


# Visual Perception Driven Picture Collage Synthesis

Zuyi Yang, Qinghui Dai, and Junsong Zhang 

© The Author(s) 2015. This article is published with open access at Springerlink.com

**Abstract** A picture collage is a composite artwork made from the spatial layout of multiple pictures on a canvas, which are collected from the Internet or user photographs. Picture collages, usually made by skilled artists, involve a complex manual process, especially when searching for competent pictures and adjusting their spatial layout to meet artistic requirements. Thus, in this paper, we present a visual perception driven method for automatically synthesizing visually pleasing picture collages. Different from previous works, we focus on how to design a collage layout, which not only provides an easier access to the theme of the canvas image, but also conforms to the human visual perception. To achieve this goal, we formulate the generation of picture collages as a mapping problem: given a canvas image, first, compute a saliency map for the canvas and a vector field for each sub-region of the canvas. Second, with a divide-and-conquer strategy, generate a series of patch sets from the canvas, where the salient map and the vector field are used to determine the patch's size and direction respectively. Third, construct a Gestalt principle based energy function to choose the most visually pleasing and orderly patch set as the final layout. Finally, with a semantic-color metric, map the picture set to the patch set and generate the final picture collage. Extensive experimental and user study results show that this method can generate visual pleasing picture collages.

**Keywords** Picture Collage, Gestalt Psychology, Saliency Map, Layout Optimization, Human Visual Perception..

## 1 Introduction

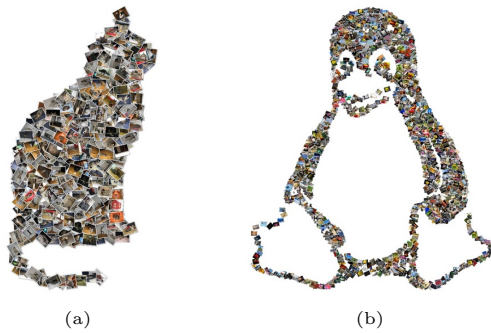
A picture collage (Fig.1) is a composite artwork made from the spatial layout of multiple pictures on a canvas, which are collected from the Internet or user photographs. Given a canvas and a set of picture elements, this art form can flexibly assign each element an appropriate position and size on the canvas, thus showing strong interest and visual appeal. Therefore, it is widely used in design, advertising, rich media creation and many other decorative illustrations.

However, the manual creation of picture collages usually requires rich design experience (Fig.1) and involves an extensive, tedious process, especially when searching for appropriate pictures and creating a visual appealing layout. There are also some automatic synthesizing methods for picture collages, which could be divided into three categories according to our experience: 1) traditional picture collages [1, 2, 6, 12, 14, 15, 21, 22], focused on creating a visual and informative summary from a given image set; 2) picture mosaics [9, 10, 18], adopted multiple sub-pictures to compose an source image and aimed to realize the visual simulation of the source image; 3) photo montages [5, 7], synthesized a new image from several photographs by cutting, gluing, rearranging and overlapping operations, to achieve a visual effect of montages. However, all of these works did not take the human visual perception into consideration as we do. Specifically, compared with the previous works, we consider two more visual perception mechanisms, Gestalt principle and visual attention, to ensure that the synthesized collage layout sufficiently conforms to human visual perception. Though some works [1, 12, 21, 22] adopted visual attention mechanism to extract the salient regions of the sub-images to preserve richer information in the collages, they did not use it to ensure the theme relevance between the collage and the canvas image as we do.

Gestalt psychology principles reflect the strategies of

<sup>1</sup> Department of Artificial Intelligence, Xiamen University, Xiamen, 361005, China. E-mail:2873509318@qq.com, 345020613@qq.com, zhangjs@xmu.edu.cn. Junsong Zhang is the corresponding author.

Manuscript received: 2014-12-31; accepted: 2015-01-30.



**Fig. 1** Picture Collages.

the human visual system about how to group objects into forms and create internal representations for them. Whenever groups of visual elements have one or several characteristics in common, they are combined to form a new larger visual object [16]. In computer vision or graphics, gestalt principles have been applied in various applications, such as image or scene abstractions [19, 20], line or pattern groupers [11, 13, 23] and so on. In this paper, we design a Gestalt-based energy function to guide the layout optimization process, which can ensure that the generated patch set is organized in a more reasonable and visual pleasing way. To the best of our knowledge, this method is the first attempt to apply Gestalt principles to the creation of the picture collages.

Our method steps are as follows: given a canvas image and a picture set, we first compute a saliency map for the canvas and a vector field for each sub-region of it; Second, for each region, along the streamlines of the vector field, we adopt a divide-and-conquer strategy to generate several patch sets and use the saliency value of the location of each patch to control its size; Finally, we design a Gestalt-based energy function to achieve the most visually pleasing and orderly layout, and design a novel semantic-color similarity metric to automatically map pictures into patches to generate the final picture collage. Also, we conduct a variety of image comparison experiments and user studies to evaluate the effectiveness of this method.

The main contributions of this paper are as follows:

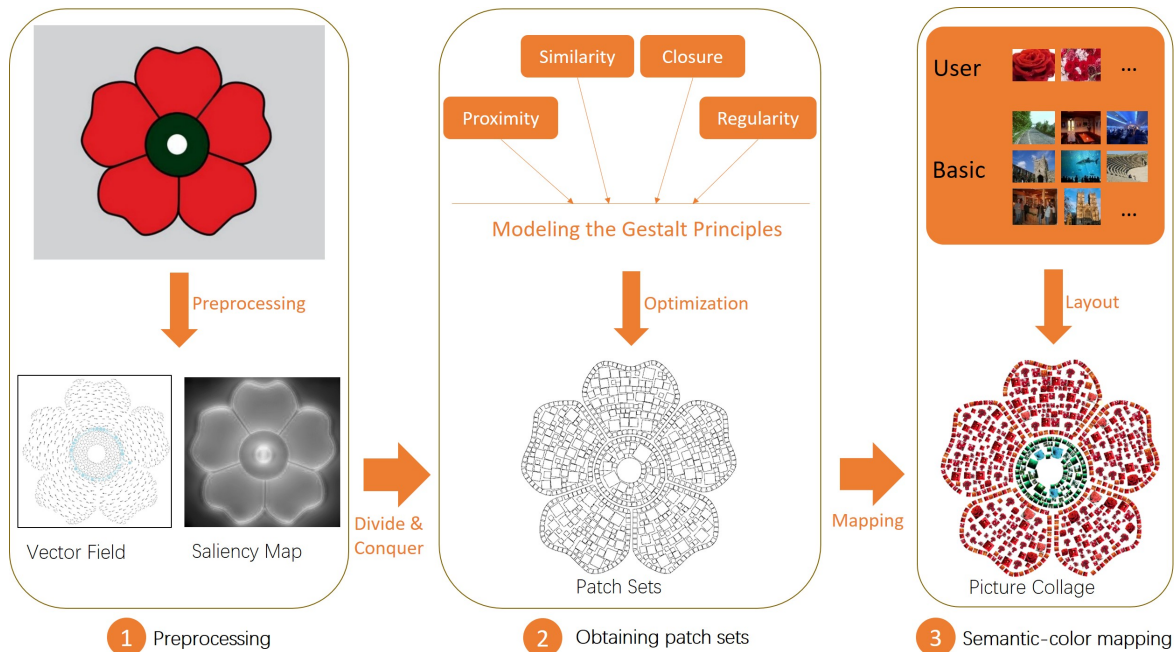
1. We propose a novel system for generating picture collages. To the best of our knowledge, our method is the first attempt to apply both Gestalt principles and image saliency to ensure that the generated collages conform to human visual perception;
2. We present a novel semantic-color similarity metric to search suitable pictures jointly considering high-level semantic and low-level color features. To compute the salient map of the canvas, we build an

eye movement based visual attention model.

## 2 Related work

In recent years, many researchers have studied the computer aided design of the picture collages in different contexts. Most of them focused on creating a visual and informative image summary from a given picture set [1, 2, 10, 12, 21, 22]. Wang et al. [21] first presented a Bayesian framework to automatically create picture collages, where the MCMC (Markov chain Monte Carlo) method was used to optimize the image arrangement. However, their work does not support the layout in arbitrary shape. Based on them, considering both semantic and visual factors, Battiato et al. [1] designed a self-adaptive image cropping algorithm and used a genetic algorithm to generate the image layout. In addition, Yu et al. [22] developed a picture collage system with a circle packing method. Furthermore, Liu et al. [12] introduced the content correlation among pictures to ensure their proximity, and extracted salient regions of pictures to make full use of the canvas space. Sharing the similar idea, we also consider these two factors to guide the generation of the element layout. However, differing from them, we compute the semantic relevance between canvas and pictures, not among pictures, and we generate the saliency map from the canvas, not from the picture elements. What's more, all the above works only adopt simple shapes, such as rectangle, circle or rhombus, as the layout canvas, and focus on creating a visual and informative summary from the given image set, to provide a quick and efficient way for browsing the image set. Different with them, our method adopt an arbitrary image as the canvas and aim to synthesize a picture collage which not only conveys the semantic and visual features of the canvas, but also conforms to human visual perception.

Photo montage pioneers another style of picture collages, where the result is synthesized from two or more images by cutting, gluing, rearranging and overlapping operations [5, 7]. Goferman et al. [5] presented a framework to produce an informative and pretty photo montage by exactly cutting the interesting regions of images in a puzzle-like manner. Huang et al. [7] created an Arcimboldo-like collage from the cutouts of multiple Internet images. They first used a mean shift clustering approach to automatically segment the input image into patches. Then, they selected a cutout for each patch with a component-aware cutout matching method. Finally, they assembled these cutouts with an affine transformation. To some extent,



**Fig. 2** Our system overview. Given the canvas image (1 top), we first generate the vector field and the saliency map (1 bottom); Then, we adopt a divide and conquer strategy to obtain a series of patch sets along the vector field, where the patch's size is controlled by the saliency of its location; Next, based on the Gestalt principles (2 top), we define an energy function to choose the best patch set (2 bottom); Finally, with a semantic-color similarity metric, we map the picture set to the patch set and generate the final picture collage (3 bottom).

the generation of our work is similar to this work, because we both aim to find a suitable picture for each patch using a similarity metric. However, their work mainly considers the color and shape similarities, while our work adopts a semantic-color metric to select suitable pictures automatically.

Besides for photo montage, picture mosaic [9, 10, 18] is also another form of visual synthesis collages. It is a synthesis where a source image is periodically divided into tiled sections (usually of equal size) and each section is replaced with one matching photograph. In the work of Jigsaw image mosaics [9], Kim et al. divided the given canvas image into several arbitrary-shaped tiles as compactly as possible and optionally deformed them slightly to achieve a more visually pleasing effect. Different from their method, we use the saliency map to search for more variable-size patches and generate the final layout in a discrete way. Most importantly, we introduce the Gestalt principles to optimize the layout to achieve a more consistent result with human visual perception.

### 3 Method

An overview of our picture collage generation is illustrated in Fig.2. Our system adopts a canvas image and a picture set as the input, and outputs

a synthesized picture collage. First, we compute a saliency map for the canvas and a smooth vector field for each sub-region of the canvas. Then, along the vector field, we partition each region into a patch set with a divide-and-conquer strategy, and adopt the saliency map to control the patch size. Additionally, by adjusting the control parameters, we could obtain a series of patch sets for each region. Next, by minimizing a Gestalt principles based energy function, we select the most visually pleasing and orderly patch set as the determined layout. Finally, with a semantic-color similarity metric, we map the pictures to the patches and achieve the final picture collage.

#### 3.1 Designing Vector Fields

As previously stated, the first step of our method is to design a desirable vector field for each sub-region of the canvas, which is used to guide the direction of the pictures. The vector field should preserve both the regional and textural features of the canvas. We calculate the vector field as the following steps: first, we decompose the canvas into several regions based on an edge detection algorithm. Then, we apply the Delaunay triangulation method to further divide each region into a triangular mesh. Finally, we interactively set heat source points in the mesh and record the heat diffusion

direction in each triangle as its direction in the vector field. After that, we sample a series of streamlines from the vector field to guide the patch direction during layout.

### 3.2 Generating Saliency Map

Before partitioning each region, we also need to compute a saliency map for the canvas, which will be used to decide where to place bigger or smaller patches. The saliency map measures the likelihood that each position in the canvas attracts the attention of a human observer.

Inspired by Judd *et al.* [8], we design an eye movement based visual attention model to obtain the saliency map. Specifically, we adopt the MIT database [3] as the source of the training data. The database contains a source image set and each source image has one corresponding eye movement data. Since the original eye movement data is represented as a series of discrete eye movement points on the map, we perform Gauss convolution on them to generate a continuous saliency map and label the obtained maps as the training ground truth. Additionally, we also extract a series of feature maps for each source image, including facial and low-level features. The training steps are as follows: first, select 100 pictures from the database and further divide them into two parts, training set (90%) and test set (10%). Second, obtain samples from the selected images. Specifically, we extract 100 positive samples and 100 negative samples from the top 30% and bottom 50% salient areas of each image respectively, and obtain 20000 samples totally. Third, compute the sample feature. In particular, for each sample, we extract the gray pixel values of its location in the feature maps and combine the extracted ones into a feature vector. Finally, based on the samples, we train a linear support vector machines (SVM) as the practical model. In the forecasting process, we extract the same feature maps from the given canvas and utilize the trained visual attention model to calculate its saliency map.

### 3.3 Obtaining Patch Sets

After obtaining the streamlines and the saliency map, we search for a patch set from each region. Region partition is a classical problem, where a set of patches with the same or various sizes are arranged into the given region. For our case, each patch is treated as a square, whose size is decided by the region area and the its location's salience.

#### 3.3.1 Problem Formulation

Given a region  $\Omega \in R^2$  and  $n$  streamlines  $SL = \{sl_i\}_{i=1}^n$ , the region partition problem in our case is to compute a best configuration for all the patches  $P = \{p_j\}_{j=1}^m$  encompassed in  $\Omega$ . What's more, it should be ensured that there is no overlap among patches and the patches are arranged based on the streamlines. In addition, we introduce coverage rate  $\tau$  to ensure that all the patches approximately cover the whole region. Specifically, the region partition problem is defined as the following optimization problem:

$$\begin{aligned} & \text{Maximize } \sum_{j=1}^m \text{Area}(p_j) \geq \tau * \text{Area}(\Omega). \\ & \text{Subject } \begin{cases} p_i \subseteq \Omega & i \in 1, \dots, m. \\ p_i \cap p_j = \emptyset & i, j \in 1, \dots, m, i \neq j. \end{cases} \end{aligned} \quad (1)$$

The arrangement of a patch set  $P = \{p_j\}_{j=1}^m$  in  $R^2$  is represented by  $C(L, D, S) = \{(l_1, \dots, l_m), (d_1, \dots, d_m), (s_1, \dots, s_m)\}$ , where  $l_j$ ,  $d_j$  and  $s_j$  are the location, the direction and the size of  $i$ th patch  $p_j$ . We designate  $C$  as a configuration. If all the patches satisfy the two constraints in Equation 1, we say that the configuration is valid. Note that one patch is looked as a closed square and thus the patch's area in the first constraint is calculated based on its size. The inclusion and intersection relations in the second constraint are computed by the relation of their pixels.

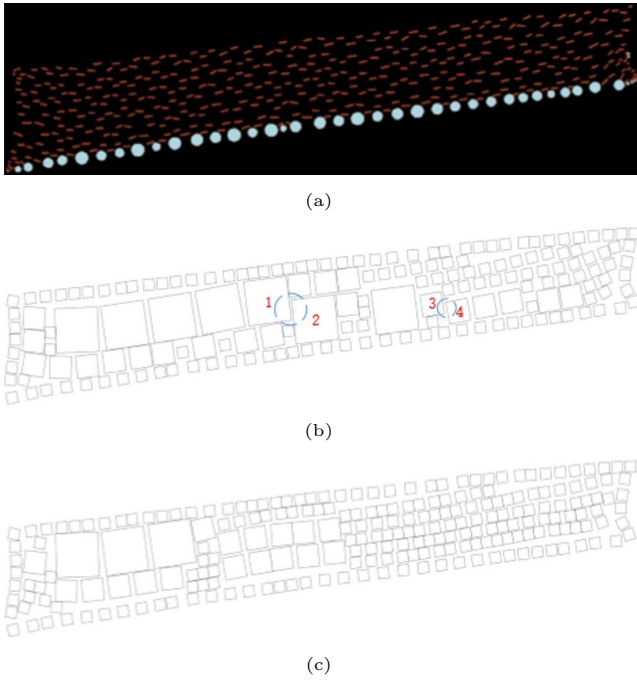
#### 3.3.2 Finding Patch

As mentioned above, we need to arrange the patches along the streamlines, so we traverse the streamlines one by one. In addition, we use a top-down search strategy to maximize the size of the patches and minimize the number of the patches, which also accelerates the search process. Namely, we first consider the patch with the biggest size to cover the region. If no patch is found, we decrease the patch size by half and search again. This operation is performed repeatedly until the constraints in Equation 1 are met.

The above algorithm can obtain a patch set that covers the region with a desired cover rate. However, it may not achieve that the patch's size follows its location's salience. Namely, according to the saliency map, a large-sized patch should be placed in a position with higher salience, and vice versa. Thus, we use a lightness threshold  $\xi(s)$  to decide whether a patch with a certain size  $s$  can be placed in a position. The lightness threshold  $\xi(s)$  is defined as:

$$\xi(s) = \text{lightness}_{MAX} - \lambda * \log_2 \frac{\text{size}_{MAX}}{s} \quad (2)$$





**Fig. 3** Comparison on the use of divide-and-conquer strategy. (a) is the vector field, where red arrows indicate the direction of the vector field and blue circles represent the heat source point. (b) is the result without divide-and-conquer strategy. (c) is the result with divide-and-conquer strategy.

where  $\lambda$  is a relaxation factor, which can be adjusted to obtain different lightness thresholds. After finding a patch, we first calculate the average lightness of all pixels in the patch. If the average lightness is greater than the lightness threshold, we think the patch is valid and add it to the patch set  $P$ . Otherwise, we abandon the patch and search for a new one along the streamline.

However, a patch obtained by the above strategy may appear to be somewhat visually disordered, because it may intersect visually with an adjacent patch with the same size, coming from **other streamlines**, such as the phenomenon indicated by the blue circle in Fig.3b. To solve this problem, we selectively traverse the streamlines. The selection criteria is based on whether more than one patch is found in the current streamline. If found, we divide other streamlines into two sets, namely, the streamlines inside (1) and outside (2) the region surrounded by the extended line of the found patches. Then, we search for patches with half the size along the first set of streamlines, and search for patches with the same size along the second set of streamlines. This divide-and-conquer strategy is detailed as follows:

1. Choose a streamline  $sl_i$  in streamlines set  $SL$  sequentially.

2. Traverse the current streamline  $sl_i$ . If finding two points  $sl_i^j$  and  $sl_i^k$  meeting  $\|sl_i^j, sl_i^k\| = s$ , ( $1 \leq j < k \leq N$ ), construct a square patch  $p$  and use the line from  $sl_i^j$  to  $sl_i^k$  as the center line of the patch. Then, go to step 3; Otherwise go to step 5.

3. If  $p$  in  $\Omega$  and  $p \cap P = \emptyset$ , calculate the average lightness  $l$  of patch  $p$  and go to step 4; Otherwise go to step 2.

4. If  $l \geq \xi(s)$ , add  $p$  to  $P$  and update  $C$  and go to step 5. Otherwise go to step 2.

5. If  $\sum_{j=1}^m Area(p_j) \geq \tau * Area(\Omega)$ , go to step 8; Otherwise go to step 6.

6. If streamline  $sl_i$  has more than one patch, divide the streamlines into two sets  $SL_1$  and  $SL_2$ .  $SL_1$  is inside the region surrounded by the extended line of the found patches, and  $SL_2$  is outside the region. Go to step 7; Otherwise go to step 1.

7. Set patch size  $\frac{s}{2}$  to  $SL_1$  and  $s$  to  $SL_2$ , go to step 1 respectively.

8. Return  $P$  and  $C$ .

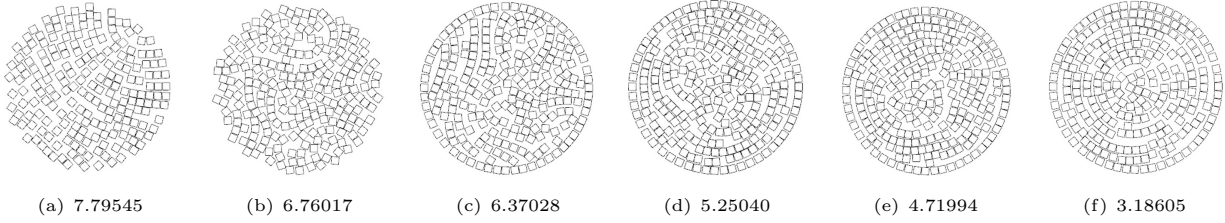
### 3.4 Modeling the Gestalt Principles

With the divide-and-conquer algorithm and the saliency map, we obtain a great patch set to approximately cover the whole region. However, due to the complexity of the vector field and the uncertainty of the distribution of different saliency maps, if using a fixed lightness threshold, the visual effect of the patch set may be in chaos. Since Gestalt psychology principles reflect the strategies of the human visual system about how to group objects into forms and create internal representations for them. Whenever groups of visual elements have one or several characteristics in common, they are combined to form a new larger visual object [16]. Thus, we use Gestalt principles to measure the degree a patch set meets human visual perception. Specifically, we choose to apply four Gestalt principles to synthesize picture collage, including proximity, similarity, closure and regularity. They are defined as follows:

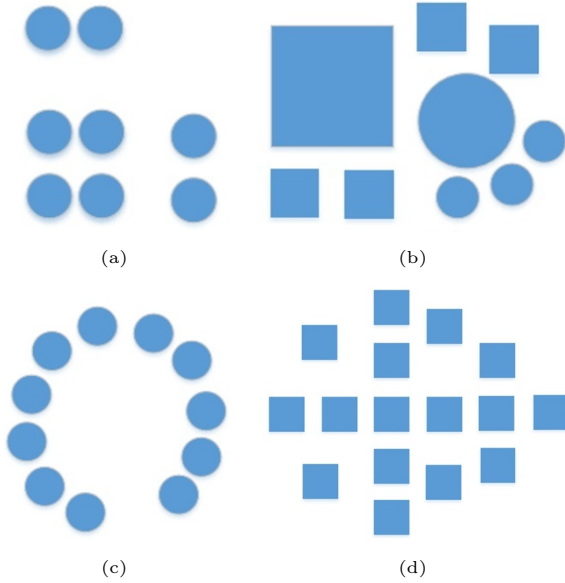
1. Proximity: the proximity principle (Fig.4a) states that when an individual perceives an assortment of objects, those objects close to each other are usually perceived as a group. We measure the proximity between the two patches  $p_i$  and  $p_j$  as follows:

$$Proximity(p_i, p_j) = \|l_i - l_j\| - (s_i + s_j). \quad (3)$$

2. Similarity: the similarity principle (Fig.4b) states that elements within an assortment of objects are usually perceptually grouped together if they are similar to each other in some qualities, such as shape,



**Fig. 5** Different patch sets and their energy values.



**Fig. 4** Gestalt Principles. (a) Proximity, (b) Similarity, (c) Closure and (d) Regularity.

size, color, orientation and so on. We measure the similarity of the  $i$ th patch  $p_i$  as follows:

$$d_s(p_i, p_j) = \begin{cases} \frac{s_i}{s_j}, & \text{if } s_i \geq s_j; \\ \frac{s_j}{s_i}, & \text{otherwise.} \end{cases} \quad (4)$$

$$d_d(p_i, p_j) = \frac{d_i \cdot d_j}{|d_i| * |d_j|}. \quad (5)$$

$$\begin{aligned} \text{Similarity}(p_i) = & \omega_{drec} * \sigma\left(\bigcup_{j \in N(p_i)} d_d(p_i, p_j)\right) \\ & + \omega_{size} * \sigma\left(\bigcup_{j \in N(p_i)} d_s(p_i, p_j)\right). \end{aligned} \quad (6)$$

3. Closure: the closure principle (Fig.4c) states that when objects are not complete, an individual usually perceives them as being a whole. We measure the closure of the  $i$ th patch  $p_i$  as follows:

$$\text{Closure}(p_i) = \frac{d_i \cdot d_j}{|d_i| * |d_j|} * \|l_i - l_j\|. \quad (7)$$

4. Regularity: the regularity principle (Fig.4d) states that objects regularly spaced tend to be grouped

together. We measure the regularity of the  $i$ th patch  $p_i$  as follows:

$$\text{Regular}(p_i) = \sigma\left(\bigcup_{j \in N(p_i)} \text{Proximity}(p_i, p_j)\right). \quad (8)$$

After formulating the above Gestalt principles, we obtain a series of patch sets by adjusting the lightness threshold and choose the most visually pleasing and orderly one by minimizing the following energy function:

$$\begin{aligned} E = \min_p \frac{1}{N} \sum_{i=1}^N (\omega_s * \text{Similarity}(p_i) \\ + \omega_c * \text{Closure}(p_i) + \omega_r * \text{Regular}(p_i)) \end{aligned} \quad (9)$$

where  $\omega_s, \omega_c, \omega_r$  are the weights of the similarity, closure and regularity principles respectively, whose sum is equal to one, and  $N$  is the number of patch set  $P$ . To verify the effectiveness of the energy function, we generated a series of visually different patch sets and calculated their energy values, as shown in Fig.5. It's easy to see that the layouts with smaller energy values are more orderly and visually pleasing.

### 3.5 Semantic-Color Mapping

In this step, we need to find a mapping between pictures and patches. Previous works usually choose the input pictures to fit the given canvas based on the color of the pictures rather than the semantics. We argue that the more important pictures should be emphasized by assigning them larger space for a more informative collage. Thus, we apply a novel semantic-color similarity metric to evaluate each picture and choose the most similar picture for each patch. Specifically, there exists a tradeoff between the high-level semantic and low-level color features. For high-level semantic feature, we would like to choose pictures highly theme-related to fit the canvas image. Whereas, for low-level color feature, we prefer to choose pictures with higher color similarity to fit the corresponding patch. Therefore, our semantic-color similarity metric can be defined as follows:

$$E_{match} = (P_i, I_j) = \omega_{sem} * D_{EJ}(W(P_i), M(I_j)) + \omega_{col} * D_{INT}(H(P_i), H(I_j)) \quad (10)$$

where  $\omega_{sem}$  and  $\omega_{col}$  are weight terms.  $D_{EJ}$  and  $D_{INT}$  measure the semantic and color similarity between  $i$ th patch  $P_i$  and  $j$ th picture  $I_j$ , whose calculations are detailed in the following paragraphs.

### 3.5.1 Semantic Similarity

If the picture element's semantics confirm to the theme of the canvas, a more informative and visual appealing collage can be synthesized. To achieve this goal, we adopt the pictures collected by Patterson et al. [17] as our basic picture set, including 14340 pictures. In addition, each picture has 102 discriminative attributes, which constitute the probability distribution vector  $W$  of the picture. For example, "Trees 1.0", "grass 0.8" and "flowers 0.8" indicate that there are 100%, 80% and 80% probability that the picture contains grass, trees and flowers respectively. Also, we manually assign the canvas image a vector of the same dimension  $M$ . Finally, we use Jaccard Distance to measure the semantic similarity between the canvas image and the picture element (see equation 11). Note that we directly use the above result to measure the semantic similarity between the patch and the picture, instead of calculating a separate one for each patch.

$$D_{EJ}(W, M) = \frac{W \cdot M}{\|W\|^2 + \|M\|^2 - W \cdot M}. \quad (11)$$

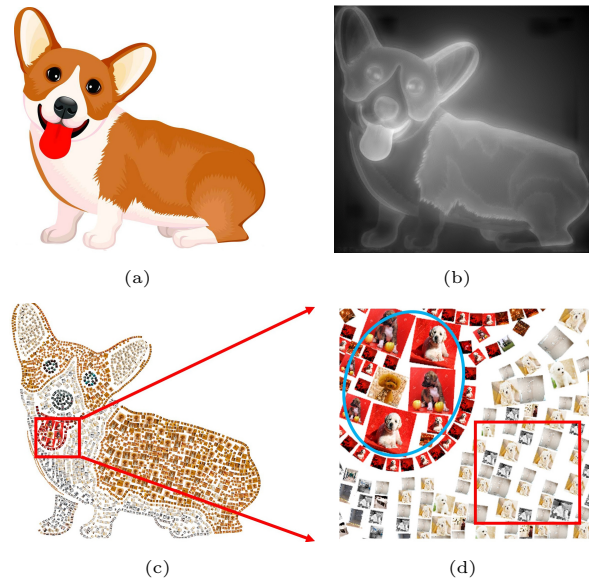
### 3.5.2 Color Similarity

We measure the color similarity using the color statistics computed from HSV color histogram. According to our experiment experience, we set the hue, saturation and value channels consist of 8, 16 and 4 bins, respectively. Let  $H_1$  and  $H_2$  denote the HSV histograms of one patch and one picture, we compute their histogram intersection distance, defined as:

$$D_{INT}(H_1, H_2) = \sum_i \min(H_1(i), H_2(i)) \quad (12)$$

where  $H(i)$  denotes the frequency of pixels that fall into  $i$ th bin and  $D_{INT}$  represents the color similarity.

As mentioned above, we hope that the large-sized patch is assigned with a more theme-related picture, because the patch's size indicates the saliency of its location. Whereas, for small-sized patches, we prefer to fit pictures with higher color similarity, which can better preserve the visual feature of the canvas. Thus, we use the patch's size to determine the weights between semantics and color:



**Fig. 6** Picture collage whose theme is 'dog'. (a) is the source image; (b) is the saliency map; (c) is the generated picture collage; (d) is the detail view of the region marked in red rectangle.

$$\omega_{sem}(s) = \lambda * \frac{s}{size_{MAX}}, \quad (13)$$

$$\omega_{col}(s) = 1 - \omega_{sem}(s) \quad (14)$$

where  $s$  is the current patch size,  $size_{MAX}$  is the default or user-set maximum size, and  $\lambda$  is the tradeoff coefficient.

## 4 Result and Evaluation

In this section, we first evaluate the performance of our method according to following three criterion:

1. Image salience: whether larger pictures are placed at the more salient positions, and vice versa;
2. Semantic-color similarity metric: whether the picture elements and the canvas are theme-related and whether the picture element's color confirms to the color of its location;
3. Gestalt principles: whether the use of Gestalt principles contributes a more visually pleasing and orderly layout.

In addition, we also compare our method with Shape Collage [4] and Arcimboldo-like collage [7], to verify the overall visual effect of our generated collages.

### 4.1 Criteria Evaluation

First, we generated a group of collages using "dog" image as the canvas to validate whether the use of saliency map is useful (see Fig.6). The salient areas

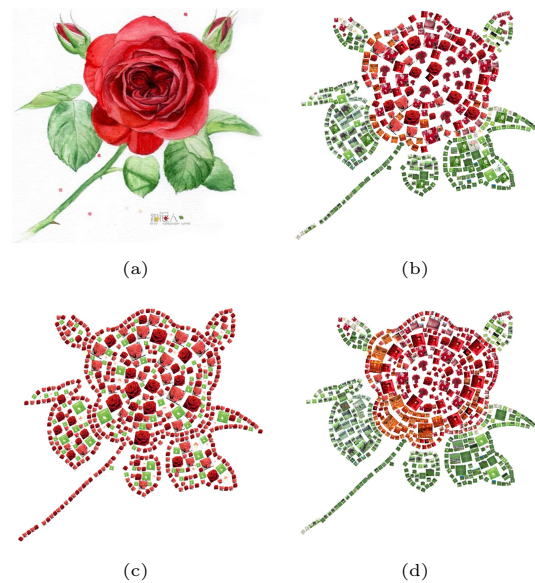


are mainly located at the back, face and mouth of the dog. Fig.6d is the detail view of the region marked in red rectangle in Fig.6c, in which there are both saliency and non-saliency areas. In Fig.6d, we can see that we successfully place some big patches on the saliency area marked with blue circle, and some small patches on the non-saliency area marked with red rectangle. Through this strategy, we can preserve the theme and visual characteristics of the canvas to the maximum.

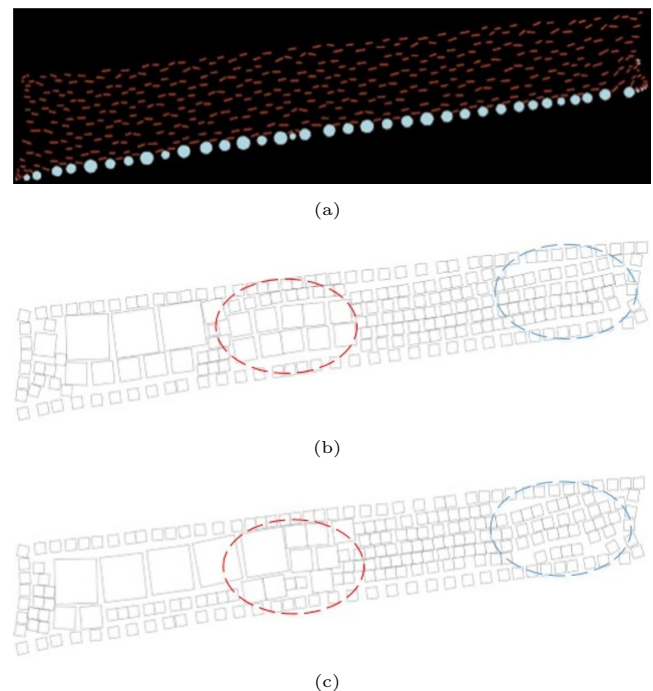
Second, we also generated three results with different considerations of semantics and color to validate the semantic-color similarity metric (see Fig.7). Specifically, we used "flowers" as the semantic attribute to search for satisfied pictures from the picture database. Fig.7a shows the canvas image, a beautiful flower; Fig.7b is the result considering both semantics and color, where  $\omega_{sem} = \omega_{col} = 0.5$ ; Fig.7c is the result only considering semantics, where  $\omega_{sem} = 1$ ; Fig.7d is the result only considering color, where  $\omega_{col} = 1$ . We can see that the constituent elements of Fig.7c are all flower topic related pictures. However, some pictures' color mismatch the color of the canvas. As for Fig.7d, though it has a strong color similarity between the elements and the canvas, it is not semantically relevant enough. On the contrary, Fig.7b combines the former two and shows a strong semantic-color similarity. This comparison shows the effectiveness of the semantic-color similarity metric. With this metric, the generated picture collage can achieve the tradeoff between retaining the semantic and visual characteristics of the canvas.

Third, we generated two patch sets with and without Gestalt principles to validate the effectiveness of the Gestalt principles. Fig.8a shows the direction of streamlines; Fig.8b and Fig.8c show the results generated with and without Gestalt principles, respectively. In Fig.8c, we can find two problems easily: (1) the patches' size in some areas varies significantly (see the area marked with red circle); (2) the patches' layout in some areas is in chaos (see the area marked with blue circle). With the Gestalt principle, we can solve these problems efficiently. The patch set in Fig.8b looks more orderly and visually pleasing and validates the effectiveness of the Gestalt principles.

Through the above experiments, our system is verified effective in considering image saliency, semantic-color metric and Gestalt principles. The saliency map and the semantic-color metric give users a better understanding to the theme and visual characteristics of the canvas image. Meanwhile, the use of the Gestalt principles contributes a more visually

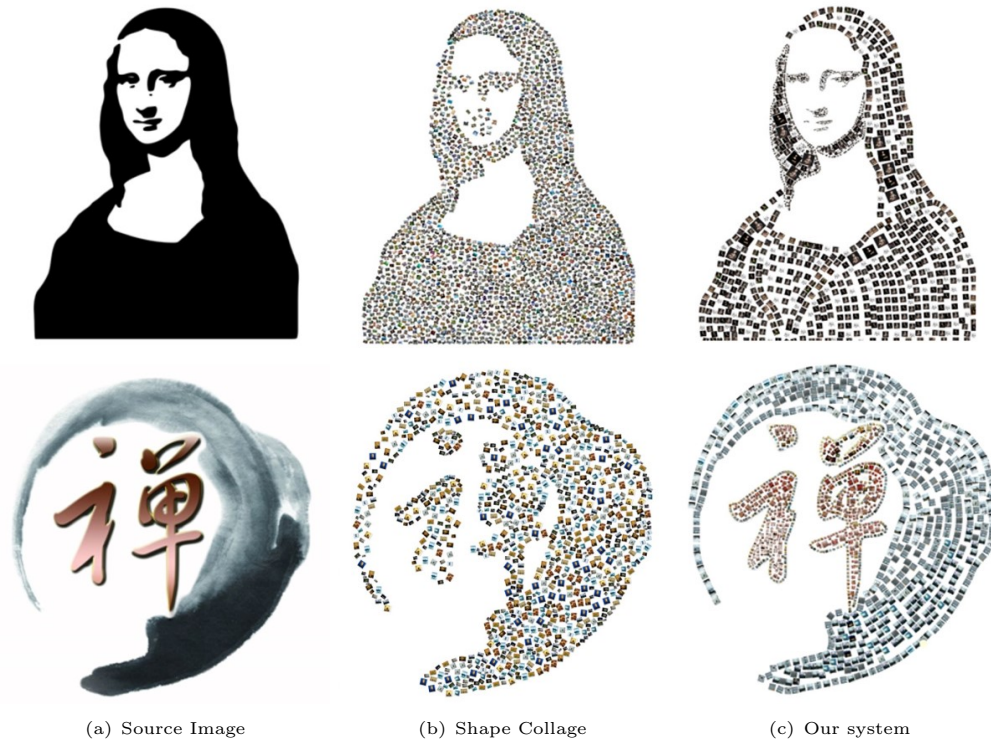


**Fig. 7** The comparison on the use of semantic-color similarity. (a) is the source image; (b) is the result generated jointly considering both the semantics and color; (c) is the result only considering semantics; (d) is the result only considering color.



**Fig. 8** The comparison on the use of Gestalt principles. (a) is the vector field, where red arrows indicate the direction of the vector field and blue circles indicate the heat source points; (b) is the result generated with Gestalt principles; (c) is the result generated without Gestalt principles.





**Fig. 9** The comparison with Shape Collage. From left to right, (a) the source image, (b) the results of Shape Collage and (c) the results of our method are shown in turn.

pleasing layout and further achieves a picture collage more consistent with human visual perception.

## 4.2 Comparisons

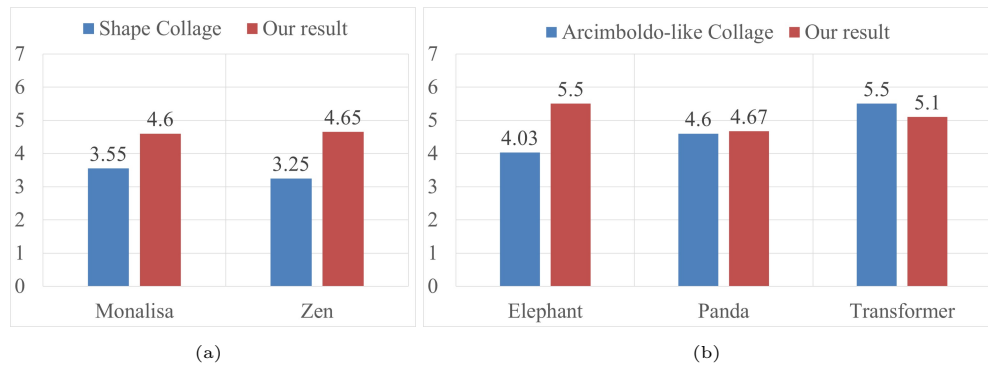
To verify the overall visual effect of the results, we compared different collages generated by three methods, Shape Collage [4], Arcimboldo-like collage [7] and our method (see Figs.9, 10 and 11). Besides, we also conducted a user study in the form of online questionnaire to quantitatively evaluate these collages. Considering the differences among these three styles, we only conducted the comparisons in the overall visual effect. In the user study, we asked twenty participants to score the results of different works, eleven female and nine male students with an average age of 22. Each participant graded the pictures on a scale from zero (poor) to seven (excellent), which evaluates their preferences on the overall visual effect of different results. We analyzed the feedback data and calculated the average scores, which are shown in Fig.10.

First, we compare our system with Shape Collage [4], a worldwide and popular collage software. It can create collages with great visual effects in a few seconds and support arbitrary shapes. More importantly, its results are more similar with ours, because we both arrange

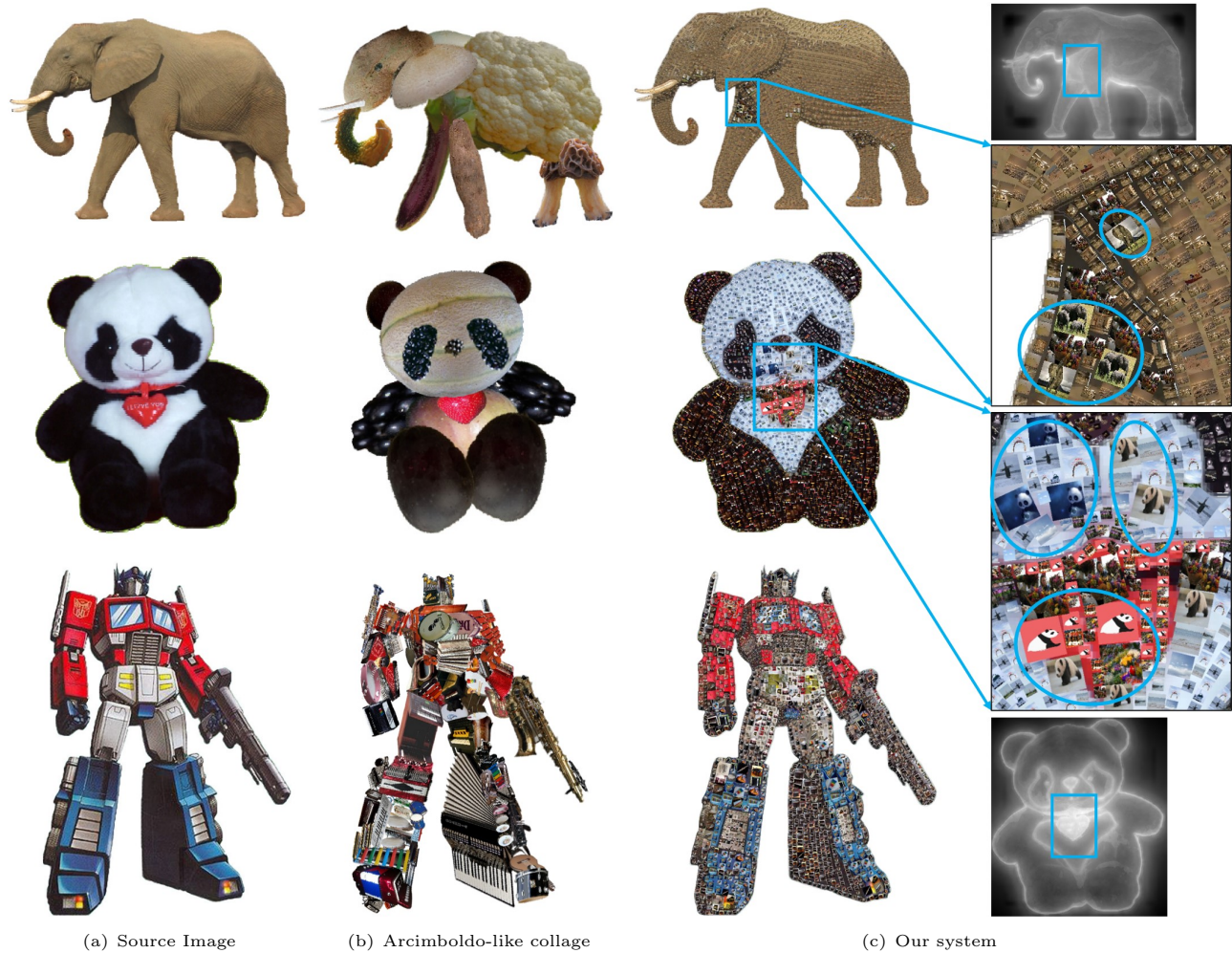
the elements in a discrete way.

As shown in Fig.9, Shape Collage performs well on the preservation of the shape of the canvas. However, it does not consider the semantic-color similarity as we do, which causes a disorganized collage layout. On the contrary, our method preserves both the visual and shape features of the canvas well. Besides, it retains more image details than Shape Collage, such as the face of Monalisa. Because it can adaptively adjust the picture element's size based on the region's area. We also show the result of the user study in Fig.10a, where the rating is significantly maintained at two levels and our method obtains generally higher scores than Shape Collage. Also, we applied ANOVA to the collected data and got ( $P_{monalisa} = P_{zen} = 0.00$ ), which shows that there are significant differences between the two groups of scores. From this comparison, we prove that our method achieves a better overall visual effect than Shape Collage.

Second, we also compare our method with Arcimboldo-like collage[7]. By filtering massive internet images, Arcimboldo-like collage combined multiple thematically-related image cutouts to represent the input canvas image. The selected cutouts were purposefully arranged such that as a whole



**Fig. 10** The average scores of the results of the three methods in the user study.



**Fig. 11** The comparison with Arcimboldo-like collage. From left to right, (a) the source image, (b) the results of Arcimboldo-like collage and (c) the results of our system are shown in turn. Besides, a detailed view of the salient areas of the elephant and panda images is also shown on the far right.

assembly to represent the input image with disguise in both shape and color, and the individual cutout still could be recognizable as its own being. Arcimboldo-like collage arranges elements in a continuous way, while our method adopts a discrete way, the streamline-based arrangement, to lay out the pictures. Therefore, for comparison, we use the original pixels of the canvas to fill the non-layout areas, namely, the gaps between the patches (Fig.11c).

Fig.11 shows the comparison between Arcimboldo-like collage [7] and our method. In Fig.11b, Multiple meaningful and thematically-related cutouts constitute the visual representation of the source images (Fig.11a), such as "vegetable" elephant, "fruit" panda and "music" transformer. Differently, our method (Fig.11c) considers more about the semantic and color relevance between the canvas and the elements. Also, we use much smaller elements to visually represent the source images. What's more, our method applies two more visual perception factors, saliency map and Gestalt principles, to guide the layout of the elements. The former is used to retain the theme features of the canvas, and the latter contributes a more visually pleasing and orderly layout. From the detailed views of the Fig.11c, we can see that some large-sized and theme-related pictures are placed in the salient regions (marked in blue rectangle). Note that though some other regions are also salient enough, such as the nose tip and edge regions of the elephant, they are too small to place large-sized pictures, or the shape of the canvas will be broken. Fig.10b shows the scores of these two works in the user study. We can see that the scores among the three pictures vary randomly between the two methods. The p values obtained from ANOVA are ( $P_{elephant} = 0.00, P_{panda} = 0.86, P_{transformer} = 0.23$ ) respectively. Therefore, we can conclude that these two methods achieve similar overall evaluation.

**Run Times:** Our system is implemented on a PC with 3.2 GHz CPU, 8 GB system memory. The time required for computation depends on the vector field generation, layout optimization and user interaction. If the time consumed in user interaction is ignored, it takes about 30 minutes to generate a result. More results are shown in Fig.12.

**Limitation:** However, our system still has a few limitations. First, it only provides a limited number of types of the vector fields. If the canvas image contains many irregular regions, the vector fields generated from these regions may be unexpected, even against user intention. Second, it is time consuming to generate a picture collage now, because the process of finding

patches and searching for suitable pictures takes a lot of time. To address this issue, one potential solution is to use GPU to speed up the algorithm. Third, our system may produce results with unsatisfactory visual effects, when there are no proper pictures to satisfy both the semantic and visual features of the canvas. Toward this end, we could extend our image database to cover more image themes to generate visual pleasing picture collages in the future.

## 5 Conclusions

In this paper, we have proposed a novel visual perception driven method for creating picture collages. Given a canvas image and a picture set, we first compute a saliency map for the canvas. Besides, we segment the canvas into several regions and calculate a vector field for each region. Second, along the vector fields, we search for several patch sets using a divide-and-conquer strategy and adopt the saliency map to determine the patch's size. Third, we construct a Gestalt principles based energy function to achieve the most visually pleasing and orderly layout. Finally, with a semantic-color metric, we map the pictures to the patches to get the final picture collage. The picture collages generated by our method not only can achieve the visual simulation of the canvas, but also enhance the semantic theme of the canvas. More importantly, this method is the first one to introduce Gestalt principles into the creation of picture collages, which makes the generated results more consistent with human visual perception. We believe that this method is a great demonstration of the combination of cognitive psychology and art computing.

## Acknowledgements

This work was supported by the National Nature Science Foundation of China (No.61772440), the Aeronautical Science Foundation of China (No.20165168007), Science and Technology on Electro-optic Control Laboratory.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.



## References

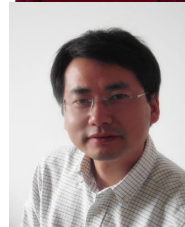
- [1] S. Battiato, G. Ciocca, F. Gasparini, G. Puglisi, and R. Schettini. Smart photo sticking. In *Adaptive Multimedial Retrieval: Retrieval, User, and Semantics*, pages 211–223, 2007.
- [2] S. Bianco and G. Ciocca. User preferences modeling and learning for pleasing photo collage generation. *Acm Transactions on Multimedia Computing Communications & Applications*, 12(1):1–23, 2015.
- [3] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark, 2015.
- [4] V. Cheung. Shape collage, 2007.
- [5] S. Goferman, A. Tal, and L. Zelnik-Manor. Puzzle-like collage. In *Computer Graphics Forum*, pages 459–468, 2010.
- [6] X. Han, C. Zhang, W. Lin, M. Xu, B. Sheng, and T. Mei. Tree-based visualization and optimization for image collection. *Library Journal*, 46(6):1286, 2014.
- [7] H. Huang, L. Zhang, and H. C. Zhang. Arcimboldo-like collage using internet images. In *SIGGRAPH Asia Conference*, page 155, 2011.
- [8] T. Judd, K. Ehinger, D. F., and T. A. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, 2009.
- [9] J. Kim and F. Pellacini. Jigsaw image mosaics. *Acm Transactions on Graphics*, 21(3):657–664, 2002.
- [10] H. Y. Lee. Automatic photomosaic algorithm through adaptive tiling and block matching. *Multimedia Tools & Applications*, 76(22):24281–24297, 2017.
- [11] C. Liu, E. Rosales, and A. Sheffer. Strokeagggregator: Consolidating raw sketches into artist-intended curve drawings. *ACM Transaction on Graphics*, 37(4), 2018.
- [12] L. Liu, H. Zhang, G. Jing, Y. Guo, Z. Chen, and W. Wang. Correlation-preserving photo collage. *IEEE Transactions on Visualization and Computer Graphics*, 24(6):1956–1968, 2018.
- [13] Z. Lun, C. Zou, H. Huang, E. Kalogerakis, P. Tan, M.-P. Cani, and H. Zhang. Learning to group discrete graphical patterns. *ACM Trans. Graph.*, 36(6), Nov. 2017.
- [14] H. L. Man, N. Singhal, S. Cho, and I. K. Park. Mobile photo collage. In *Computer Vision and Pattern Recognition Workshops*, pages 24–30, 2010.
- [15] T. Mei, B. Yang, S. Q. Yang, and X. S. Hua. Video collage: presenting a video sequence using a single image. *Visual Computer*, 25(1):39–51, 2009.
- [16] L. Nan, A. Sharf, K. Xie, T. T. Wong, O. Deussen, D. Cohen-Or, and B. Chen. Conjoining gestalt rules for abstraction of architectural drawings. *Journal of Computer-Aided Design & Computer Graphics*, 30(6):1–10, 2012.
- [17] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.
- [18] D. Pavic, U. Ceumern, and L. Kobbelt. Gizmos: Genuine image mosaics with adaptive tiling. *Computer Graphics Forum*, 28(8):2244–2254, 2009.
- [19] Y. Qi, J. Guo, Y. Li, H. Zhang, T. Xiang, and Y. Song. Sketching by perceptual grouping. In *2013 IEEE International Conference on Image Processing*, pages 270–274, 2013.
- [20] Y. Qi, Y. Song, T. Xiang, H. Zhang, T. Hospedales, Y. Li, and J. Guo. Making better use of edges via perceptual grouping. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1856–1865, 2015.
- [21] J. Wang, L. Quan, J. Sun, X. Tang, and H. Y. Shum. Picture collage. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 347–354, 2006.
- [22] Z. Yu, L. Lu, Y. Guo, R. Fan, M. Liu, and W. Wang. Content-aware photo collage using circle packing. *IEEE Transactions on Visualization & Computer Graphics*, 20(2):182–195, 2013.
- [23] J. Zhang, Y. Chen, L. Li, H. Fu, and C.-L. Tai. Context-based sketch classification. In *Proceedings of the Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering, Expressive '18*, New York, NY, USA, 2018. Association for Computing Machinery.



**Zuyi Yang** Zuyi Yang received his bachelor's degree in Artificial Intelligence Department, Xiamen University, China, in 2018. Currently he is a Master candidate in Xiamen University, China. His research interests include Computer Graphics and Layout Optimization.



**Qinghui Dai** Qinghui Dai received his master's degree in Artificial Intelligence Department, Xiamen University, China, in 2018. Currently he is a technology staff member of Tencent. His research interest is Computer Graphics.



**Junsong Zhang** Junsong Zhang received the Ph.D. degree in computer science from State Key Lab of CAD&CG, Zhejiang University, Hangzhou, China, in 2008. He is currently an Associate Professor at the Mind, Art and Computation Group, Artificial Intelligence Department, Xiamen University, Xiamen, China. His main research interests include Computer Graphics, Human Computer Interaction, and Cognitive Science.





Fig. 12 More Results.