# Correlation-Preserving Photo Collage

Lingjie Liu, Hongjie Zhang, Guangmei Jing, Yanwen Guo, Zhonggui Chen, and Wenping Wang, *Fellow, IEEE*

**Abstract**—A new method is presented for producing photo collages that preserve content correlation of photos. We use deep learning techniques to find correlation among given photos to facilitate their embedding on the canvas, and develop an efficient combinatorial optimization technique to make correlated photos stay close to each other. To make efficient use of canvas space, our method first extracts salient regions of photos and packs only these salient regions. We allow the salient regions to have arbitrary shapes, therefore yielding informative, yet more compact collages than by other similar collage methods based on salient regions. We present extensive experimental results, user study results, and comparisons against the state-of-the-art methods to show the superiority of our method.

**Index Terms**—Photo collage, image saliency, irregular shaped packing, image classification

## 1 Introduction

Smartphones equipped with high-resolution cameras enable people to take high quality photos more easily than before. An album often contains photos about various memorable events, such as a holiday trip or a Christmas party, or about a common theme or related to a common object, such as life in a city or a cute pet. This has brought about challenges in displaying a group of photos in a semantically structured way. To tackle this problem, photo collage has been proposed, as an effective means to summarize and display a photo group. Its goal is to create a compact, informative and aesthetic summary representation for a group of photos.

Manually creating a good collage requires high-level skills and can be a painstaking task. Thus, a few attempts have been made to provide automatic solutions. Some commercial software tools, including AutoCollage (a plug-in of Windows Live), Google Picasa, and Photovisi, etc., have been developed.

Most existing approaches to photo collage boil down to solving complicated optimization problems [1], [2], [3], [4], [5], [6], [7], with an objective function measuring the quality of a collage by various criteria. Typically, the presentation of each photo in a collage is determined by several geometric parameters and a layer index, overall hundreds of parameters need to be computed via optimization. The nonlinear nature of the objective function and the large parameter space to be determined together make photo collage generation a difficult problem for a photo collection with a large number of photos as shown in Fig. 1.

We present a novel method for generating a photo collage that is compact, visually pleasing, and informative. In contrast to previous methods that formulate photo collage as a complicated optimization problem, our method generates a collage in two relatively simple steps. In the first step, we partition the canvas into a set of regions according to the extracted cutouts of input photos, Then in the second step, we display the cutout of each photo in the region it is assigned to. By doing so, we avoid complex optimization on photo presentation. Our region partition algorithm allows cutouts to have arbitrary shapes. This yields more compact collages than produced by other similar region-partition based collage methods such as [8], which uses circles to approximate cutouts and so is not suitable for objects that cannot be tightly bounded by circles, such as pedestrians, cars, and tall buildings.

Another novel feature of our method, which is a main contribution, is to place photos of similar contents clustered together in the collage, which we call *content correlation*. Photos in a collection often present a certain degree of content correlation. Imagine one has a one-day San Francisco city tour. The photos taken are often related to each other, because many are probably about the same common landmark, such as the Civic Center, Golden Gate Bridge, or Fisherman's Wharf. Then, naturally, we may want to make correlated photos grouped close to each other in the collage. We propose that placing photos with similar contents together helps the viewer grasp the theme of the collage better. As a means of visual presentation of photos, such as a way of photo collage also helps the viewer quickly identify and compare similar photos (often of the same objects) for efficiently appreciating or studying a photo collection. This has been confirmed by our extensive comparison and user study, as will be presented later.


- *L. Liu and W. Wang are with the Department of Computer Science, The University of Hong Kong, Hong Kong, P.R. China. E-mail: liulingjie0206@gmail.com, wenping@cs.hku.hk.*
- *H. Zhang and Y. Guo are with the National Key Lab for Novel Software Technology, Nanjing University, Nanjing, Jiangsu 210023, P.R. China. E-mail: hjzhang@smail.nju.edu.cn, ywguo@nju.edu.cn.*
- *G. Jing was with the Department of Computer Science, The University of Hong Kong, Hong Kong, P.R. China. E-mail: gmjing2010@gmail.com.*
- *Z. Chen is with the Department of Computer Science, Xiamen University, Xiamen 361005, P.R. China. E-mail: chenzhonggui@xmu.edu.cn.*


*Manuscript received 16 June 2016; revised 20 Feb. 2017; accepted 26 Mar. 2017. Date of publication 12 May 2017; date of current version 27 Apr. 2018. (Corresponding author: Yanwen Guo.)*
*Recommended for acceptance by S.-M. Hu.*
*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.*
*Digital Object Identifier no. 10.1109/TVCG.2017.2703853*

Fig. 1. A photo collage produced by our method, consisting of 60 photos taken at 12 tourist attractions of Hong Kong.

To achieve clustering based on content correlation, we use deep learning techniques to analyze correlation of photo contents and, further, to cluster the photos based on the correlation of photo contents. This is realized by iterating the following two steps: (1) partitioning a canvas region into many cells for placing the photo cutouts; and (2) running a combinatorial optimization technique to reshuffle the positions of misplaced cutouts, if any, on the canvas. As a feature not possessed by other existing photo collage methods and softwares, the property of preserving content correlation by our approach enables the viewer to easily appreciate and grasp collage contents. Fig. 1 summarizes 60 photos taken during a tour in Hong Kong, and Fig. 2 shows the same collage with the cluster boundaries highlighted. Note that photos of the same attractions are grouped together.

Our method is capable of producing a compact and correlation-preserving collage comprising arbitrarily shaped cutouts of salient regions of input photos. Two main contributions are made in this paper.

- We present an effective method that groups photos according to the similarity of photo contents. This is helpful to efficient presentation of the collage contents.
- Our method is capable of packing salient regions of arbitrary shapes. This helps to achieve compact photo collages without loss of relevant information.

Fig. 2. The same collage as in Fig. 1, with the cluster boundaries highlighted in red.

## 2 Related Works

*Photo Collection Summarization.* The problem of photo collection summarization, which selects a subset of photos from a large collection that best represents the entire set, is considered in [9], [10]. A prototype system for automatically organizing images using concept hierarchies is developed in [11]. In [12], a framework for automatically selecting a summary set of photographs from a large collection of georeferenced photos is proposed. In [13], a scene summarization approach to processing online image collections is proposed via feature clustering, while the problem of jointly summarizing large sets of Flickr images and YouTube videos for storyline reconstruction is investigated in [14], and a similar problem is explored in [15], [16]. These methods mainly focus on selecting the images that best represent the whole collection. In contrast, our goal is to produce a visually pleasing collage that seamlessly assembles the important regions of a group of selected photos. Therefore, the above algorithms may be used for photo selection in a preprocessing step to our photo collage method.

*Photo Collage.* AutoCollage [7] defines the problem of creating the collage of representative cutouts from a set of images as an energy minimization problem which aims to obtain a label for each pixel on the canvas. In [1], the 2D spatial arrangement of rectangular images is optimized in a Bayesian framework in order to maximize the visibility of salient regions. Several other methods formulate the problem as different objective functions, and propose to use more advanced optimization techniques [2], [3], [4]. The problem of creating Digital Tapestry [6] from a photo collection is formulated as a multi-class labeling problem, which is modeled using a Markov Random Field and optimized

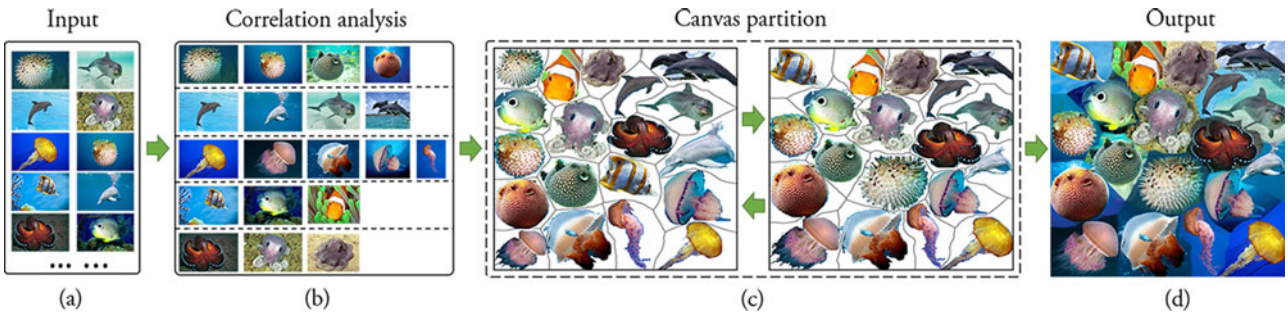



Fig. 3. The pipeline of our method. (a) The input is a collection of photos. (b) Photo correlation is analyzed and correlated photos are grouped together. (c) Photo cutouts are packed on the canvas via an iterative correlation-preserving packing process. (d) The final collage.

using an expansion move algorithm based on graph cut. The narrative collage [17] hierarchically arranges photos according to the basic narrative elements like faces, time and places. In each level of the hierarchical collage, the representative photos are arranged sequentially. The puzzle-like collage [18] aims to assemble photo cutouts of arbitrary shapes in the manner of puzzle assembly.

The photo collage method based on circle packing [8] is most related to work in that both methods first partition the canvas into disjoint cells and display each image in a designated cell. However, there are two important differences. First, the circle packing based method represents the photo cutouts as circles, thus resorting to a circle packing algorithm. Because of its use of circular cutout shapes, the method performs poorly for elongated objects which cannot be compactly bounded by circles, such as standing human figures, cars, and tall buildings. In contrast, our method allows the salient regions of input photos to be contained by arbitrarily shaped cutouts, leading to better utilization of limited canvas space. Second, our method first classifies the input photos according to content correlation and then places correlated photos close to each other, while the collage method based on circle packing [8] does not take content correlation into consideration, resulting in a random placement.

Mosaic and photomontage are problems that bear similarity to photo collage. Mosaic arranges a collection of small images in a way that suggests a large image when seen from a distance. In the Jigsaw image mosaics [19], image tiles of arbitrary shapes are used to compose an arbitrarily-shaped picture. Slight deformation is allowed for tighter packing. Digital photomontage [5] generates a single composite image by interactively piecing together multiple photos of the same scene. In [20], photo-realistic composite pictures are generated by seamlessly stitching together photos from the Internet to conform with the given freehand sketches and text labels.

*Convolutional Neural Networks.* The convolutional neural networks (CNNs) are a series of specially configured neural networks, inspired by biological findings of neural science in [21]. The CNNs are characterized by their stacked convolution and pooling layers for local connection and weight sharing. In recent years, a large number of new architectures of CNNs have shown remarkable performance in object detection and recognition, such as GoogLeNet [22], VGG [23], the Deep Residual Network [24], and so on. In this paper, we employ the VGG-16 model for its capability in extracting semantic features. Our tests show that the VGG-16 model facilitates keeping content correlation of photos in creating our collage representation.

## 3 Overview

Given as input a collection of photos, our method proceeds in three main steps, as illustrated in Fig. 3. Note that when there are some photos that are visually highly similar to each other, only one representative is selected to be included in the collage. We assume that this has been done in a pre-processing step.

*Correlation Analysis.* Our method then utilizes deep learning techniques to analyze photo correlation. The CNN model [23] is generalized to extract semantic description and to obtain a class label for every photo. Photo correlations are embedded into the initial arrangement on the canvas using t-Distributed Stochastic Neighbor Embedding (t-SNE) [25]. Photos with higher content similarity are grouped together, and photos with the same class label should stay close which imposes a constraint to canvas partition.

*Canvas Partition.* We use cutouts of the input photos, rather than the original photos, for collage. The cutout of a photo is based on the salient region of the photo, which is extracted using saliency detection [26], combined with face detection by the Viola-Jones face detector [27]. To avoid cluttering, some margin is added around the detected salient region to form the cutout. Since the salient region is in general of arbitrary shape, the cutout has an arbitrary shape. The photo cutouts, along with their class labels and the initial embedding based on their semantic description, are fed into our canvas partition algorithm that iteratively improving canvas partition to achieve optimal packing and clustering.

*Collage Assembly.* In the last step, we map each cutout onto its corresponding region yielded by canvas partition. A saliency-based blending operation [8] is used to create a seamless transition between adjacent images.

## 4 Correlation Analysis

Content correlation is common among photos. For example, a collection of photos taken during a city tour exhibits such correlation when there are different photos of the same landmarks, and photos taken in a zoo are relevant to others in that they are about the same animals. We propose to use correlation of photo contents in determining their collage layout, that is, to keep correlated photos close in the final collage.

### 4.1 CNN-Based Semantic Analysis

We stress that it is not our aim in the present paper to solve the problem of semantic understanding. Rather, we use deep learning techniques to develop an effective solution to

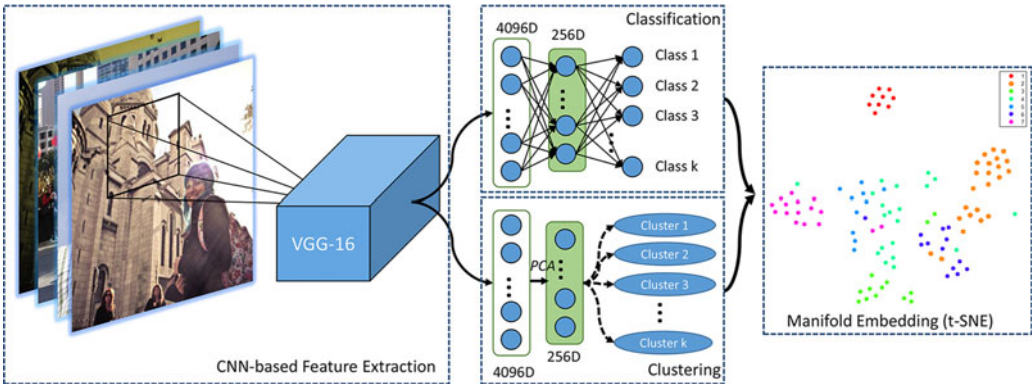



Fig. 4. The neural network used for correlation analysis. For a specific theme, the photos pass through VGG-16 and an additional 2-layer network. For general themes, the photos pass through VGG-16 and k-means classification. The 256D feature vectors are finally embedded onto the 2D canvas by t-SNE to provide an initial photo arrangement.

the generation of photo collages based on content correlation. For this purpose, we use a recently proposed CNN model, VGG-16, and extend it to suit our application, as illustrated in Fig. 4. Specifically, we first extract a 4,096D feature vector for each photo using a pre-trained VGG-16 model [23]. The 4,096D feature vector is further reduced to 256D by Principal Component Analysis (PCA) without sacrificing much accuracy. Thus, the content of a photo is represented by a 256D feature vector and the class label of each photo is assigned by applying $k$-means to the 256D feature vectors of photos. To simplify exposition, we shall use the term "class label" to denote "cluster label" in $k$-means.

The above CNN model meets the basic requirement of semantic extraction. It is, however, not suitable for photos of some special themes, such as traveling and family photos, because the VGG-16 model has not been trained for the recognition of these kinds of objects. So we demonstrate the efficacy of our framework for photo collage with content correlation for these themes of traveling and family. To achieve this, an additional network for photos of such a theme. This additional network is designed as a 2-layer fully connected neural network. It accepts a 4,096D feature vector produced by the VGG-16 as input and outputs the corresponding category in a specific domain. The categories that can be classified by the network are related to a specified theme, such as traveling. The size of the hidden layer is set to be 256 with the $\tanh$ function as activation function. This is optimized using back-propagation on the training set, and 20 percent of the training set is used for validation to avoid over-fitting. For each photo, we extract a 256D feature vector in the hidden layer as its representation and use the most likely category as its class label.

We use traveling photos to validate the efficacy of this additional network. We have trained networks for recognizing famous attractions of five cities as shown in our experiments. For every city, we chose 6 to 12 sightseeing spots recommended by "Google Image—Points of Interest" for the city. Note that training an additional network for recognizing only sightseeing spots of the five cities requires much less data than training a deep network. For each sightseeing spot, we used 500 photos crawled from the Internet as the training data, after manually culling wrong matches. The photos used for testing were provided by volunteers, which are quite different from the photos for training.

The benefit of the additional network can be seen in the test results shown in Fig. 7. Here, for a collection of photos taken during a trip to Beijing, using the VGG-16 model without the additional network incurs quite a few wrong assignments of image labels. In comparison, the VGG-16 with the additional network shows better performance in classifying these photos.

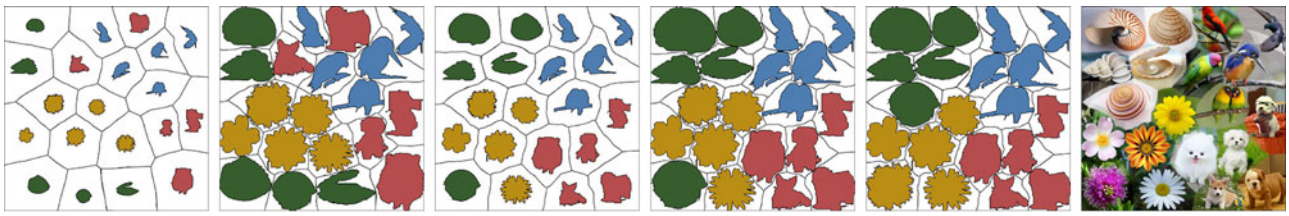
### 4.2 Photo Embedding

The extracted 256D feature vectors of the input photos are used to determine the initial positions for the photos on canvas so that photos with similar feature vectors (i.e., highly correlated in their contents) are placed close to each other. This amounts to computing an embedding on the 2D plane of the feature vectors in a high-dimensional space. This is done by utilizing t-SNE, a state-of-the-art dimension reduction method. The t-SNE technique is capable of capturing much of the local structure of the high-dimensional data as well as revealing global structures, such as the presence of clusters. It works by first constructing a probability distribution to model the similarity of high-dimensional data, and then mapping the distribution into a low-dimensional space by minimizing their distance. For more details about the mechanism of the t-SNE, please refer to [25].

Note that the t-SNE is a local optimization method, and it produces different outputs in different runs, so it suffers from uncertainty and lack of global optimality. To improve the results of t-SNE, we use the class labels of images to rectify the photo positions on the canvas, as will be detailed in Section 5.2. As shown in Fig. 8, although different outputs of t-SNE as the initialization of canvas partition will lead to different collages, these collages all meet the requirements of being compact, informative and correlation-preserving.

## 5 Irregular Shape Based Canvas Partition

To make the best use of a given canvas space, it is essential to arrange all cutouts compactly without overlapping, while keeping cutouts of those correlated photos with the same class label close to each other in the final collage.



(a) Initialization (b) Before iteration (c) $\frac{1}{2}$ iteration (d) 1 iteration (e) 2 iterations (f) The final collage

Fig. 5. Canvas partition. (a) An initial arrangement of cutouts is given by the embedding results produced by t-SNE. Different colors of the cutouts indicate different class labels. (b) The packing result by performing canvas tessellation and subregion optimization. (c) Halfway of the first iteration: Shrinking the cutouts until no overlap exists. (d) The first iteration is completed. (e) After two iterations, the cutouts are packed compactly and those of the same class are gathered together. Misplaced cutouts have been swapped using the Hungarian algorithm. (f) A seamless collage is obtained, with blending applied between adjacent images.

### 5.1 Problem Formulation

Given $N$ cutouts $\{R_i\}_{i=1}^N$ of the input photos with class labels and their initial placement on the canvas, our goal is to partition the canvas, $C$, in such a way that all the cutouts are tightly arranged on it without overlapping, while those cutouts with the same class label are grouped together on the canvas. We allow the cutouts to have arbitrary shapes, not necessarily convex. We measure tightness by the coverage rate, defined as the ratio of the sum of areas (number of pixels) covered by the salient cutout regions to the total area of the canvas. Thus, our task is to seek an optimal arrangement of cutouts to maximize the coverage rate of the canvas while maintaining the clusters of cutouts in the same class.

We assign a transformation, denoted as $\mathbf{T}_i = \{\mathbf{t}_i, \theta_i, s_i\}$, to each cutout, where $\mathbf{t}_i$, $\theta_i$ and $s_i$ denote translation, rotation, and scaling, respectively. This transformation is obtained from optimization based on the current status of the $n$ cutouts $R_i$. It is used to transform $R_i$ into a better configuration that increases coverage. Note that rotation $\{\theta_i\}$ is optional in our method, since rotated photos may not be preferred by users. Furthermore, the arrangement needs to meet the constraint that the cutouts should not overlap with each other.

### 5.2 Optimization

The problem described above is essentially a variant of the packing problem that allows a limited degree of scaling. Inspired by the work of surface mosaic synthesis with irregular tiles [28], we adapt its continuous optimization approach to 2D planes to solve our optimization problem. To group correlated photos together, we initialize positions of cutouts based on photo correlation and maintain position correlation in each iteration.

We iteratively update cutout configuration $\mathbf{T}_i$, with the goal of increasing canvas coverage and maintaining position correlation. We start with an initial arrangement of cutouts of sufficiently small sizes on the canvas $C$. More specifically, we scale the t-SNE result to the size of 2D canvas as the initial arrangement and all cutouts are then shrunk until no overlap exists among cutouts. In each iteration, the canvas is partitioned into a set of non-overlapping regions with each subregion containing one cutout. We optimize the transformation of each cutout independently to increase its coverage within its containing subregion. Then we perform necessary shuffling to make correlated photos stay close to each other. The pseudo-code of our canvas partition algorithm is given in Algorithm 1. The process is shown in Fig. 5, and we explain the details in the following section.

#### 5.2.1 Canvas Tessellation and Subregion Optimization

With the initial arrangement of cutouts, we would like to partition the canvas into a set of non-overlapping subregions with each subregion containing one cutout. We sample points on the boundary of the polygonal cutout tiles and then build a triangulation with these sampling points as vertices using Delaunay triangulation. Afterwards, chordal axis transformation (CAT) [29] is extracted from the triangulation. We call these polygonal regions as CAT regions. In Fig. 6, we show a cutout, marked in blue, and its corresponding CAT region, marked in gray.

**Algorithm 1.** Irregular Shape Based Canvas Partition

**Input:** $C$, Canvas
$\alpha$, a threshold for checking coverage increase
$\{R_i\}_{i=1}^N$, a set of cutouts with class labels
$T_0$, the t-SNE result
$N_{max}$, the maximum iteration number
**Output:** Partition of canvas $C$ and an arrangement of cutouts

```
Scale T_0 on 2D canvas and shrink the input cutouts until no overlap exists.
Do canvas tessellation and subregion optimization.
Compute the coverage rate as A_0.
repeat
    Reshuffle the positions of misplaced cutouts to guarantee that all cutouts in same class have position correlation, then shrink cutouts until no overlap exists;
    Alternate the process of tessellating canvas and optimizing subregions, until the increase of coverage rate < α;
until No misplaced cutout can be found or the number of iteration N > N_max;
return The partition of C and arrangement of cutouts.
```

We then optimize the position, size and orientation (optional) of each cutout individually within its CAT region to maximize canvas coverage. To avoid overlaps, we restrict each cutout within its CAT region by constraining every point sampled from the cutout to the interior of the CAT region. Referring to Fig. 6, let $R_i$ and $Q_i$ denote the polygons of an cutout and its CAT region, separately, and let $\mathbf{p}_k$ denote a sampling point on the boundary of $R_i$. Let $E_k$ denote the subset of edges of $Q_i$ which intersect at least one triangulation edge adjacent to $\mathbf{p}_k$. Then we restrict $\mathbf{p}_k$ within $Q_i$ by constraining $\mathbf{p}_k$ in the inner sides of the edges in $E_k$.

In each iteration of maximizing the scaled copy of the cutout, the following objective function is maximized:

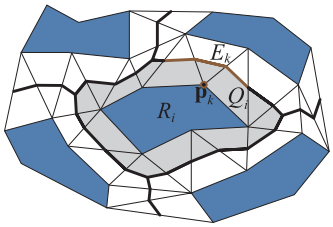



Fig. 6. Subregion optimization.

$$\begin{aligned} &\text{Maximize} \quad f(\mathbf{t},\theta,s)=s \\ &\text{subject to} \quad \Delta(\mathbf{p}_k(\mathbf{t},\theta,s),\mathbf{e}_j)\geq 0, \\ &\qquad \text{for} \;\; 1\leq k\leq M, \mathbf{e}_j\in E_k, -\frac{\pi}{6}-\Theta\leq\theta\leq\frac{\pi}{6}-\Theta, \end{aligned} \tag{1}$$

where $\mathbf{p}_k(\mathbf{t},\theta,s)$ is the new position of $\mathbf{p}_k$ by applying the transformation $\mathbf{T}=\{\mathbf{t},\theta,s\}$. $\Delta(\mathbf{p}_k,\mathbf{e}_j)$ is the signed area of the triangle specified by the point $\mathbf{p}_k$ and the edge $\mathbf{e}_j$. $\Delta(\mathbf{p}_k,\mathbf{e}_j)>0$ and only if $\mathbf{p}_k$ is in the inner side of $\mathbf{e}_j$, and $\Theta$ is the sum of incremental rotation angles in the former iterations.

Note that the objective function is simply the scaling factor, which measures the area coverage. In [28], the rotation angle is restricted to a range for every iteration to satisfy the local containment constraint for $R_i$, while we constrain the sum of incremental rotation angles for all iterations to the range as $[-\frac{\pi}{6},\frac{\pi}{6}]$, because the transformed $R_i$ should not rotate too much with respect to the original orientation to avoid overly tilted display of photos. Optionally, the user can switch off rotation by setting $\theta$ as 0.

We solve the above optimization problem efficiently using the interior point method. When rotation variable is fixed, the problem is reduced to a linear programming problem with the absence of $\theta$. In our experiments, the transformation is initialized to be $\mathbf{T}=\{\mathbf{t},\theta,s\}=\{(0,0),0,1\}$.

We compute the CAT regions again after updating the layout of cutouts, and alternate between subregion optimization and canvas partitioning iteratively until the increase of coverage rate is smaller than a threshold.

### 5.2.2 Position Reshuffling

Sometimes, due to erroneous initialization by t-SNE or drifting during canvas tessellation and subregion optimization in the preceding steps, some cutouts may get mixed up with the cutouts in other classes, resulting in incorrect clustering. We reshuffle those misplaced cutouts to improve the clustering as follows.

First, the misplaced cutouts need to be detected. Specifically, we first compute the center of every class by averaging the centroids of cutouts in the class. For each cutout, we compute the distance between the centroid of the cutout and the center of its class. For each class, we set a distance threshold. If the distance of a cutout to its class center is larger than the threshold of its class, the cutout is considered as misplaced. Empirically, the threshold of a class is set as 1.5 times the distance between the averaged centroid of cutouts in this class and the class center.

Next, we rearrange the misplaced cutouts, if there are any. Suppose that $N$ misplaced cutouts are taken out to create $N$ vacant positions on the canvas. We compute the distance between each vacant position and the center of the class of each misplaced cutout. We then want to find the optimal match between the $N$ misplaced cutouts and the $N$ vacant positions to minimize the sum of these distances. This is an instance of the perfect matching problem that can be formulated as follows:

$$\text{Minimize} \quad f([x_{ij}]_{N\times N})=\sum_{i=1}^{N}\sum_{j=1}^{N}c_{ij}x_{ij}, \tag{2}$$

where $[x_{ij}]_{N\times N}$ is constrained to be a permutation matrix, and $c_{ij}$ is the distance from the vacant position $j$ to the center of a class containing the misplaced cutout $i$.

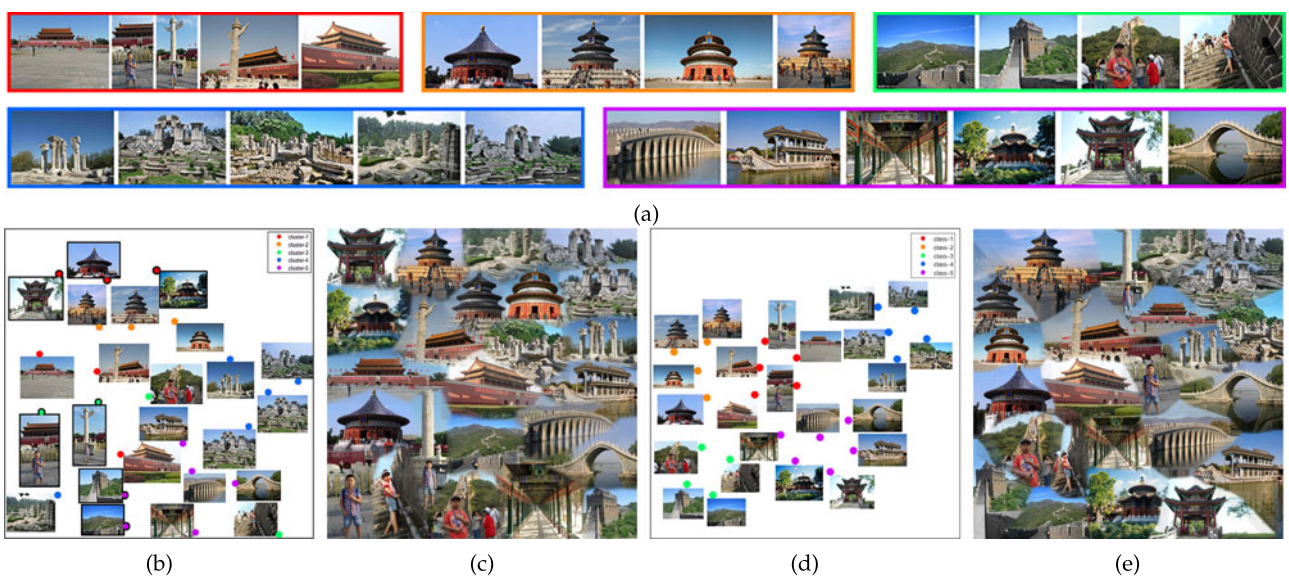
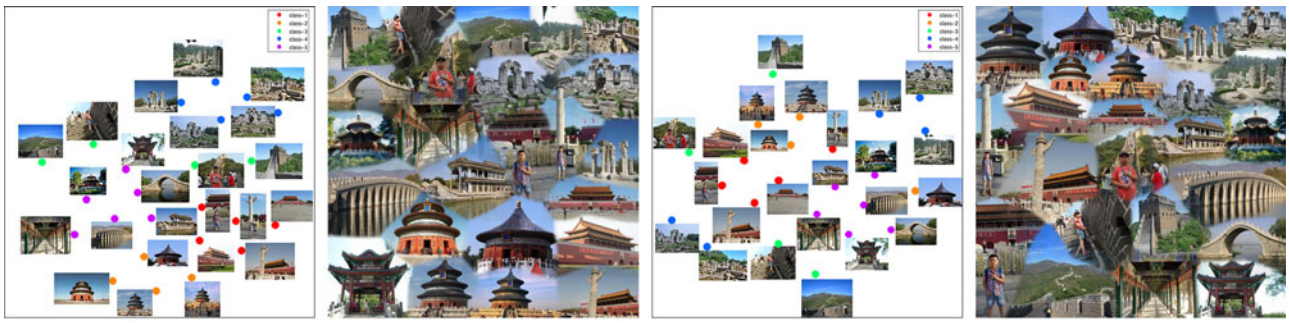



Fig. 7. Comparisons of collage results for the photo collection in (a). (b) and (c) show the initial embedding and the resulting collage without the additional network for classifying traveling photos. (d) and (e) show the initial embedding and the resulting collage with the additional network for classifying traveling photos. In (b) and (d), the color of a point indicates the label of the photo next to it, and those misclassified photos are framed with black boundaries.



(a) (b) (c) (d)

Fig. 8. Comparisons of the collage results generated with two different t-SNE results as initialization. (a) and (c) are photo embedding results, and (b) and (d) are the corresponding collages.

This is the classical assignment problem. Since $N$ is small in our problem, the solution can be efficiently found using the Hungarian algorithm [30].

After optimizing the positions of the misplaced cutouts, we shrink all the cutouts to a size small enough to avoid overlaps. Figs. 5b and 5c show the results before and after reshuffling and shrinking. We iterate the above process of canvas partition and position reshuffling, until no misplaced cutout can be found or the number of iterations is larger than a threshold $N_{max}$. In our experiment, $N_{max}$ is set to be 10. We finally display the cutout of each image in the region it is assigned to. The final seamless collage is obtained after applying blending [8] to smoothen the shared boundary between adjacent images, as shown in Fig. 5f.

Note that the above reshuffling scheme would be inadequate if there is only one misplaced photo, since there does not exist another misplaced one to swap with. Such a case may occur, though only infrequently. The problem could be resolved by performing a chain of swaps to move the misplaced piece to the cluster it should belong to. However, balancing simplicity and generality, we have not implemented this feature. So our current implementation does have this limitation of being unable to deal with the case where there is only one misplaced photo.

## 6 Experiments

### 6.1 Our Results

We have tested our approach on groups of photos of a variety of themes, such as animals, models, and traveling. The collages produced by our method are shown in the last column of Fig. 9 and the third column of Fig. 10. It can be seen that our method produces informative, compact and visually pleasing results for images of both general themes and specific traveling themes.

*Collage of General Themes.* In Fig. 9, the collage results of photos of different animals, including dogs, cats, monkeys, elephants, and rabbits are shown in the first row. It can be seen that the collage produced by our method is compact and free of visual artifacts. Boundaries between adjacent images are appropriately positioned. The same kind of animals, such as dogs or elephants, are grouped together and are packed closely in the final collage. This makes the themes of the photo collection clear and easy to grasp.

The collage result for models by our method is shown in the last column of the second row. In our result, photos of the models in similar costumes are properly clustered. The results of different collage methods for different transportation vehicles are shown in the third row. In our result, photos of the same kind of transportation vehicles are clustered together and cutouts of all photos are displayed in an intact manner.

*Collage of Traveling Photos.* For the traveling photos as shown in Fig. 10, we further train an additional network following the steps as described in Section 4. We present the trained networks for five cities: San Francisco, Paris, Rome, Beijing and Hong Kong. The CNN model integrated with an additional network proves to be able to extract more accurate semantics for attraction-specific themes than without the additional network, as shown in Fig. 7. A simple measure of classification performance is the number of misclassified images which are marked with black frames in Figs. 7b and 7d. And in Fig. 7d, the images with the same label are grouped together, thus preserving the correlation of photo contents. In contrast, this property is not observed in Fig. 7b.

In Fig. 8, we give two more examples of using the CNNs integrated with an additional network to show the advantages of our extended network. It can be seen that correlation-preserving photo collages are obtained even with different t-SNE results as initialization. t-SNE, as a local optimization method, cannot guarantee that the same optimal result is obtained for each different runs. Note that the scheme of reshuffling misplaced photos proves effective in improving the result with respect to image labels.

In Fig. 10, our result in the first row shows a photo collage of a tour to San Francisco. Photos of the same attractions in the city are packed closely in the result (left to right and top to bottom: the Lombard Street, Palace of Fine Arts Theater, Civic Center, Union Square, Golden Bridge, and Fisherman's Wharf). The results in the second row summarize the photos during a couple's holiday in Paris. In our result, photos of those tourist attractions visited are grouped together and prominently displayed on the final collage (left to right and top to bottom: Notre Dame Cathedral, Louvre, Place de la Concorde, Eiffel Tower, and Arc de Triumphal). Our result in the third row shows the collage for photos of a girl's visit to Rome. To train the attraction-specific network, Colosseum, Pantheon Rome, St peter's basilica, Piazza Navona, Piazza Venezia, and Roman Forum are used as the attraction labels, and photos crawled from the Internet are used for training. In the collage by our


Fig. 9. The collage results for photos of general themes, including animals (top), models (middle), and transportation vehicles (bottom). From left to right: The results by Microsoft AutoCollage, circle packing, a variant of method without content-correlation preservation, and our correlation-preserving method, respectively. The last two are provided to help appreciate the importance of preservation content correlation in collage presentation.

method, photos on the same attractions stay close. This shows that the combinatorial optimization works as expected and our classifier can group the photos into different attraction-specific classes.

*Collage Using Time Information.* Time can be useful in generating the collages for a group of photos taken in the order of time for story telling, for instance, the photos recording the growth of a child as shown in Fig. 11. Often time information is available from the EXIF metadata of images. Therefore one may cluster photos according to the time information, in addition to appearance features.

To achieve this, we first classify photos by different time slots according to the time information. The temporal feature of a photo is encoded as an indicator vector in which all elements are set as 0 except that one element is 1, and this indicator is used to uniquely identify the time slot. For each photo, the indicator vector is concatenated with the 256D feature vector generated by CNNs. To make elements of the concatenated vector commensurable, each element is further normalized by subtracting the mean and dividing the standard deviation of the corresponding elements of all input photos. Next, we use $k$-means to group photos into different semantic clusters. The remaining steps to generate the final collage are the same as the steps without using the time information. Fig. 11 compares the collages generated with and without time information, showing that time information indeed helps produce a better clustered photo collage.

*Collage with Photo Rotation.* Our approach can be used to generate photo collage with photo rotation. To achieve this, we only need to feed the rotated cutouts of photos into our region partition algorithm. Fig. 12 shows such an example. In this example, we set the rotation in the range between $-\frac{\pi}{6}$ to $\frac{\pi}{6}$ to avoid severely cluttered result.

*Timing of Computation.* All our results are generated on an Intel Corei7 CPU@3.46 GHz PC with 24G RAM. The running time mainly depends on the number of photos and image resolutions. To speed up processing, saliency detection and canvas partition are performed on down-sampled images and the scaled canvas, and the original images are used for image mapping. We implemented our approach on GPUs since the major computation of our framework including saliency detection and image mapping can be easily parallelized. For a group of 20 photos, it takes around 20 seconds to create the collage.


Fig. 10. The collages of traveling photos for some cities. From top to bottom: San Francisco, Paris, and Rome. From left to right: The results by Microsoft AutoCollage, circle packing, our correlation-preserving method, and our result repeated with cluster boundaries highlighted, respectively.

### 6.2 Comparisons

AutoCollage [7] and the collage method based on circle packing (which is abbreviated as *circle packing* in the following) [8] produce collages with the similar style. So we compare with these two methods. AutoCollage is part of Microsoft AutoCollage,[1] which is a plug-in of Windows Live. The source code of circle packing [8] was provided by the authors of the original paper [8]. Figs. 9 and 10 show the results by AutoCollage, circle packing, and our method.

*Comparison with AutoCollage.* In the results of AutoCollage, some important foreground objects are severely occluded by their neighbors, as observed in the first columns of Figs. 9 and 10. For instance, in the AutoCollage result of the "animals" theme, some animal faces are occluded, while a large area of background (near the bottom) are included in the collage. In the result of models, the head of the model wearing in grey in one image (middle right) is mixed with its neighbor.

As aforementioned, AutoCollage minimizes an objective function quantifying the criteria of a good collage with certain constraints. The optimization procedure may get stuck in poor local minima and so produces sub-optimal results. This leads to the noticeable visual artifacts shown in their results. In contrast, salient regions are displayed without occlusions in our results, due to our use of the divide-and-conquer strategy. That is, given the cutouts of all photos we first partition the canvas into a set of disjoint subregions. The cutout of each photo is guaranteed to be visible by displaying it in the subregion it is assigned to.

*Comparison with Circle Packing.* The cutout of each image is approximated as a circle by the circle packing method, and such an approximation leads to serious problems when it is applied to cutouts containing elongated objects. As can be seen from the second columns of Figs. 9 and 10, most collage results by circle packing either contain too much redundant background or have excessive trimming. For example, the bodies of most models cannot be completely displayed.

Our method allows photo cutouts to be of arbitrary shapes, which results in better utilization of canvas space and better fitting of cutouts with their containing subregions, as shown in our result. As a result, the cutouts of all models remain intact after packing. In contrast, in the collages by

1. http://microsoft-autocollage-2008.en.softonic.com/download.

Fig. 11. Top: Photo collages generated with (left) and without (right) time information. Bottom: The corresponding results with cluster boundaries highlighted in red.

circle packing of the animals and transportation vehicles shown in Fig. 9, the heads of some animals and the bodies of steamships are occluded by their neighboring cutouts.

*Benefit of Preserving Content Correlation.* Our approach has the additional advantage of preserving photo correlations. To show the significance and necessity of preserving photo correlation, we also compare it with a variant of our method without content correlation as shown in the third column of Fig. 9. This variant is obtained by setting a random photo arrangement as initialization and partitions the canvas without using position reshuffling. It is obvious that photos packed by Autocollage, circle packing, and this version with enforcing correlation appears disorganized in terms of photo theme. That is, photos of different themes are mixed together.

In contrast, correlated photos are packed closely by our method. In Fig. 9, for our result of transportation vehicles, the semantic clusters from left to right and from top to bottom are easily recognized: "aircrafts", "automobiles", "trains", and "ships", where different clusters are delineated with neat boundaries. The advantages of our correlation-preserving collage can also be observed in the collages of animals (left to right and top to bottom: dogs, monkeys, rabbits, cats, and elephants) and models in different kinds of costume. In Fig. 10, our results with cluster boundaries highlighted are given in the last column, which clearly demonstrate that photos in the same classes are clustered together by our approach.

### 6.3 User Study

We conducted a user study to evaluate the performance of our method. A web interface was designed for the user study. Fifty people who are unfamiliar with our research participated in the survey. Each participant was shown the ten groups of collages with several different themes, such as

Fig. 12. Our collage results on animals with (left) and without (right) photo rotation.

animals, models, transportation vehicles, and traveling in Paris, Rome, San Francisco, and Beijing, etc. Six of them are shown in Figs. 9 and 10, and the rest examples are included in the supplemental materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2017.2703853.

The user study consists of two parts. In the first part, we compare our method with two other methods, AutoCollage and circle packing [8]. In the second part, we display the results with and without content correlation and ask the subject which one looks better.

*Part I.* The first part was designed to find out the subjects agree that whether our approach is informative, visually appealing and superior over competitive methods. For each photo group, we first showed the participant the input photos, followed by three results produced by the three methods in a random order. For each collage, the participant was asked to answer the questions Q1∼Q3 by giving a numerical score from 1 to 5, with 1 for *strongly no* and 5 for *strongly yes*. The last question of each group asks the subjects to choose the best collage from the three collages.

1) Is it good to preserve the foreground of each image?
2) Do you find the summary visually appealing?
3) Is it a good visual summary of the set of photos?
4) Which one do you prefer overall?

Fig. 13 shows the averaged scores on questions Q1∼Q3 and the normalized votes on Q4 in percentage for the ten collage groups. Overall, our method is preferred to AutoCollage and circle packing on all the ten groups.

We further analyze the results of the user study statistically using a paired-sample, two-tailed $t$-test. In the paired-sample $t$-test with the mean of zero, a small $p$-value ($p \leq 0.05$) indicates a significant difference between two groups of data. As shown in Table 1, there is a statistically significant difference in subjects choosing our method over the other two methods on each of the four questions.

For the results of building and fish, the users' average rating on AutoCollage is relatively high among ten collage groups. It may be due to similar appearance of the same objects. These photos of the same objects were taken from different directions and they appear similar because of the symmetric structures.

For the results of San Francisco, the users' votes on AutoCollage is relatively close to the rating on ours on Q1∼Q3. The similar case holds for the results on Paris and Rome. Looking at the photos of Rome, most of the famous

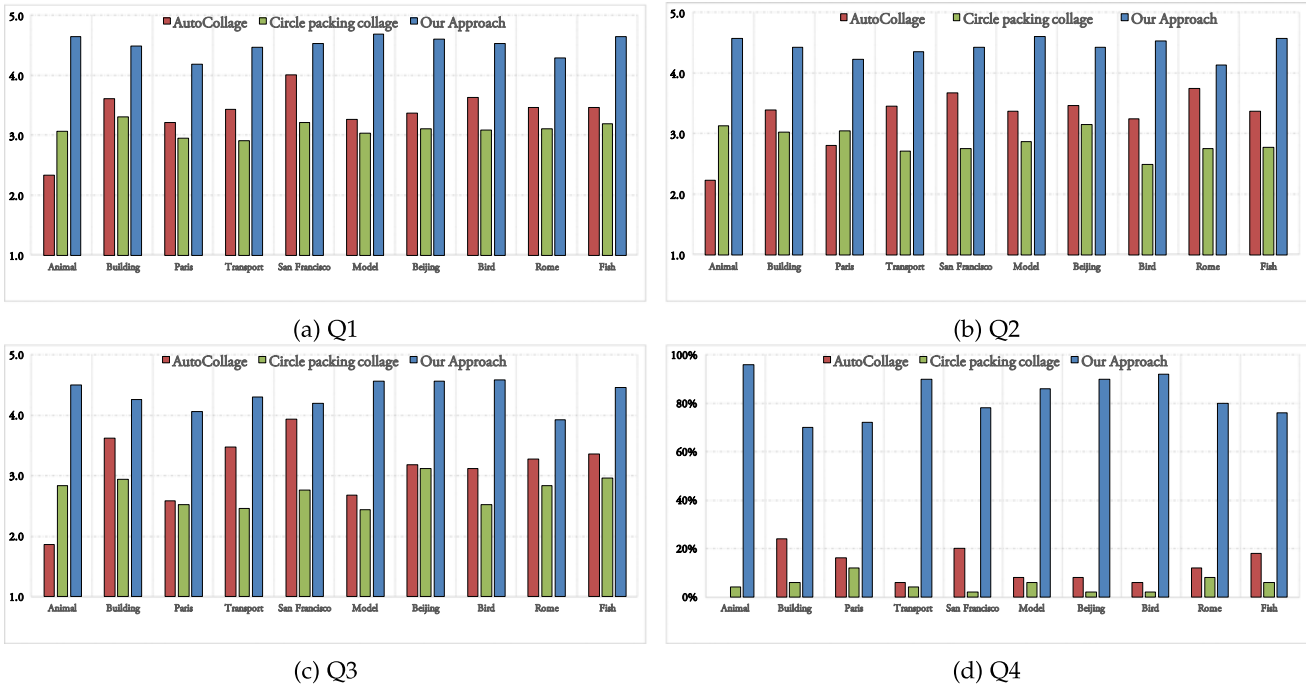



Fig. 13. Results of Part I. Normalized votes for the ten collage groups on questions Q1 (a), Q2 (b), Q3 (c), and Q4 (d).

landmarks such as Colosseum and Pantheon have complex structures. The collages then look cluttered and disordered, making it difficult for the participants to make choices.

In particular, we told the participants which collages were ours after the study, and further asked several participants the reasons of their choices. Some participants said that our collages were visually more pleasing and had a better utilization of the limited canvas space, such as the collages for the animals, models, transportation vehicles, and most traveling photos. In comparison, visual artifacts in the results of AutoCollage and circle packing, for example, the occluded or truncated foreground on photos in Figs. 9 and 10, influenced their choices. The participants agreed that a satisfactory collage should be assembled seamlessly and compactly, without noticeable visual artifacts. Furthermore, they stated that our results were easy-to-follow because the similar photos were grouped together.

We further asked some participants how they compared the results of San Francisco, Paris and Rome. Most of them said that they were not familiar with the three cities. Some felt that the contents of the collages produced by all the three methods were somewhat too cluttered so it was a bit difficult for them to choose the best one. What they said coincides with our analysis on the reason why users' ratings of Autocollage and our method were relatively close on these three cities, though our results are still clearly preferred.

*Part II.* The second part of the user study is to verify whether users prefer to have collages with similar contents grouped together. For the ten collage groups, the participant was shown the correlation-absent collage and the correlation-preserving collage in a random order, and he was required to choose which one he liked. The normalized votes for the ten groups are shown in Fig. 14. And the $p$-value of a paired-sample, two-tailed t-test is $4.56e-7$ on the two results with or without content correlation. We further asked the participants why they preferred the correlation-preserving collages. They stated that the version with content correlation was much easier to catch the theme at the first glance and to understand the components in the collage.

The couple in the traveling photos of Paris said they liked the correlation-preserving collage much better than collages without content correlation. They also said that they had a much easier time recalling their happy moments in the trip when viewing the correlation-preserving collage. Throughout this part of the user study, the users' preference on the correlation-preserving collages and their reasons reflect the superiority of clustering similar photos in the collage.

TABLE 1
The $p$-Value of a Paired-Sample, Two-Tailed t-Test

| Questions | Q1 | Q2 | Q3 | Q4 |
|---|---|---|---|---|
| AC versus Ours | 5.69e-05 | 3.42e-04 | 4.34e-05 | 2.28e-07 |
| CP versus Ours | 1.78e-08 | 8.40e-08 | 8.62e-10 | 4.01e-09 |

*AP: AutoCollage; CP: circle packing; Ours: our approach.*

## 6.4 Limitations

As stated above, a clear limitation is that our current implementation is unable to deal with the case where there is only one misplaced photo after position reshuffling in the step of canvas partition. Another limitation of our method is the lack of consideration of composition aesthetics. It

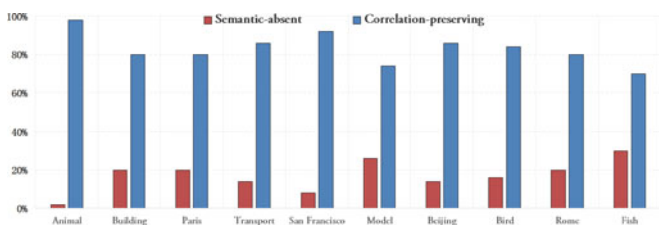



Fig. 14. Results of Part II. Normalized votes for the ten collage groups.

(a) (b)

Fig. 15. Failure cases. (a) Incorrect object sizes. (b) A bad clustering result for furniture images with highlighted (red) cluster boundaries (left to right, top to bottom: Decorations, vases, lights, tables, and chairs), and the boundaries of misclassified cutouts are marked in blue.

would be useful to take into account those popular composition rules that are frequently used by professional photographers, such as rule-of-thirds, the consistency of relative component sizes, etc. This would help further improve our collage results.

Even though our correlation analysis is built upon the state-of-art deep learning techniques, it may still produce incorrect image classification. The accuracy of classification will affect our results as shown in Figs. 7b and 15b. Our results can be improved as the techniques on object recognition and image classification continue to develop.

## 7 Conclusion

We have presented a new method for generating photo collages that are compact, informative and visually pleasing. Our method is capable of grouping photos in the 2D canvas space according to the semantic correlation of photo contents. We believe this feature would greatly enhance the efficiency for a viewer to appreciate the collage content. We pack extracted salient regions of the photos directly and allow the the salient regions to have arbitrary shapes. This leads to efficient use of the limited canvas space. The effectiveness of our method is validated through extensive comparisons against previous methods and by the user study.

## Acknowledgments

Yanwen Guo's work was partially supported by the National Natural Science Foundation of China (61373059 and 61672279) and the NSF of Jiangsu Province (BK20150016). Zhonggui Chen's work was partially supported by National Natural Science Foundation of China (61472332). Wenping Wang's research was partially supported by the Research Grant Council of Hong Kong (17208214). Yanwen Guo is the corresponding author.

## References

[1] 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.

[2] 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.

[3] Y. Wei, Y. Matsushita, and Y. Yang, "Efficient optimization of photo collage," Microsoft Research, Redmond, WA, USA, MSRTR-2009-59, 2009.

[4] Y. Yang, Y. Wei, C. Liu, Q. Peng, and Y. Matsushita, "An improved belief propagation method for dynamic collage," *Visual Comput.*, vol. 25, no. 5–7, pp. 431–439, 2009.

[5] A. Agarwala, et al., "Interactive digital photomontage," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 294–302, 2004.

[6] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake, "Digital tapestry," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, pp. 589–596.

[7] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake, "Autocollage," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 847–852, 2006.

[8] Z. Yu, L. Lu, Y. Guo, R. Fan, M. Liu, and W. Wang, "Content-aware photo collage using circle packing," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 2, pp. 182–195, Feb. 2014.

[9] P. Sinha, H. Pirsiavash, and R. Jain, "Personal photo album summarization," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 1131–1132.

[10] 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.

[11] P. Clough, H. Joho, and M. Sanderson, "Automatically organising images using concept hierarchies," in *Proc. Multimedia Workshop Running ACM SIGIR Conf.*, 2005, pp. 33–39.

[12] A. Jaffe, M. Naaman, T. Tassa, and M. Davis, "Generating summaries for large collections of geo-referenced photographs," in *Proc. 15th Int. Conf. World Wide Web*, 2006, pp. 853–854.

[13] I. Simon, N. Snavely, and S. M. Seitz, "Scene summarization for online image collections," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[14] 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.

[15] T. Chen, A. Lu, and S.-M. Hu, "Visual storylines: Semantic visualization of movie sequence," *Comput. Graph.*, vol. 36, no. 4, pp. 241–249, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0097849312000337

[16] G. Jing, Y. Hu, Y. Guo, Y. Yu, and W. Wang, "Content-aware video2comics with manga-style layout," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2122–2133, Dec. 2015.

[17] L. Zhang and H. Huang, "Hierarchical narrative collage for digital photo album," *Comput. Graph. Forum*, vol. 31, no. 7, pp. 2173–2181, 2012.

[18] S. Goferman, A. Tal, and L. Zelnik-Manor, "Puzzle-like collage," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 459–468, 2010.

[19] J. Kim and F. Pellacini, "Jigsaw image mosaics," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 657–664, 2002.

[20] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. no. 124.

[21] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiology*, vol. 160, no. 1, 1962, Art. no. 106.

[22] C. Szegedy, et al., "Going deeper with convolutions," in *Proc. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9. [Online]. Available: http://arxiv.org/abs/1409.4842

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Int. Conf. Learn. Representations*, vol. abs/1409.1556, 2015.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 770–778, Jun. 2016, doi: 10.1109/CVPR.2016.90.

[25] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

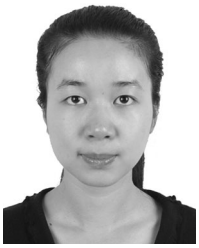
[26] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2083–2090.

[27] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

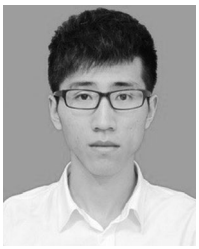
[28] W. Hu, Z. Chen, H. Pan, Y. Yu, E. Grinspun, and W. Wang, "Surface mosaic synthesis with irregular tiles," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 3, pp. 1302–1313, Mar. 2016. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2015.2498620

[29] L. Prasad, "Morphological analysis of shapes," *CNLS Newslett.*, vol. 139, no. 1, pp. 1997–2007, 1997.

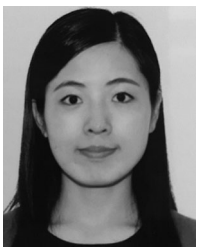
[30] E. Lawler, *Combinatorial Optimization: Networks and Matroids*. Fort Worth, TX, USA: Saunders College Publishing, 1976.



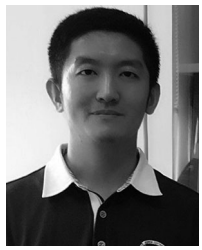
**Lingjie Liu** received the BEng degree in computer science and technology from the Huazhong University of Science and Technology. She is working toward the PhD degree in computer science supervised by Prof. Wenping Wang with the University of Hong Kong. Her research interests include 3D reconstruction and image processing.



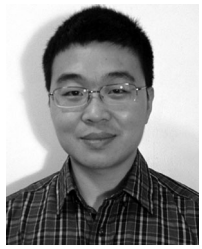
**Hongjie Zhang** is working toward the PhD degree in the National Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Jiangsu, China. His research interests include computer vision and graphics, digital image processing, and pattern recognition.



**Guangmei Jing** received the BEng degree from the University of Science and Technology of China, Hefei, China, in 2010 and the PhD degree from the University of Hong Kong, in 2015. She worked as a research assistant with the State Key Laboratory of CAD & CG, Zhejiang University, Hangzhou, China, in 2010. Her research interests include image/video processing and computer vision.



**Yanwen Guo** received the PhD degree in applied mathematics from the State Key Lab of CAD & CG, Zhejiang University, China, in 2006. He is currently a full professor in the National Key Laboratory for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Jiangsu, China. He was a visiting professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, in 2006 and 2009, respectively, the Department of Computer Science, The University of Hong Kong, in 2008, 2012, and 2013, respectively, and a visiting scholar in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, from 2013 to 2015. His research interests include image and video processing, vision, and computer graphics.



**Zhonggui Cheng** received the BSc and PhD degrees in applied mathematics from Zhejiang University, in 2004 and 2009, respectively. He is an associate professor in the Department of Computer Sciences, School of Information Science and Engineering, Xiamen University, China. His research interests include computer graphics and computational geometry.

**Wenping Wang** received the PhD degree in computer science from the University of Alberta, in 1992. He is chair professor and head of Computer Science Department, University of Hong Kong. His research covers computer graphics and geometric computing. He has published more than 120 journal papers in these fields. He is a fellow of the IEEE.