limited satisfactory layouts as the training data that will generate the layout structure. Furthermore, narrating a story through a compact collage is another interesting and challenging issue. We plan to investigate a narrative mode, which does not depend on the spatial-temporal information of an image collage.

Acknowledgments

We thank the anonymous reviewers for valuable comments and Yuan Liang for preparing some comparison results. This work was supported by National Key R&D Program of China under no. 2018YFC0807500, and by National Natural Science Foundation of China under nos. 61832016, 61672520 and 61702488, and by Ministry of Science and Technology under no. 108-2221-E-006-038-MY3, Taiwan and by CASIA-Tencent Youtu joint research project.

References

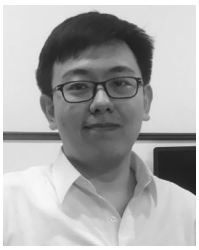
[1] L. Tan, Y. Song, S. Liu, and L. Xie, "ImageHive: Interactive content-aware image summarization," *IEEE Comput. Graphics Appl.*, vol. 32, no. 1, pp. 46–55, Jan./Feb. 2012.

[2] C. Yang, J. Shen, J. Peng, and J. Fan, "Image collection summarization via dictionary learning for sparse representation," *Pattern Recognit.*, vol. 46, no. 3, pp. 948–961, 2013.

[3] K. Plicanic, *Album Moxie: The Savvy Photographer's Guide to Album Design and More with InDesign*. San Francisco, CA, USA: Peachpit Press, 2013.

[4] J. Wagemans et al., "A century of Gestalt psychology in visual perception I. Perceptual grouping and figure-ground organization," *Psychological Bull.*, vol. 138, no. 6, pp. 1172–1217, Nov. 2012.
[5] E. Y. Kim and E. Ko, "Canonical image selection based on human affects in photographic images," *Image Vis. Comput.*, vol. 54, pp. 83–98, 2016.
[6] H. Chang, F. Yu, J. Wang, D. Ashley, and A. Finkelstein, "Automatic triage for a photo series," *ACM Trans. Graphics*, vol. 35, no. 4, pp. 148:1–148:10, Jul. 2016.
[7] Y. Cong, J. Liu, G. Sun, Q. You, Y. Li, and J. Luo, "Adaptive greedy dictionary selection for web media summarization," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 185–195, Jan. 2017.
[8] J. E. Camargo and F. A. González, "Multimodal latent topic analysis for image collection summarization," *Inf. Sci.*, vol. 328, pp. 270–287, 2016.
[9] P. Brivio, M. Tarini, and P. Cignoni, "Browsing large image datasets through Voronoi diagrams," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 6, pp. 1261–1270, Nov./Dec. 2010.
[10] 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.
[11] L. Liu, H. Zhang, G. Jing, Y. Guo, Z. Chen, and W. Wang, "Correlation-preserving photo collage," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 6, pp. 1956–1968, Jun. 2018.
[12] E. Gomez-Nieto et al., "Similarity preserving snippet-based visualization of web search results," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 3, pp. 457–470, Mar. 2014.
[13] E. Gomez-Nieto, W. Casaca, D. Motta, I. Hartmann, G. Taubin, and L. G. Nonato, "Dealing with multiple requirements in geometric arrangements," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 3, pp. 1223–1235, Mar. 2016.
[14] E. Gomez-Nieto, W. Casaca, L. G. Nonato, and G. Taubin, "Mixed integer optimization for layout arrangement," in *Proc. 26th Conf. Graphics Patterns Images*, 2013, pp. 115–122.
[15] A. Lau and A. Vande Moere, "Towards a model of information aesthetics in information visualization," in *Proc. 11th Int. Conf. Inf. Vis.*, 2007, pp. 87–92.
[16] A. Vande Moere and H. Purchase, "On the role of design in information visualization," *Inf. Vis.*, vol. 10, no. 4, pp. 356–371, Oct. 2011.
[17] S. Bianco and G. Ciocca, "User preferences modeling and learning for pleasing photo collage generation," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 12, no. 1, pp. 6:1–6:23, Aug. 2015.
[18] Y. Liang, X. Wang, S. H. Zhang, S. M. Hu, and S. Liu, "PhotoRecomposer: Interactive photo recomposition by cropping," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 10, pp. 2728–2742, Oct. 2018.
[19] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 297–306.
[20] P. Sinha, S. Mehrotra, and R. Jain, "Summarization of personal photologs using multidimensional content and context," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, 2011, Art. no. 4.
[21] G. Kim, L. Sigal, and E. P. Xing, "Joint summarization of large-scale collections of web images and videos for storyline reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 4225–4232.
[22] A. Lidon, M. Bolaños, M. Dimiccoli, P. Radeva, M. Garolera, and X. Giro-i-Nieto, "Semantic summarization of egocentric photo stream events," in *Proc. 2nd Workshop Lifelogging Tools Appl.*, 2017, pp. 3–11.
[23] 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 Psychologica*, vol. 99, no. 2, pp. 141–161, 1998.
[24] D. C. L. Ngo, A. Samsudin, and R. Abdullah, "Aesthetic measures for assessing graphic screens," *J. Inf. Sci. Eng.*, vol. 16, no. 1, pp. 97–116, 2000.
[25] J. Geigel and A. C. Loui, "Automatic page layout using genetic algorithms for electronic albuming," in *Internet Imaging II*. Bellingham, WA, USA: SPIE, 2000, pp. 79–91.
[26] S. Lok, S. Feiner, and G. Ngai, "Evaluation of visual balance for automated layout," in *Proc. 9th Int. Conf. Intell. User Interfaces*, 2004, pp. 101–108.
[27] 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. 33:1–33:22, Feb. 2016.
[28] R. Hübner and M. G. Fillinger, "Comparison of objective measures for predicting perceptual balance and visual aesthetic preference," *Frontiers Psychol.*, vol. 7, 2016, Art. no. 335.
[29] J. Wang, L. Quan, J. Sun, X. Tang, and H.-Y. Shum, "Picture collage," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 347–354.
[30] T. Liu, J. Wang, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Picture collage," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1225–1239, Nov. 2009.
[31] S. Ma and C. W. Chen, "Automatic creation of magazine-page-like social media visual summary for mobile browsing," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2016, pp. 469–473.
[32] Y. Wei, Y. Matsushita, and Y. Yang, "Efficient optimization of photo collage," Tech. Rep. MSR-TR-2009-59, May 2009. [Online]. Available: https://www.microsoft.com/en-us/research/publication/efficient-optimization-of-photo-collage/
[33] Y. Yang, Y. Wei, C. Liu, Q. Peng, and Y. Matsushita, "An improved belief propagation method for dynamic collage," *Vis. Comput.*, vol. 25, no. 5, pp. 431–439, 2009.
[34] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," *ACM Trans. Graphics*, vol. 25, no. 3, pp. 847–852, Jul. 2006.
[35] G. P. Nguyen and M. Worring, "Interactive access to large image collections using similarity-based visualization," *J. Vis. Languages Comput.*, vol. 19, no. 2, pp. 203–224, 2008.
[36] T. Dwyer, K. Marriott, and P. J. Stuckey, "Fast node overlap removal," in *Proc. Int. Symp. Graph Drawing*, 2005, pp. 153–164.
[37] E. R. Gansner and Y. Hu, "Efficient node overlap removal using a proximity stress model," in *Proc. Int. Symp. Graph Drawing*, 2008, pp. 206–217.
[38] H. Strobelt, M. Spicker, A. Stoffel, D. Keim, and O. Deussen, "Rolled-out wordles: A heuristic method for overlap removal of 2D data representatives," *Comput. Graphics Forum*, vol. 31, no. 3pt3, pp. 1135–1144, 2012.
[39] S. Liu et al., "Composing semantic collage for image retargeting," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5032–5043, Oct. 2018.
[40] Z. Yu, L. Lu, Y. Guo, R. Fan, M. Liu, and W. Wang, "Content-aware photo collage using circle packing," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 2, pp. 182–195, Feb. 2014.
[41] 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.
[42] M. Chen, F. Xu, and L. Lu, "Manufacturable pattern collage along a boundary," *Comput. Vis. Media*, vol. 5, no. 3, pp. 293–302, Sep. 2019.
[43] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 3319–3327.
[44] A. Stirling, "A general framework for analysing diversity in science, technology and society," *J. Roy. Soc. Interface*, vol. 4, no. 15, pp. 707–719, 2007.
[45] 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. 26th ACM Int. Conf. Multimedia*, 2018, pp. 879–886.
[46] D. Cohen-Or, O. Sorkine, R. Gal, T. Leyvand, and Y.-Q. Xu, "Color harmonization," *ACM Trans. Graphics*, vol. 25, no. 3, pp. 624–630, 2006.
[47] E. L. Lawler, *Combinatorial Optimization: Networks and Matroids*. North Chelmsford, MA, USA: Courier Corporation, 1976.
[48] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[49] B. Zhou, A. 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.
[50] V. Cheung, "Shape collage," 2013. [Online]. Available: https://shapecollage.com/
[51] 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.
[52] J. Kiess, S. Kopf, B. Guthier, and W. Effelsberg, "A survey on content-aware image and video retargeting," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 3, pp. 76:1–76:28, Jul. 2018.

Xingjia Pan received the BSc degree in automation and finance from Nankai University, Nankai, China in 2015. He is currently working toward the PhD degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include computer graphics, computer vision, and machine learning.



Fan Tang received the BSc degree in computer science from North China Electric Power University, Beijing, China in 2013, and the PhD degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China in 2019. He is currently a researcher with Fosafer. His research interests include image synthesis and image recognition.

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

Chongyang Ma received the BS degree from the Fundamental Science Class (mathematics and physics), Tsinghua University, Beijing, China in 2007, and the PhD degree in computer science from the Institute for Advanced Study, Tsinghua University, in 2012. He is currently a research lead with Kwai Inc. His research interests include computer graphics and computer vision.

Yiping Meng received the BSc degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, China in 2013, and the MEng degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese of Sciences, Beijing, China in 2017. She is currently a research outreach manager with Didi Chuxing. Her research interests include image synthesis and image recognition.

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

Tong-Yee Lee received the PhD degree in computer engineering from Washington State University, Pullman, United States in May 1995. He is currently a chair professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan, ROC. He leads the Computer Graphics Group, Visual System Laboratory, National Cheng-Kung University (http://graphics.csie.ncku.edu.tw/). His current research interests include computer graphics, non-photorealistic rendering, medical visualization, virtual reality, and media resizing. He is a senior member of the IEEE and a member of the ACM.

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