Measuring and Predicting Visual Importance of Similar Objects

Yan Kong, Weiming Dong, *Member, IEEE*, Xing Mei, *Member, IEEE*, Chongyang Ma, Tong-Yee Lee, *Senior Member, IEEE*, Siwei Lyu, *Member, IEEE*, Feiyue Huang, and Xiaopeng Zhang, *Member, IEEE*

Abstract—Similar objects are ubiquitous and abundant in both natural and artificial scenes. Determining the visual importance of several similar objects in a complex photograph is a challenge for image understanding algorithms. This study aims to define the importance of similar objects in an image and to develop a method that can select the most important instances for an input image from multiple similar objects. This task is challenging because multiple objects must be compared without adequate semantic information. This challenge is addressed by building an image database and designing an interactive system to measure object importance from human observers. This ground truth is used to define a range of features related to the visual importance of similar objects. Then, these features are used in learning-to-rank and random forest to rank similar objects in an image. Importance predictions were validated on 5,922 objects. The most important objects can be identified automatically. The factors related to composition (e.g., size, location, and overlap) are particularly informative, although clarity and color contrast are also important. We demonstrate the usefulness of similar object importance on various applications, including image retargeting, image compression, image re-attentionizing, image admixture, and manipulation of blindness images.

Index Terms—Similar objects, visual importance, listwise ranking

1 Introduction

Similar objects are ubiquitous in both natural and artificial scenes. They can be also found in many different photographs and art works, particularly those that feature plants, animals, food, and architecture (Fig. 1). Identifying and extracting these objects from images are useful for biological and artificial vision systems. In particular, the spatial distribution and visual appearance of similar objects in a scene may vary significantly. Thus, the relationship between these objects is analyzed to provide strong cues for high-level scene understanding and to facilitate many useful applications in image processing, computer vision, and computer graphics.

Capturing multiple instances of objects of the same class in an image has become an active area of research. Many approaches have been proposed for automatic, robust, and/or accurate extraction of salient regions and multiple similar objects from the input photos [1], [2], [3], [4]. However, most of these methods do not analyze further the extracted similar objects in terms of their relative distribution and appearance variation in the scene. Thus, these methods have limitations in enabling a comprehensive understanding of the scenes and an intuitive manipulation of images.

In this study, we propose to quantitatively measure and predict the visual importance of similar objects in an image (e.g., Fig. 2), which has remained an open problem. We demonstrate the generality and applicability of visual importance for high-level image understanding and editing through various examples (Section 7). However, given the current image processing tools and computer vision techniques, computing the visual importance of similar objects is not easy. The reason is that annotating accurately different meaningful tags for similar objects in an image is usually difficult, and measuring their visual importance based on the existing saliency detection methods is not straightforward (Section 2).

Our key idea is to build a large database of images containing multiple similar objects and to compute the visual importance by learning from user annotated ranking data (see Fig. 3). In particular, we first collect a *dataset* of 808 example images (Section 3). The similar objects in each image are interactively segmented and manually ranked by 71 different users according to their visual observations about the relative object importance. Inspired by datasets such as ImageNet [5] and Places [6], our images are captured "in the wild" by collecting them from online repositories rather than collecting them in a laboratory to support real-world applications. Then, we develop a probabilistic model to *measure* the visual importance of similar objects from the collected user ranking data (Section 4). For each example image in the database, we compute the score of each object, which represents the relative visual importance within the image. We adopt the learning-to-rank algorithm

- *Y. Kong, W. Dong, X. Mei, and X. Zhang are with the NLPR-LIAMA, Institute of Automation, Chinese Academy of Sciences, Beijing, China. E-mail: kc4271@gmail.com, {Weiming.Dong, Xing.Mei, Xiaopeng.Zhang}@ia.ac.cn.*
- *C. Ma is with the University of Southern California, CA. E-mail: chongyang.ma@usc.edu.*
- *T.-Y. Lee is with the National Cheng-Kung University, Taiwan. E-mail: tonylee@mail.ncku.edu.tw.*
- *S. Lyu is with the University at Albany-SUNY, Albany, NY. E-mail: slyu@albany.edu.*
- *F. Huang is with the Tencent, China. E-mail: garyhuang@tencent.com.*

Manuscript received 25 May 2015; revised 23 Dec. 2015; accepted 29 Dec. 2015. Date of publication 0 . 0000; date of current version 0 . 0000.
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.2016.2515614

Fig. 1. Album of a Flickr member. Similar objects and repetitive patterns commonly occur in photography in daily life.

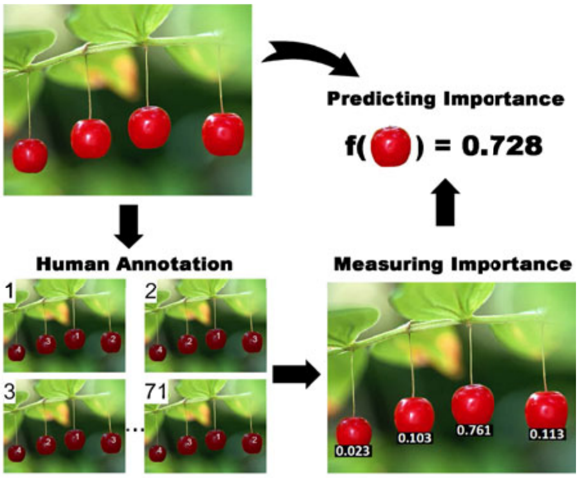


Fig. 3. Measuring and predicting importance of similar objects.

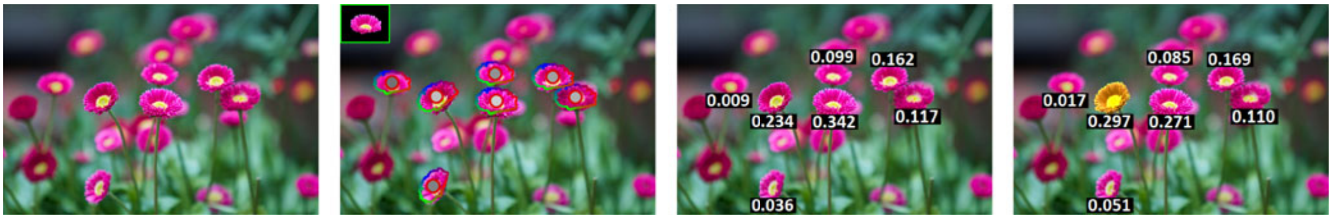
and a regression method with the measured data as the training set to *predict* the visual importance in a new image (Section 5). We further investigate what makes an object important among multiple similar instances and provide a range of features for qualitative analysis of object importance.

We demonstrate the effectiveness of our approach by comparing it to state-of-the-art saliency detection methods and two other baseline methods on importance prediction (Section 6). We also provide a wide range of high-level image manipulations enabled by our approach, which would otherwise be difficult or inaccurate to achieve via the existing methods, including image retargeting, image compression, image re-attentionizing, image admixture, and change blindness images (Section 7).

2 Related Work

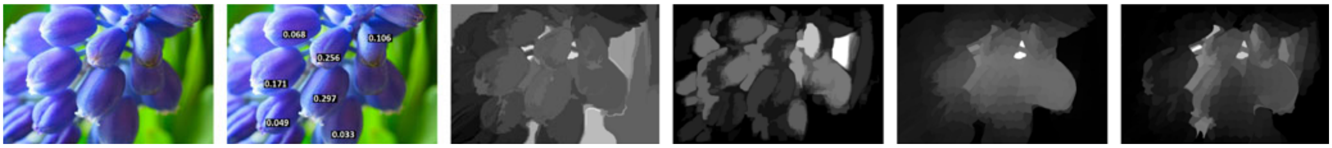
Similar object detection. Detection and segmentation of similar objects from an image have been extensively explored in recent years. Cheng et al. [1] presented a user-assisted approach in identifying approximately the repeated objects in an image, via boundary band matching to locate candidate elements and active contours to obtain object boundaries. Schweitzer et al. [7] presented a template matching technique based on Walsh transform. Huang et al. [8] presented a graph-based method to cut out repeated elements with similar colors. Xu et al. [9] designed a scribble-based tool to select interactively similar shapes by roughly stroking through the elements. Cai and Baciu [3] followed the classical region growing image segmentation scheme and used a mean-shift clustering to group local image patches. Kong et al. [2] used template matching to obtain candidate instances by separately using shape and color descriptors. Then they adopted a joint optimization scheme to decide the final locations. Erhan et al. [4] proposed a saliency-inspired neural network model for object detection that can capture multiple instances of the same class. Our work can benefit from the progress of this research direction.

Image importance prediction. Some recent studies focused on the problem of understanding and predicting the perceived importance of image content in a normal scene. Boiman and Irani [14] addressed the problem of detecting irregularities in visual data. A salient or visually irregular object can be detected from repetitive objects. Elazary and Itti [15] regarded object naming order in *LabelMe* dataset [16] as a measure for objects of interest and indicated that selecting interesting objects in a scene was largely constrained by low-level visual attributes. In ESP game [17], gamers help determine the content of images by providing meaningful labels. Intuitively, visually important elements are labelled earlier than less important ones in this process. Judd et al. [18] collected eye-tracking data to train a model of saliency using machine learning to predict regions or objects on which humans are interested. Spain and Perona [19] investigated factors to predict the order in which objects are mentioned. They also demonstrated that the relative semantic importance or saliency of an object and the order in which a user tags the image have a close relationship. Berg et al. [20] used descriptions written by people as indicators of importance, and several factors related to human perceptions, including attributes related to image segmentation and semantics, were proposed. Hwang and Grauman [21]



(a) Input Image (b) Similar object detection (c) Object visual importance (d) Image editing (re-attentionizing)

Fig. 2. After detecting and segmenting the similar objects from an input image, we can predict the visual importance of each object. This information can be used in several object-based image editing applications. The numbers on the objects represent the importance value.



(a) Original image (b) Predicted importance (c) GC [10] (d) BMS [11] (e) GMR [12] (f) RBD [13]

Fig. 4. Saliency detection results by state-of-the-art methods. When the principle part of an image is similar objects, it's difficult to evaluate the importance of each object by saliency detection, even not easy to get the most important one.

proposed an unsupervised learning procedure that discovered the relationship between how humans tag images (e.g., the order in which words are mentioned) and the relative importance of objects. Zitnick and Parikh [22] investigated the semantic information in abstract images created from collections of clip art. They measured the mutual information between visual features and the semantic classes to discover which visual features are most semantically meaningful and relatively important. Our task of predicting the importance of similar objects in a scene differs from that of these previous works in the following aspects: (1) Providing semantic descriptions for this kind of scenes as importance indicators is difficult for people. (2) Any auxiliary information, such as content labels or tags for importance study, does not exist. (3) Thus far, no suitable database for this specific task exists. (4) Our prediction results can be readily utilized in image-editing applications.

Saliency detection. Saliency detection aims to find the most informative and interesting region in a scene. Bottom-up saliency methods rely on some prior knowledge about salient objects and backgrounds, such as contrast and compactness. Different saliency methods characterize the prior knowledge from different perspectives. Cheng et al. [10] proposed to first segment the image and then compute the global contrast of those segments as the saliency, thereby highlighting the entire object regions. Yang et al. [12] classified saliency detection into a graph-based ranking problem, which propagates labels on a sparsely connected graph to characterize the overall differences between the salient objects and the background. Zhang and Sclaroff [11] characterized an image through a set of Boolean maps to measure the surroundedness cues for perceptual figureground segregation. Zhu et al. [13] introduced boundary connectivity to characterize the spatial layout of image regions with respect to image boundaries and then integrated multiple low-level cues into a principled optimization framework to obtain saliency maps. Vig et al. [23] generated a large number of instances of a richly parametrized bio-inspired hierarchical model family, and those that were predictive of image saliency were selected. In the present study, we compare our method to the aforementioned five state-or-the-art saliency detection methods for object importance ranking evaluation (see Fig. 4 and further comparisons in Section 6). We show that our method outperforms saliency in both predicting the importance permutation of the objects and deciding the most/least important object.

Similar pattern manipulation. Similar patterns in an image can be manipulated for various high level goals. After extracting similar objects from the input image, Cheng et al. [1] presented applications for image rearrangement, editing transfer, deformation propagation, and instance replacement. Ma et al. [24] synthesized repetitive elements according to a small input exemplar and a large output domain, which can preserve both individual element properties and their aggregate distributions. Zhang et al. [25] found separable objects in the target image based on the analysis of element separability. Then, they replaced individually structured objects from groups. Dong et al. [26] summarized the content of an image by carving out unimportant instances from a set of similar objects.

3 Similar Object Importance Dataset

Although some objects in an image can be identified easily as more important than others, quantifying this expectation has not been addressed. To learn to predict visual importance, we create the *Similar Object Importance Dataset* (**SOID**), a collection of real-world images consisting of multiple similar objects and a group of object permutations annotated in terms of object visual importance.

3.1 Dataset Collection

SOID is designed to cover visual scenes with enough diversity and to support real-world applications by containing visual content *in the wild*, i.e., images extracted from the Web rather than those captured or generated in a controlled setting. We have downloaded 1,106 images from internet by using keywords, such as "similar objects," "similarity," and "repetition," and by searching categories of photos that are likely to contain similar objects, such as flowers, animals, and food. Each image contains at least three similar objects that form the dominant or important content of the image. All images in the dataset are optionally resized such that both the widths and the heights are less than 640 pixels. We manually remove some noisy examples from the dataset if they do not contain similar patterns, have low quality, and/or show severely clustered objects. In this way, we end up with a dataset of 808 samples. Our collection of photos covers various topics, including plants, food, animals, humans, and other indoor and outdoor scenes. Fig. 5 shows some representative samples of these images. We select these scenes because they are common and represent the overall statistics of the collection.

We develop an interactive tool to segment the similar objects from each image by adapting *SimLocator* framework to facilitate the data annotation and analysis process [2]. For each image, we first select a template object from all the instances and use *SimLocator* to detect automatically and to segment other instances from the background. We also implement an interactive local selection interface based on Paint Selection [27] to refine the segmentation results of *SimLocator*. This object segmentation tool also boosts image manipulation applications for object-level editing.

Fig. 5. Representative examples of our images in Similar Object Importance Dataset. The photos all contain multiple similar objects.

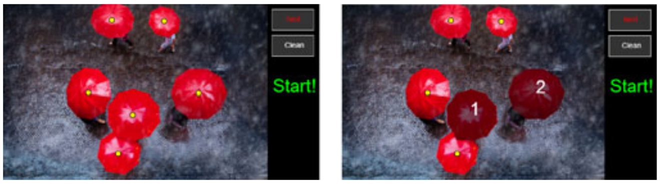
Given that a viewer cannot easily provide a semantic description or specify an accurate importance value for an object as in [20], we adopt a ranking model that people can easily participate in (Section 3.2). Then, we use a ranking model to measure the object importance based on human-annotated permutations (Section 4).

3.2 Human Annotation

As defined in [19], the *importance* of an object in a particular image is its probability to be noticed first by a viewer. Inspired by this definition, we design an interactive system for data collection. The images in our dataset are randomly displayed one by one. For each image, we extract all the similar objects with the help of our similar object detection system. Each object is highlighted by a yellow point (Fig. 6a). We ask the viewers to click the labeled objects in the image sequentially according to their visual importance from high to low (Fig. 6b). The participant should click first the object that he/she considers as the most important, and then the second most important one, until the least important one. After the participant finishes ranking the objects, he/she can click the "next" button to rank the next image, and we obtain a permutation for the similar objects in each image. In our system, we always lock the ranking interface for 3 seconds before the image can be annotated when the system switches to a new image to avoid the transition effect between images (i.e., the user tends to select the object near the last object in the previous image).

A total of 71 participants (38 men and 33 women, aged 20 to 45 years) from different backgrounds have performed the object ranking task. In the final form of SOID, each example data set is composed of the original image, the similar objects segmented from the image, and all the object importance permutations annotated by the viewers.

Subject agreement. We employ the nonparametric statistic Kendall's W (also known as Kendall's coefficient of concordance) [28] to evaluate the validation of the annotations by assessing the subject agreement for an image on the full importance order of the objects. Kendall's W is a normalization of Friedman test statistics, which can be used to assess agreement among raters. Its value ranges from 0 (no agreement) to 1 (complete agreement). For our problem, a similar object i is supposedly given the rank $r_{i,j}$ by the jth participant, where n objects and m participants exist. Then, the total rank given to object i is $R_i = \sum_{j=1}^{m} r_{i,j}$, and the mean value of these total ranks is $\overline{R} = \frac{1}{2} \cdot \frac{mn(n+1)}{n} = \frac{m(n+1)}{2}$. The sum of squared deviations, S, is defined as $S = \sum_{i=1}^{n} (R_i - \overline{R})^2$. Then, Kendall's W is defined as $W = \frac{12S}{m^2(n^3-n)}$.



(a) Interface

(b) User ranking

Fig. 6. Human annotation.

We randomly pick 50 images from SOID and show their Kendall's W values in Fig. 7. The average Kendall's W value of all the images in SOID is 0.566, which reaches the concordance rate of "moderate" ($W \in [0.5, 0.7)$). The value is not very high because some images have objects with visual importance rates that are not distinctly different. Therefore, the orders of these objects in the annotated permutations are likely to be different, thereby directly decreasing the value of Kendall's W. However, most subjects are concordant with the most and the least important objects. In particular, we first find in each image the object that is ranked as top-1 important the most times, and the object that is voted as last-1 important the most times. Then, we compute the percentages of the corresponding ranked times to the number of annotators, and the two percentages are used to evaluate the agreement of the subjects to the most and the least important object. The average values for all images in SOID are 0.690 and 0.686, which expresses the high confidence of the participants in choosing the most and the least important objects. In some images, the most and least important objects cannot be distinguished easily even by people. Thus, this level of agreement is good enough for our algorithm to obtain good measurement and prediction results.

4 Measuring Object Importance

The goal of this section is to quantitatively measure the visual importance of similar objects based on the collected SOID dataset. As discussed in [29], Bradley-Terry model [30] for pairwise comparisons is a simple and well-investigated method to describe the probabilities of the possible outcomes when individuals are judged against one another in pairs. We draw inspiration from this and adopt Plackett-Luce (PL) model [31], a generalized Bradley-Terry model that can handle multiple comparisons, to compute the visual importance of objects in an image.

The PL model assumes that ordering is a process of choosing the object without replacement according to its visual importance. By letting n, π_i, s_{π_i} denote the quantity of objects, the index of the ith important object chosen by the viewer, and the visual importance of object π_i, respectively, we determine the probability of choosing object π_i as follows:

$$\mathrm{P}(\pi_i) = \frac{s_{\pi_i}}{s_{\pi_i} + s_{\pi_{i+1}} + \cdots + s_{\pi_n}}. \qquad (1)$$

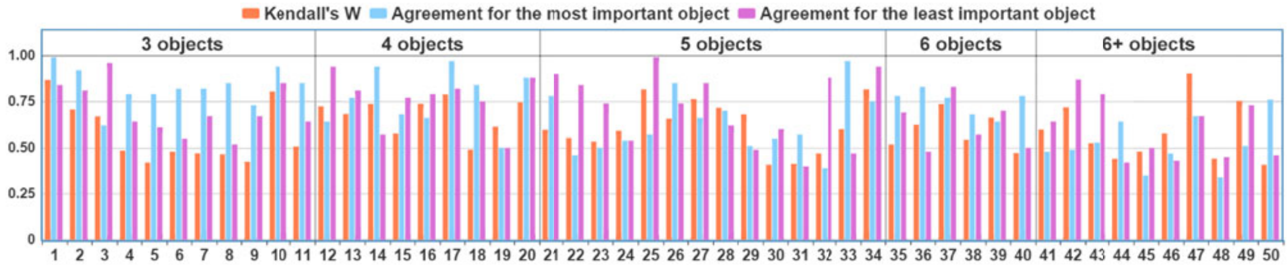


Fig. 7. Assessing subject agreement for annotations. We randomly pick 50 images from SOID and show their assessment values.

Equation (1) is a generalization of the Bradley-Terry model for the pairwise comparison of alternatives, which specifies the probability that "a wins against b" in terms of

$$\mathrm{P}(a \succ b) = \frac{s_a}{s_a + s_b}.$$

Intuitively, the larger the s_a is compared with s_b, the higher is the probability of choosing a. Likewise, the larger the parameter s_{π_i} in Equation (1) is compared with the parameters $\{s_{\pi_j}, j \neq i\}$, the higher is the probability of choosing the object π_i as more important.

Using the PL model, we formulate the probability of producing an object permutation π as

$$\mathrm{P}(\pi|s) = \prod_{i=1}^{n} \frac{s_{\pi_i}}{\sum_{j=i}^{n} s_{\pi_j}}. \tag{2}$$

For each image, by assuming that the user-annotated object orders $\boldsymbol{\pi} = \{\pi^{(1)}, \dots, \pi^{(K)}\}$ are provided by K different viewers independently, the probability of producing all these orders is

$$\mathrm{P}(\boldsymbol{\pi}|s) = \prod_{t=1}^{K} \prod_{i=1}^{n} \frac{s_{\pi_i^{(t)}}}{\sum_{j=i}^{n} s_{\pi_j^{(t)}}}.$$

Finally, we maximize the log-likelihood function to obtain the maximum likelihood estimation of the similar object importance s:

$$\log \mathcal{L}(s; \boldsymbol{\pi}) = \sum_{t=1}^{K} \sum_{i=1}^{n-1} \left[\log \left(s_{\pi_i^{(t)}} \right) - \log \left(\sum_{j=i}^{n} s_{\pi_j^{(t)}} \right) \right].$$

We use the minorization and maximization algorithm [29] to solve this problem.

Fig. 8 shows some examples of measuring the importance of similar objects in the given images. The objects with large numbers/scores are more important than the others. We can determine from these results some hints that help us design feasible features for object importance prediction. For example, we can see in the peaches image that the most important object is more visible than the other objects, whereas the two most important objects in the macaron image are in the middle of the scene.

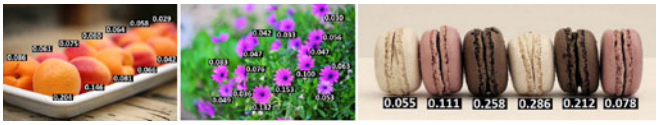


Fig. 8. Measured object importance. The numbers are object importance value which is normalized to $(0, 1)$.

5 Predicting Object Importance

Is predicting automatically the importance of each similar object from a photograph without manual annotation possible? We resort to data-driven approaches where object importance is predicted by the combination of several image features. We assume that good pattern recognition algorithms that can accurately detect and segment objects will emerge in the near future. Thus, we consider features that may be computed from the image once a contour of each object is available. We also try to determine what makes an object visually important in an image.

Given an image $\mathbf{I}$ and its n similar objects, our goal is to score and rank these objects according to their visual properties. We adopt the listwise learning-to-rank methodology [32] to rank the importance of each object. The previous experiments in the machine learning field have demonstrated that the listwise approach usually performs better than pointwise and pairwise approaches when dealing with ordering problems [32]. However, learning-to-rank approaches are more suitable for ordering than for scoring the objects, which may be inadequate for specific applications. Therefore, we also provide an effective regression tool based on the random forest to score object importance.

5.1 Features

We devise features to convey information on the characteristics of objects and the composition of the photo. We hope that these features would capture the key factors that make a specific similar object important in a particular image.

Position. We consider the position of an object in the image as its feature. As shown in Fig. 9 (left), we take the masks of the top three important objects (pixel values are 1

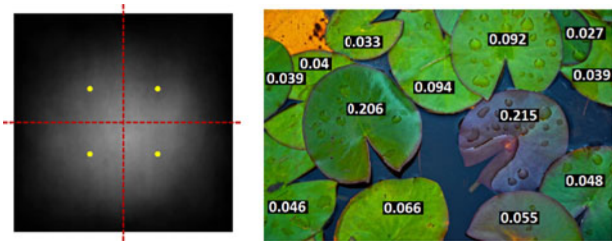


Fig. 9. Density of important objects. We can notice that the distribution is high in the central third of the image by looking at the mean number of the important objects per image covering a particular pixel (photos resized to 50×50). Furthermore, the distribution is both left-right and top-bottom symmetric. The image on the right is a typical example. The important objects are symmetrically distributed in the scene.

if they belong to the object; otherwise, 0) for all the images; the important objects are averaged, and an object density map is created [33]. We notice that the object density map is approximately symmetric in both horizontal and vertical directions, with most of the density residing in the center of the image. Therefore we measure distances from the object mask to the center point, horizontal midline, vertical midline, and four points that divide the image into thirds (marked in yellow in Fig. 9). We use these distances to encode the location of the object in the image. We also measure the sum of the distances from the object center point to the center points of the other objects. This feature represents the possibility that an object is near the center of a cluster.

Area. We consider the relative size of an object. We use the ratio between the area of the object and the area sum of all the objects, and the area rank within a local region. For the latter, we use the center of each object as the center point. We also use five times the length of the radius of its circumcircle as radius to generate a circular area as the local region. The area rank of this object is set to be the number of objects whose areas are smaller than this object. Then, the area rank of this object is also normalized by dividing the number of objects in the local region.

Color contrast. We consider the color contrast of an object to other objects and to the background as two features. For each pixel in an object, we first compute its histogram-based contrast to all pixels in other objects according to the mechanism in GC saliency [10]. Then, the average of the contrast of all pixels in the object is used as the color contrast of the object to other objects. Similarly, we use the average of pixel color contrast to the background pixels as the color contrast of the object to the background. The details of computing the histogram-based contrast can be found in [10].

Overlap. We consider the completeness of an object. We use the percentage of its area overlapped by other objects as a feature. Inspired by the object completion method in [1], we deform and map the complete reference template onto this object. We also regarded the uncovered pixels within the template as the occluded area. Then, we compute the ratio between the number of occluded pixels and the number of pixels in the deformed template as our overlap feature. During the data collection process, we also allow a user to indicate manually if an object is complete.

Blur. We consider the defocus blur of an object. We use the method in [34] to generate a defocus map, and the mean of the blur across the object is used as the feature.

Different from the method in [19], which devises many features for Lasso, the method we used has considerably fewer features, which we selected because the functional redundancy of the features may decrease the analyzability of the result.

5.2 Predicting by Ranking

We now describe our ranking-based importance prediction method. We adopt the *listwise* learning-to-rank methodology [32]. m images supposedly exist in the training set. Each image $\{\mathbf{I}^{(i)}\}_{i=1}^{m}$ contains n similar objects. Its features are extracted according to Section 5.1 and combined to a feature vector $\{\varphi_j^{(i)}\}_{j=1}^{n}$.

We learn a ranking function that can minimize the empirical loss $L(\mathbf{h})$, as follows:

$$L(\mathbf{h}) = \frac{1}{m}\sum_{i=1}^{m}\phi(\mathbf{h}(\varphi^{(i)}), y^{(i)}),$$

where $\phi(\mathbf{h}(\varphi), y)$ is the loss function. $y^{(i)}$ is the object index list of $\mathbf{I}^{(i)}$, and $y_j^{(i)}$ is the index of the object at position j (measured in Section 4 as the ground truth). The ranking function $\mathbf{h}$ assigns a score to each similar object (by employing a scoring function g), sorts the similar objects in a descending order of the scores, and finally creates the index $\hat{y}$ of the ranked object list. That is, $\mathbf{h}(\varphi^{(i)})$ is decomposable with respect to similar objects and is defined as

$$\mathbf{h}(\varphi^{(i)}) = sort(g(\varphi_1^{(i)}), \ldots, g(\varphi_n^{(i)})), \tag{3}$$

where $sort(\cdot)$ denotes the sorting function, and $g(\cdot)$ is the scoring function. In this paper, we define $g(\cdot)$ as $g(\cdot) = \exp(w^T\varphi)$, where w is the weight vector to be learned.

We experiment with two different loss functions, including *ListMLE* for likelihood loss and *ListNet* for cross entropy loss. Specifically, ListMLE maximizes the sum of the likelihood function with respect to all the training object orders, and its loss function is defined as:

$$L(g;\varphi,y) = -\frac{1}{m}\sum_{i=1}^{m}\log P(y^{(i)}|\varphi^{(i)};g),$$

$$P(y^{(i)}|\varphi^{(i)};g) = \prod_{j=1}^{n}\frac{g(\varphi^{(i)}_{y_j^{(i)}})}{\sum_{l=j}^{n} g(\varphi^{(i)}_{y_l^{(i)}})}.$$

The loss function of cross entropy used in ListNet is defined as

$$L(g;\varphi,\psi) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{\forall\pi\in\mathcal{Y}_k} P(\pi^{(i)}|\varphi^{(i)};\psi^{(i)})\log P(\pi^{(i)}|\varphi^{(i)};g),$$

$$P(\pi^{(i)}|\varphi^{(i)};\psi) = \prod_{j=1}^{n}\frac{\psi^{(i)}_{\pi_j^{(i)}}}{\sum_{l=j}^{n}\psi^{(i)}_{\pi_l^{(i)}}}, P(\pi^{(i)}|\varphi^{(i)};g) = \prod_{j=1}^{n}\frac{g(\varphi^{(i)}_{\pi_j^{(i)}})}{\sum_{l=j}^{n} g(\varphi^{(i)}_{\pi_l^{(i)}})},$$

where π is one permutation and $\mathcal{Y}_k$ contains all the permutations of top k ($k = 3$ in all our experiments). ψ is the object importance values for an image (measured in Section 4 as the ground truth).

For both ListMLE and ListNet, we use a one-layer linear neural network model parameterized by w without the bias b, and the model is trained using stochastic gradient descent for optimization. We initialize the weights in the linear neural network from a zero-mean Gaussian distribution with standard deviation 0.1. The learning rate is initialized at 0.001 and kept unchanged in all of the training cycles. We train the network for 1,000 cycles. The derivation we use for ListMLE is

$$\Delta w = -\sum_{t=1}^{n}\left[\varphi^{(i)}_{\pi_t^{(i)}} - \frac{\sum_{l=t}^{n}\exp\left(w^T\cdot\varphi^{(i)}_{\pi_l^{(i)}}\right)\cdot\varphi^{(i)}_{\pi_l^{(i)}}}{\sum_{l=t}^{n}\exp\left(w^T\cdot\varphi^{(i)}_{\pi_l^{(i)}}\right)}\right],$$

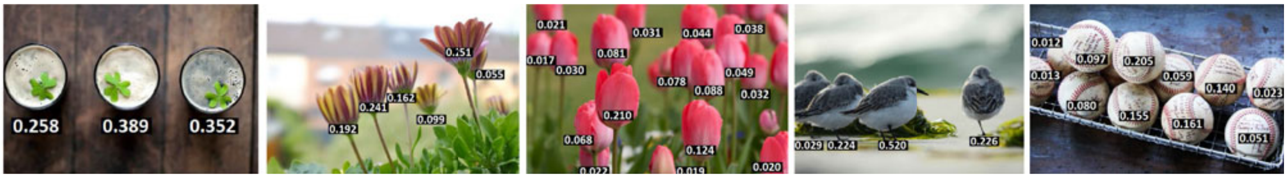


Fig. 10. Prediction of object importance. The numbers are object importance values which are normalized to $(0, 1)$.

and the derivation for ListNet is

$$\Delta w = -\sum_{\forall \pi \in \mathcal{Y}_k} P(\pi^{(i)} | \varphi^{(i)}; \psi) \cdot \sum_{t=1}^{k} \left[\varphi^{(i)}_{\pi^{(i)}_t} - \frac{\sum_{l=t}^{n} \exp\left(w^T \cdot \varphi^{(i)}_{\pi^{(i)}_l} \right) \cdot \varphi^{(i)}_{\pi^{(i)}_l}}{\sum_{l=t}^{n} \exp\left(w^T \cdot \varphi^{(i)}_{\pi^{(i)}_l} \right)} \right].$$

Ranking prediction. Once a listwise model has been trained, it can be used for ranking object importance on new images. In particular, given an image $\{\mathbf{I}^{(i)}\}_{i=1}^{m}$ and a set of n similar objects extracted from $\mathbf{I}^{(i)}$, $\{\varphi_j^{(i)}\}_{j=1}^{n}$ is the corresponding vector set. We compute a score $g(\varphi_j^{(i)})$ for each object and obtain the ranks by sorting their scores with Equation (3).

Predicting by pairwise ranking. Pairwise methodology is not focused on accurately predicting the rank list of all objects. It usually simplifies ranking to classification on object pairs to determine which object in a pair is preferred. If all the object pairs are correctly classified, all these objects are correctly ranked. We test Ranking SVM [35] to solve the object importance prediction problem, which applies the SVM technology to perform pairwise classification. Given a similar object feature pair $(\varphi_u^{(i)}, \varphi_v^{(i)})$, and the corresponding ground truth label $y_{u,v}^{(i)}$, if $s_u^{(i)} > s_v^{(i)}, y_{u,v}^{(i)} = 1$; otherwise $y_{u,v}^{(i)} = 0$. The mathematical formulation of Ranking SVM is:

$$\min \frac{1}{2} ||w||^2 + \lambda \sum_{i=1}^{m} \sum_{u,v: y_{u,v}^{(i)}=1} \xi_{u,v}^{i}$$
$$s.t. \;\; w^T(\varphi_u^{(i)} - \varphi_v^{(i)}) \geq 1 - \xi_{u,v}^{(i)}, \text{ if } y_{u,v}^{(i)} = 1; \; \xi_{u,v}^{(i)} \geq 0, i = 1, \ldots, n.$$

5.3 Predicting by Random Forest

We use random forest [36] to regress the object importance value with the extracted features. Denote B as the number of trees in the random forest, and M is the number of features extracted from the image. For each tree, we sample two-thirds of training objects as our bootstrapped training data set, and use $\sqrt{M}$ features to fit the training set. Finally, we combine the prediction of B trees as our predicted importance:

$$f_{importance}(o_j) = \frac{1}{B} \sum_{b=1}^{B} f^b(o_j),$$

where $f^b(\cdot)$ represents the regression tree. In our experiment, we use 5,922 objects in 808 images as the training set, where $B = 500$. The overall ranking for each object o_j are then obtained by sorting their importance values.

We show some object importance prediction results in Fig. 10, and the feature coefficients for predicted importance in Fig. 11. Since the optimization strategies of ListNet and ListMLE are similar, we calculate the correlation of the two list of feature coefficients and find that they are quantitatively consistent (the details of Kendall's τ metric are described in Section 6.1.1).

6 Evaluations

We experiment with our importance prediction methods (i.e., ListNet, ListMLE, Ranking SVM and random forest) on the SOID database (described in Section 3). Each image is associated with a set of interactively segmented similar object masks, the importance scores of each similar object, and a "ground truth" permutation of the objects sorted by their importance scores (generated in Section 4). We use fivefold cross-validation for each of the experiments in this section, and the average results are reported.

6.1 Ranking Accuracy

We examine the quality of our importance prediction results by assessing how well they correlate with the ground-truth ranking. To the best of our knowledge, our work is the first to approach the problem of ranking visual importance of similar objects in a scene. We compare our method with the following two baseline ranking methods.

Linear regression. In this method, we use linear regression (pointwise strategy) [37] to estimate the feature coefficients.

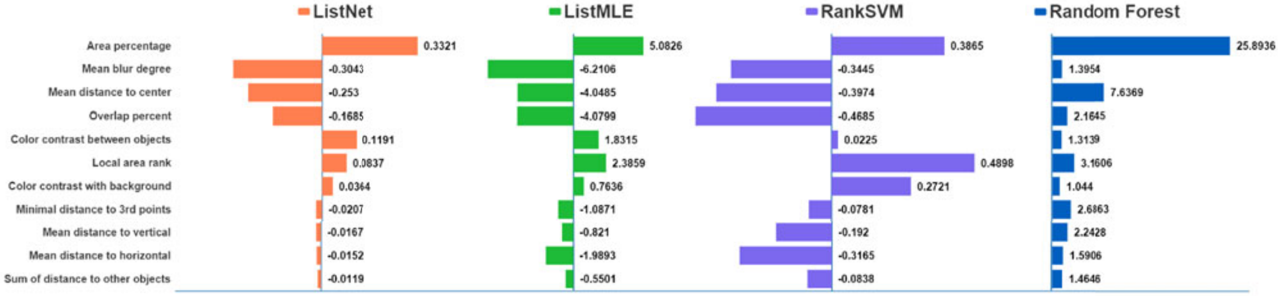


Fig. 11. Coefficients for importance prediction. Kendall's τ correlation coefficient of the feature lists of ListNet and ListMLE is 0.854.

TABLE 1
Rank Correlation and Ranking Quality

	Kendall's τ	Weighted Kendall's τ	Spearman's ρ	nDCG$_1$	nDCG$_3$	nDCG$_5$	nDCG full list
ListNet	0.7391	0.8911	0.8337	0.9417	0.9622	0.9647	0.9805
ListMLE	0.7432	0.8835	0.8327	0.9369	0.9652	0.9671	0.9817
RankSVM	0.7321	0.8796	0.8250	0.9365	0.9604	0.9634	0.9796
Random Forest	0.7272	0.8765	0.8210	0.9319	0.9541	0.9580	0.9774
Linear Regression	0.6578	0.8119	0.7529	0.8874	0.9212	0.9253	0.9453
Mean Distance to Center	0.4577	0.6000	0.5495	0.7937	0.8458	0.8550	0.9075
BMS	0.3065	0.4116	0.3826	0.6968	0.8072	0.8283	0.9055
eDN	0.4247	0.5742	0.5230	0.7703	0.8452	0.8589	0.9233
GC	0.4583	0.6082	0.5563	0.7993	0.8623	0.8751	0.9325
GMR	0.3862	0.5365	0.4817	0.7522	0.8427	0.8617	0.9235
RBD	0.3821	0.5222	0.4767	0.7579	0.8474	0.8624	0.9237

We include all the information of object segmentation and object features in the training process.

Mean distance to center. In this method, we use the single feature "mean distance to center" to rank the visual importance of each similar object, given that people usually focus their attention on the central part of a scene. This feature is also one of the dominant attributes particularly when the object areas in the scene are comparable.

Moreover, we consider the general saliency strategies that are frequently used to detect regions that are informative in an input image. In particular, we compare our method against five recent state-of-the-art saliency detection methods: BMS [11], GMR [12], eDN [23], RBD [13], GC [10]. For each object in an input image, we use the mean of the saliency value within the object as the importance score to rank the object importance by saliency. Then, the permutation is obtained by sorting the objects according to the importance scores. We then compare their ranking results with those generated by our importance prediction methods.

6.1.1 Correlation with Ground-Truth Ranking

We compute their rank correlation to evaluate how well our ranking results agree with the ground truth ranking. In particular, we experiment with the three rank correlation metrics.

Kendall's τ coefficient. Kendall's rank correlation coefficient [38], commonly known as Kendall's τ coefficient, is a statistic used to measure the association between two measured quantities. Given a set of elements $O = \{o_i, i = 1, \ldots, n\}$ and two ranking functions r_1 and r_2, Kendall's τ coefficient is computed as

$$\tau(r_1, r_2) = \frac{\sum_{ij} \delta[r_1(i,j) = r_2(i,j)] - \sum_{ij} \delta[r_1(i,j) \neq r_2(i,j)]}{0.5 \cdot n \cdot (n-1)},$$

where δ denotes the indicator function. $r(i,j)$ outputs 1 if the ranking function r gives o_i a higher rank than o_j, otherwise, the output is 0. This metric penalizes a pair of elements if their relative orders given by the two ranking functions disagree.

Weighted Kendall's τ coefficient. Inspired by [39], we also experiment with the weighted Kendall's τ rank correlation metric:

$$\tau(r_1, r_2) = \frac{\sum_{ij} \alpha_{ij} \delta[r_1(i,j) = r_2(i,j)] - \sum_{ij} \alpha_{ij} \delta[r_1(i,j) \neq r_2(i,j)]}{\sum_{ij} \alpha_{ij}},$$

where the weight α_{ij} is defined as

$$\alpha_{ij} = \max\{s(i), s(j)\} \cdot |s(i) - s(j)|,$$

where s scores are the object importance measured in Section 4. Intuitively, the weighted Kendall's τ rank correlation reduces the penalty on the discordant pairs of objects when their importance scores are close to each other. At the same time, it emphasizes the penalty on the pairs that contain important objects.

Spearman's ρ coefficient. In statistics, Spearman's rank correlation coefficient, or Spearman's ρ coefficient, is a nonparametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described with a monotonic function. For our problem, Spearman's ρ is computed as

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n \cdot (n^2 - 1)},$$

where d_i is the difference between the rank of the ith object in the predicted result and in the ground truth.

Table 1 (left side) shows the average rank correlation of the test data. The results show that the object importance ranking results from our methods have significantly higher correlations with ground truth ranking than those from the baseline methods and the saliency detection methods. Listwise ranking methods (ListNet and ListMLE) achieve the best performance in all metrics.

6.1.2 Normalized Discounted Cumulative Gain (nDCG)

Discounted cumulative gain (DCG) is a measure of ranking quality, which uses a graded relevance scale of elements in a ranking result set to measure the usefulness, or gain, of an element based on its position in the result list. The gain is accumulated from the top of the result list to the bottom with the gain of each result discounted at low ranks. The DCG accumulated at a particular rank position p is defined as

$$\text{DCG}_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i)},$$

where rel_i is the graded relevance of the result at position i, which in our problem is the importance score at position i in the ground-truth permutation. We then use nDCG to measure the ranking quality of our methods

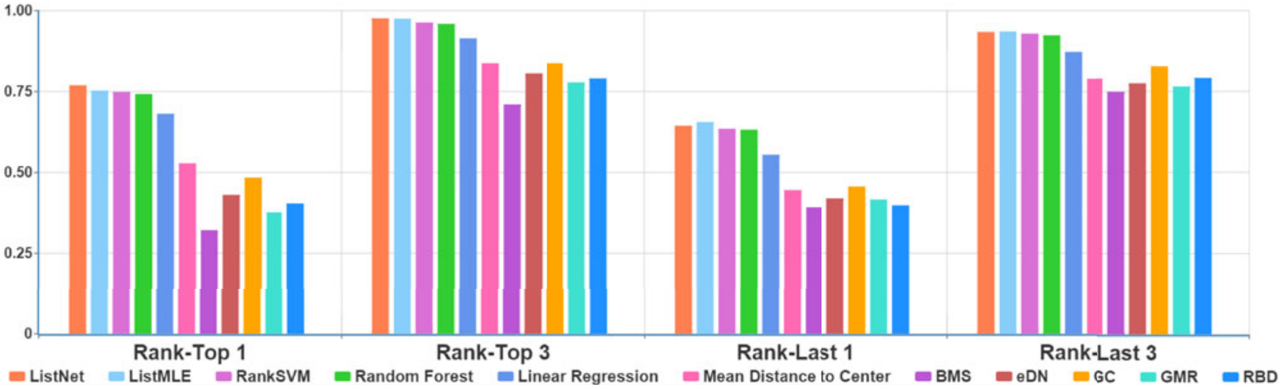


Fig. 12. Rank-n accuracy. The ranking predicted by our model is significantly better than the baseline ranking of saliency detection in selecting the most important and least important similar objects.

$$n\text{DCG}_p = \frac{\text{DCG}_p}{i\text{DCG}_p},$$

where iDCG is the ideal DCG, which is obtained by sorting the objects as a result permutation by importance, and then producing the maximum possible DCG until position p. In a perfect ranking algorithm, DCG_p is the same as $i\text{DCG}_p$ producing an nDCG of 1.0.

Table 1 (right side) shows the ranking quality on the test data. The results show that the object importance ranking from our method has significantly higher quality than those from the baseline methods and saliency detection methods, whereas listwise methods achieve the best performance in all metrics.

6.1.3 Importance Prediction Performance

In this experiment, we examine the performance of our importance prediction results to retrieve the most important object and to judge the least important object for a given image. For evaluation, we measure the rank-n accuracy. Given the ranking results for all objects in testing images, the rank-n accuracy is computed as the percentage of the test data for which the actual most important object is ranked within the top n positions, and the actual least important object is ranked within the last n positions. We also consider the least important object because accurately judging the unimportant objects in some applications (e.g., image retargeting) is as useful as retrieving the important ones.

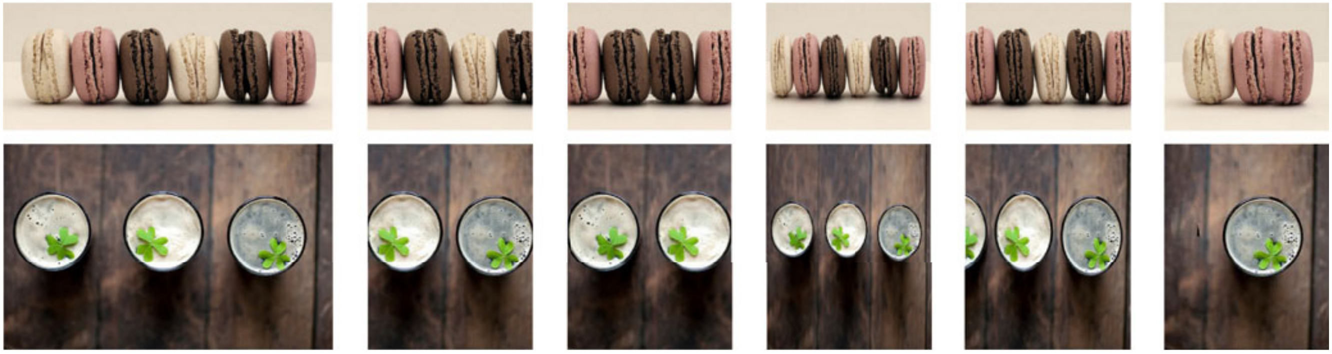
Fig. 12 shows rank-top 1, rank-top 3, rank-last 1, and rank-last 3 accuracy from our methods, as well as those from the baseline and saliency detection methods. The figure shows that the ranking results from our methods can provide significantly better object importance prediction accuracy than those from the baseline and saliency detection methods.

7 Applications

We have implemented our algorithm in C++ on a computer with Intel Core i7 CPU at 3.9 GHz, 8 GB RAM, and Geforce GTX 670. Our interactive similar object extraction tool typically takes 2-10 seconds to process a 900×900 image. After extracting all the objects from the given image, our importance prediction process (including object feature extraction) usually takes 5-8 seconds depending on the number of objects and the size of the given image. This information makes various high-level image editing applications possible, which we describe next. In some examples, we compare our results of using the object importance information predicted by listwise ranking (ListNet) with the results of using the object importance information derived from linear regression (baseline method) and saliency maps. Those saliency maps are randomly picked from the results of the five state-of-the-art saliency detection methods, which are used for comparison in Section 6. The numbers on the objects represent the importance order or importance value.

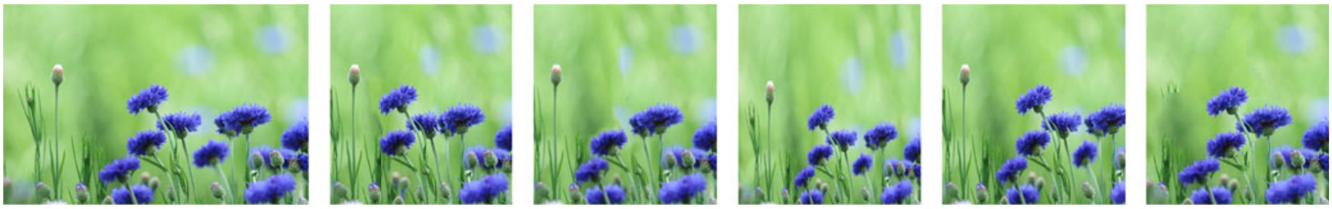
Image retargeting. Content-aware image retargeting has become a useful tool because of the diversity of display devices and versatility of image sources. Image retargeting can be more effectively achieved with an understanding of image semantics. Recently, Dong et al. [26] presented a summarization-based image retargeting algorithm that can manipulate an image at object-level. During the operation, "unimportant" objects instead of pixels or patches, are entirely removed, thereby preserving the shape of the remaining objects. As shown in Figs. 13, 14, and 15, our object importance information can be directly integrated into their object carving process and help to keep the important objects in the resized images. The results generated by the retargeting program through our object importance information are better than those generated through linear regression or saliency-based object importance information. The main reason is that our object importance prediction method can accurately recognize the important and unimportant objects, whereas the baseline methods cannot. In Figs. 13c, 14c, 15c, and 15f, the important objects in the retargeted images are lost because the baseline methods wrongly mark the important objects as "unimportant" in the original image. For image retargeting, we only need to know the order of the objects.

Image compression. Detecting the salient regions of an image can facilitate the application of image compression. A popular approach to reduce the size of a compressed image involves selecting a small number of interesting regions in it and encoding them in priority [43]. We adopt the compression process of JPEG image format to compress images adaptively with multiple similar objects. Three experiments are also designed to validate the effectiveness of using our object importance information. For an input image, we compare our result obtained through adaptive compression (denoted as AC method, in which the



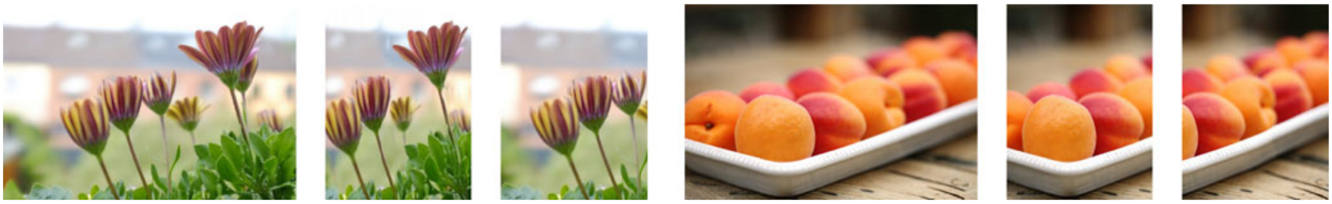
(a) Original image (b) Our result (c) By linear regression (d) AAD [40] (e) F-MultiOp [41] (f) Shit-Map [42]

Fig. 13. Image retargeting application. The remaining objects in our results are more visually and semantically vivid than the ones in (c).



(a) Original image (b) Our result (c) By saliency (d) AAD [40] (e) F-MultiOp [41] (f) Shift-Map [42]

Fig. 14. Image retargeting application. An important flower is wrongly recognized as unimportant by saliency and then removed in (c).

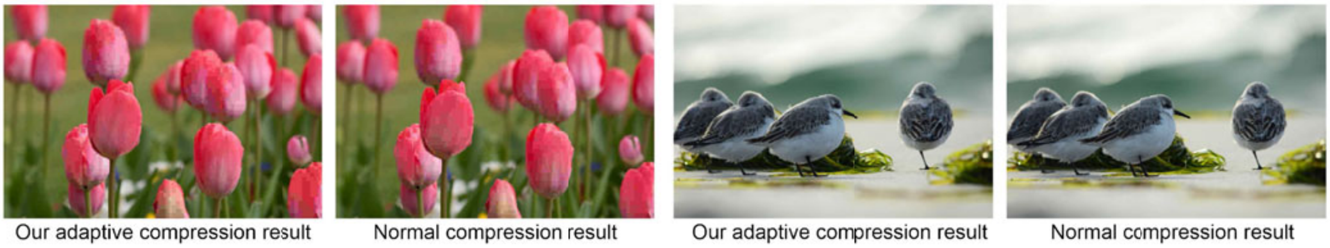


(a) Original image (b) Our result (c) By saliency (d) Original image (e) Our result (f) By saliency

Fig. 15. Image retargeting application. Important objects are wrongly recognized as unimportant by saliency and then removed in (c).

important objects are compressed to a relatively high quality and the unimportant objects are compressed to a relatively low quality) with the result obtained through normal compression (denoted as NC method, in which the similar objects are compressed to the same quality) or with the result obtained through AC method generated through saliency-based object importance (denoted as ACS method). In Experiment 1, we compress the original image to the same size through AC and NC. In Experiment 2, we use AC to compress each image to 80 percent size of the result through NC. In Experiment 3, we compress the original image to the same size through AC and ACS. In each experiment, we conduct the same user study with an online evaluation system. In each page, we show two compressed images and ask the participant to choose the image with good visual quality or to click “similar” if he/she is unsure. In both experiments, we show 73 pairs of images to the participants and record the total number of times that each item has been chosen. A total of 44 people (23 men and 21 women, aged 20 to 45 years) from different backgrounds participate in the user study. In Experiment 1, the participants choose our results, NC results, and “similar” for 63.5, 12.6, and 23.9 percent times, respectively. Thus, we can conclude that the image compression results generated with our method are better than those generated with NC when the file sizes are the same. In Experiment 2, the participants choose our results, NC results, and “similar” for 17, 21.3, and 61.7 percent times, respectively. Thus, we can conclude that our method can generate image compression results encompassing quality similar to that of the results generated with NC when file sizes of our results are smaller than the files sizes of NC results. In Experiment 3, the participants choose our results, ACS results, and “similar” for 70.2, 9.1, and 20.7 percent times. Thus, we can conclude that the image compression results generated with our object importance information are better than those generated with saliency-based importance information. We show the comparison of image compression results in Figs. 16 and 17.

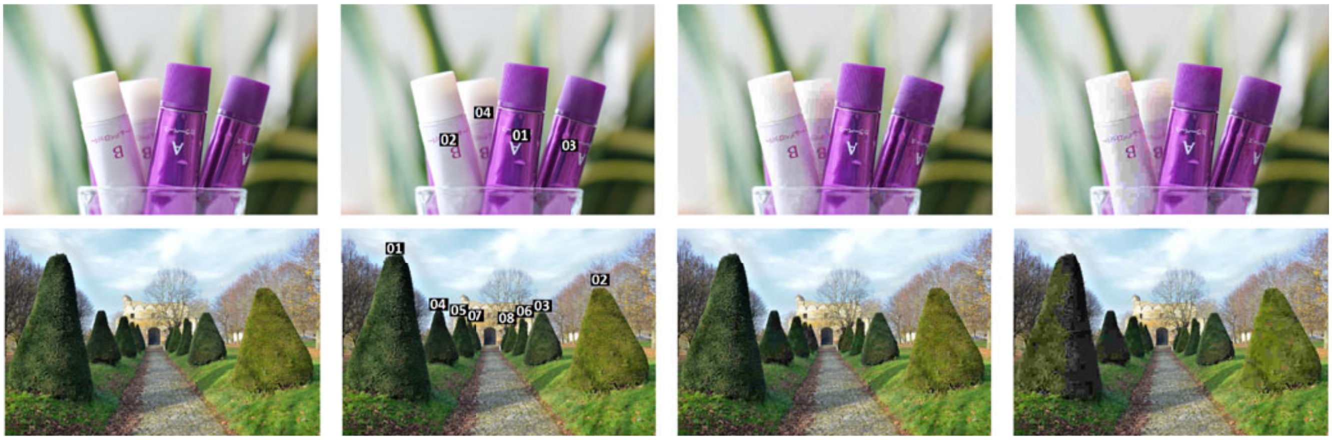
Image re-attentionizing. A psychological experiment showed that redundant regions of an image are often skipped, and instead, the viewers focus on the rare regions in the image [44]. An image re-attentionizing tool [45] can endow the target region in an image with the ability to attract human visual attention. The objects that are visually attractive in an image can be determined by using our importance prediction model. By contrast, the factors effectively affect the visual importance of an object are also determined. Thus, we can easily increase or decrease the


Our adaptive compression result
Normal compression result
Our adaptive compression result
Normal compression result

(a) The images are compressed to the same size.

(b) The images are compressed to different sizes (ours are smaller).

Fig. 16. In adaptive image compression, important objects are encoded first. Compared with the results of the normal method, our results have (a) the same file sizes but better visual qualities and (b) smaller file sizes than the original but similar visual quality.



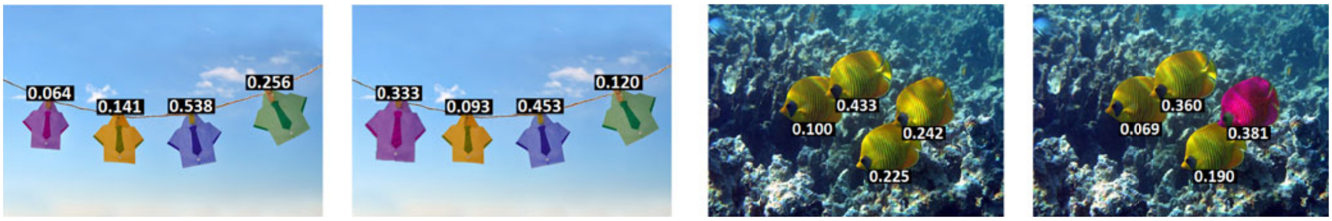
(a) Original image

(b) Object importance

(c) Our compression results

(d) Compression according to saliency

Fig. 17. Adaptive image compression. Our results have better visual quality than those obtained with saliency-based importance information.



(a) Original image

(b) Re-attentionizing by scaling

(c) Original image

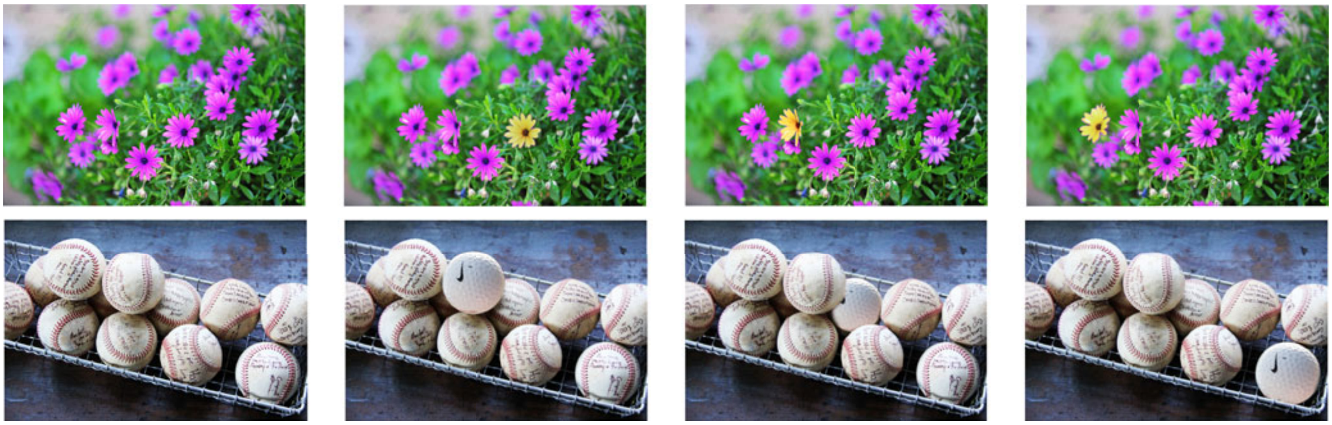
(d) Re-attentionizing by re-colourization

Fig. 18. Image re-attentionizing. We change the visual importance of objects through different editing operations, according to different factors.

visual attractiveness through simple editing operations. Figs. 2d and 18 show some results of image re-attentionizing. Our importance prediction subscribes a valuable quantitative reference for users during the re-attentionizing operations. Fig. 19 shows that an image re-attentionizing operation is employed to generate the broken pattern effect (discontinuation or interruption in the flow of the similar pattern), which commonly used in photography to break the monotony of repetition and to make the photo appear catchy. Our results in Fig. 19b are visually more pleasing than the others because we re-attentionize the most important object in each example. The baseline linear regression method fails to predict accurately the most important object in Fig. 19a.

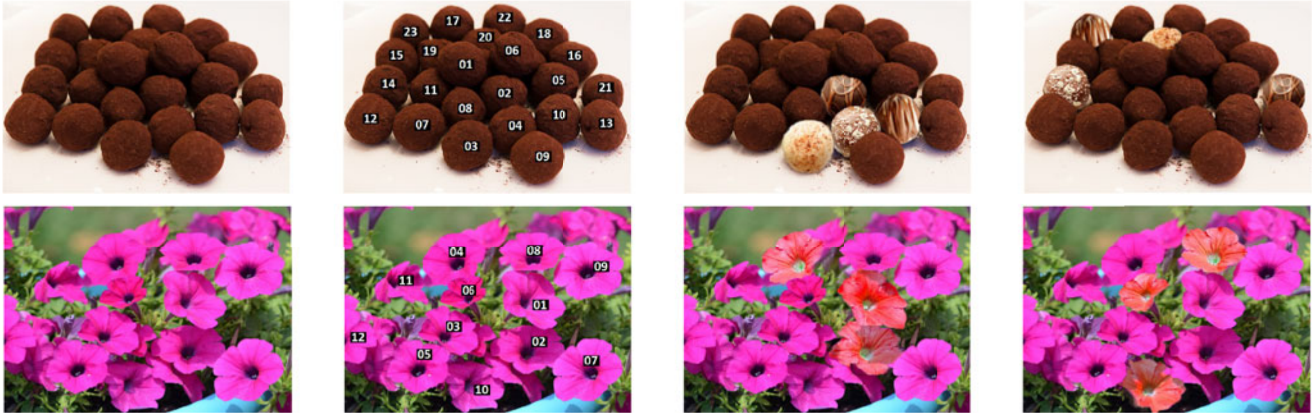
Image admixture. For image admixture [25], considering the visual importance of the objects is a good scheme to choose the suitable elements for the replacement operation when different object groups are combined. After predicting the object importance (see Fig. 20b), we generate object admixture results by separately replacing some important objects (see Fig. 20c) as well as some unimportant objects (see Fig. 20d) through the element replacement method in [25]. The exotic objects visually dominate the scene when important objects are replaced. However, compositions are still balanced when unimportant objects are replaced. In conclusion, using object importance in object selection makes the results of the image admixture controllable and visually vivid (the objects are randomly replaced in [25]).

Change blindness images. Change blindness is a psychological phenomenon in which the excessively large changes made to an image may not be noticed by observers. Change blindness is related to visual attention, and changes in locations with low saliency are unlikely to be detected [46]. As shown in Fig. 21, we can utilize the information of object importance to create the change blindness images. A



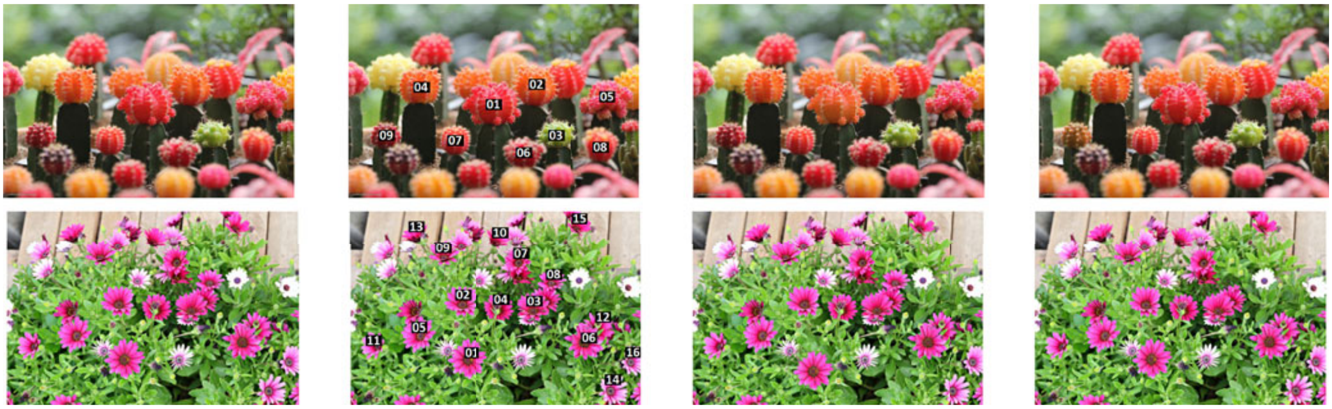
(a) Original image (b) Our broken pattern result (c) Broken pattern by linear regression (d) Re-attentionizing a random object

Fig. 19. Generating a broken pattern through image re-attentionizing (achieved by re-colorization and object replacement). In (b) and (c), we separately re-attentionize the most important object predicted through our method and linear regression. In (d), we re-attentionize a random object.



(a) Original image (b) Object importance (c) Replacing importance objects (d) Replacing unimportant objects

Fig. 20. We can choose suitable elements for replacement in image admixture by utilizing the object importance information.



(a) Original image (b) Object importance (c) Changing an importance object (d) Changing an unimportant object

Fig. 21. Change blindness images. We can perform tasks in different degrees of difficulty by changing objects with different importance values.

difficult image can be created if we change a relatively unimportant object. We also conduct a user study to test the effect of changing objects with different importance values, by recording the cost of time that users recognize as the difference between the original and new images. According to the statistics, the average recognition times in Fig. 21c are 3 seconds (cactus) and 7 seconds (flowers). The average recognition times in Fig. 21d are 15 seconds (cactus) and 24 seconds (flowers).

8 Conclusion and Future Work

In this study, we investigate the problem of measuring the visual importance of similar objects in a scene, and a method is proposed to predict this problem once the objects are segmented from the background. Our prediction algorithm works without object identity, i.e., we can always determine important identities without knowing what the objects are.

We asked a large number of participants to understand how humans perceive similarity by sorting similar objects in photographs that exhibit different scenes. For each of the 808 images, we collect 71 independent sorted lists. This dataset allows us to measure the object importance with a ranking model. We adopt two listwise ranking methods, one pairwise method, and a regression method based on random forests to separately evaluate the visual importance from object features computed from the image.

We determined a ranking of how informative different object-related image features are in predicting importance. The size, position, and overlap percentage of similar objects are useful for natural images. The degree of sharpness is also important if some regions of the image are blurred. However, as a machine learning system, predicting accuracy will be affected by inaccurate features. For example, it is not easy for our blur feature to distinguish blur effects caused by defocus from blur textures, so an important object with blur texture may be predicted as unimportant in a scene containing similar objects with different textures (all objects are clear, some objects contain sharp textures and others contain blur textures). Moreover, color contrast feature may be ineffective for an object with only-user-noticeable color difference in a small area (e.g., a cake has a cherry but the others do not have). Fortunately, there is internal relevance between our features, such as area and overlap (an incomplete object often has a small area), and blur and mean distance to center (clear objects are often placed near the center of the image), so one feature will be a supplement if its relevant feature is not very accurate. In our algorithm, the ranking of object visual importance results from interactions among multiple features, so to get a good result we usually do not require every feature is accurately computed.

Our experiments show that high-importance objects with state-of-the-art automatic saliency detection schemes cannot be easily isolated. Our system relies on an interactive similar object detection system. We believe that at present a fully automatic method that can robustly detect similar objects from a single natural image does not exist. Developing a good similar object detection method is out of the scope of this study. However, progress in this area can clearly benefit computer vision and computer graphic applications. Scene classification and understanding of the image may also improve importance prediction. We plan to collect numerous images for SOID and to train different ranking models for different classes of images based on their specific characteristics. The accuracy of the prediction may be increased based on this submodeling strategy.

In our experiments, the listwise-based ranking models achieve the best performance on object importance prediction. At present the training processes of these two models are both developed based on single-layer neural networks. However, the performance of the models is expected to be improved by deep learning framework through multi-layer neural networks. Moreover, thanks to the rapid development convolutional neural network technologies, like AlexNet [47] and R-CNN [48], we can detection objects and extract features directly from a raw image input. In the future, we plan to incorporate these technologies into our framework and learn an end-to-end model to rank similar objects directly.

We also apply our method to many applications, such as image retargeting, adaptive image compression, image reattentionizing, and finding suitable elements for replacement and blindness images change. Our method can also be used to evaluate image retargeting methods based on the proposed method. Many application results can be found in our supplementary file, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2016.2515614. Overall, these results show the good potential of the proposed method.

Acknowledgments

The authors thank the anonymous reviewers for their constructive comments. This work was supported by National Natural Science Foundation of China under nos. 61271430, 61372184, 61331018, and 61471359, by Supporting Program for Sci & Tech Research of China under No. 2015BAH53F00, and by CASIA-Tencent BestImage joint research project. The work of Tong-Yee Lee was supported by the Ministry of Science and Technology with contract nos. MOST-103-2221-E-006-106-MY3 and MOST-104-2221-E-006-044-MY3, Taiwan. The work of Siwei Lyu was supported by National Science Foundation under nos. IIS-0953373 and CCF-1319800. Y. Kong and W. Dong contributed equally to this work.

References

[1] M.-M. Cheng, F.-L. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, “RepFinder: Finding approximately repeated scene elements for image editing,” *ACM Trans. Graph.*, vol. 29, no. 4, pp. 83:1–83:8, 2010.

[2] Y. Kong, W. Dong, X. Mei, X. Zhang, and J.-C. Paul, “SimLocator: Robust locator of similar objects in images,” *The Visual Comput.*, vol. 29, no. 9, pp. 861–870, 2013.

[3] Y. Cai and G. Baciu, “Detecting, grouping, and structure inference for invariant repetitive patterns in images,” *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2343–2355, Jun. 2013.

[4] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, “Scalable object detection using deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2155–2162.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 248–255.

[6] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[7] H. Schweitzer, R. Deng, and R. F. Anderson, “A dual-bound algorithm for very fast and exact template matching,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 459–470, Mar. 2011.

[8] H. Huang, L. Zhang, and H.-C. Zhang, “RepSnapping: Efficient image cutout for repeated scene elements,” *Comput. Graph. Forum*, vol. 30, no. 7, pp. 2059–2066, 2011.

[9] P. Xu, H. Fu, O. K.-C. Au, and C.-L. Tai, “Lazy selection: A scribble-based tool for smart shape elements selection,” *ACM Trans. Graph.*, vol. 31, no. 6, pp. 142:1–142:9, Nov. 2012.

[10] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, “Global contrast based salient region detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[11] J. Zhang and S. Sclaroff, “Saliency detection: A Boolean map approach,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.

[12] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, “Saliency detection via graph-based manifold ranking,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3166–3173.

[13] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2814–2821.

[14] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 17–31, 2007.

[15] L. Elazary and L. Itti, "Interesting objects are visually salient," *J. Vis.*, vol. 8, no. 3, pp. 3:1–3:15, 2008.

[16] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1–3, pp. 157–173, 2008.

[17] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2004, pp. 319–326.

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[19] M. Spain and P. Perona, "Measuring and predicting object importance," *Int. J. Comput. Vis.*, vol. 91, no. 1, pp. 59–76, Jan. 2011.

[20] A. Berg, T. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi, "Understanding and predicting importance in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3562–3569.

[21] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 134–153, Nov. 2012.

[22] C. L. Zitnick and D. Parikh, "Bringing semantics into focus using visual abstraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3009–3016.

[23] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2798–2805.

[24] C. Ma, L.-Y. Wei, and X. Tong, "Discrete element textures," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 62:1–62:10, Jul. 2011.

[25] F.-L. Zhang, M.-M. Cheng, J. Jia, and S.-M. Hu, "ImageAdmixture: Putting together dissimilar objects from groups," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 11, pp. 1849–1857, Nov. 2012.

[26] W. Dong, N. Zhou, T.-Y. Lee, F. Wu, Y. Kong, and X. Zhang, "Summarization-based image resizing by intelligent object carving," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 1, pp. 111–124, Jan. 2014.

[27] J. Liu, J. Sun, and H.-Y. Shum, "Paint selection," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 69:1–69:7, 2009.

[28] M. G. Kendall and B. B. Smith, "The problem of m rankings," *The Ann. Mathematical Statist.*, vol. 10, no. 3, pp. 275–287, Sep. 1938.

[29] D. R. Hunter, "MM algorithms for generalized Bradley-Terry models," *Ann. Statist.*, vol. 32, no. 1, pp. 384–406, 2004.

[30] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, pp. 324–345, 1952.

[31] J. I. Marden, *Analyzing and Modeling Rank Data*. Boca Raton, FL, USA: CRC Press, 1996, vol. 64.

[32] Y. Lan, T.-Y. Liu, Z. Ma, and H. Li, "Generalization analysis of listwise learning-to-rank algorithms," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 577–584.

[33] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, pp. 1–26, Nov. 2008.

[34] S. Zhuo and T. Sim, "Defocus map estimation from a single image," *Pattern Recog.*, vol. 44, no. 9, pp. 1852–1858, Sep. 2011.

[35] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 133–142.

[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[37] D. A. Freedman, *Statistical Models: Theory and Practice*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[38] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, Jun. 1938.

[39] Y. Liu, J. Wang, S. Cho, A. Finkelstein, and S. Rusinkiewicz, "A no-reference metric for evaluating the quality of motion deblurring," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 175:1–175:12, Nov. 2013.

[40] D. Panozzo, O. Weber, and O. Sorkine, "Robust image retargeting via axis-aligned deformation," *Comput. Graph. Forum*, vol. 31, no. 2pt1, pp. 229–236, May 2012.

[41] W. Dong, G. Bao, X. Zhang, and J.-C. Paul, "Fast multi-operator image resizing and evaluation," *J. Comput. Sci. Technol.*, vol. 27, no. 1, pp. 121–134, 2012.

[42] Y. Pritch, E. Kav-Venaki, and S. Peleg, "Shift-map image editing," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 151–158.

[43] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

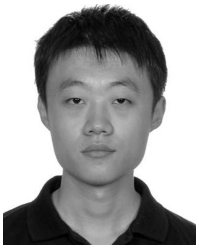
[44] F. Attneave, "Some informational aspects of visual perception," *Psychological Rev.*, vol. 61, no. 3, pp. 183–193, 1954.

[45] T. Nguyen, B. Ni, H. Liu, W. Xia, J. Luo, M. Kankanhalli, and S. Yan, "Image re-attentionizing," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1910–1919, Dec. 2013.

[46] L.-Q. Ma, K. Xu, T.-T. Wong, B.-Y. Jiang, and S.-M. Hu, "Change blindness images," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 11, pp. 1808–1819, Nov. 2013.

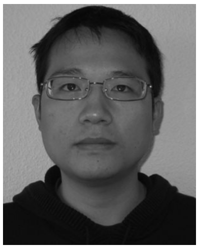
[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[48] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 580–587.

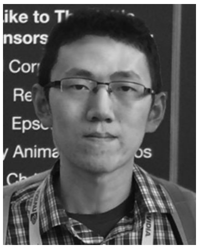


Yan Kong received the BSc degree in computer scienc from Beijing Jiaotong University, PR Chinae in 2011 and is currently working toward the PhD degree from the Sino-European Lab in computer science, automation and applied mathematics (LIAMA) and the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation, Chinese Academy of Sciences. His research interests include image synthesis and image analysis.

Weiming Dong received the BSc and MSc degrees in computer science in 2001 and 2004, respectively, both from Tsinghua University, PR China, and the PhD degree in computer science from the University of Lorraine, France, in 2007. He is an associate professor in the Sino-European Lab in computer science, automation, and applied mathematics (LIAMA) and the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation, Chinese Academy of Sciences. His research interests include image synthesis and image analysis. He is a member of ACM and the IEEE.



Xing Mei received the BSc degree in electronic engineering in 2003 from the University of Science and Technology of China (USTC), and the PhD degree in 2009 from CASIA. He is an assistant professor in the Sino-European Lab in computer science, automation and applied mathematics (LIAMA) and the National Laboratory of Pattern Recognition (NLPR) at the Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests include image processing, computer vision, and computer graphics. He is a member of the IEEE.



Chongyang Ma received the BS degree from fundamental science class (mathematics and physics) of Tsinghua University in 2007 and the PhD degree in computer science from the Institute for Advanced Study of Tsinghua University in 2012. He is a postdoctoral scholar in the Department of Computer Science at the University of Southern California. Before he joins USC, he spent one year working as a postdoctoral fellow in the Department of Computer Science at the University of British Columbia.

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

Siwei Lyu graduated from Peking University in the BS degree in information science and the MS degree in computer science in 1997 and 2000, respectively, and received the PhD degree in computer science in 2005 from Dartmouth College. He is an associate professor at the Computer Science Department of University at Albany, State University of New York. Prior to joining University at Albany, he was a post-doctoral research associate at the Howard Hughes Medical Institute and the Center for Neural Science of New York University from 2005 to 2008, and an assistant researcher at Microsoft Research Asia. His main research interests include digital image forensics, computer vision, and machine learning. He received the IEEE Signal Processing Society Best Paper Award in 2010, and the US National Science Foundation (NSF) CAREER Award in 2010. He is a member of the IEEE.

Feiyue Huang received his BSc and PhD degrees in computer science in 2001 and 2008, respectively, both from Tsinghua University, PR China. He is the director of the Social Network Platform Department, Tencent. His research interests include image understanding and face recognition.

Xiaopeng Zhang received the MSc degree in mathematics from Northwest University in 1987, and the PhD degree in computer science from Institute of Software, Chinese Academy of Sciences, in 1999. He is a professor in the Sino-European Lab in computer science, automation and applied mathematics (LIAMA) and the National Laboratory of Pattern Recognition (NLPR at the Institute of Automation, Chinese Academy of Sciences. His main research interests are computer graphics and pattern recognition. He received the National Scientific and Technological Progress Prize (Second Class) in 2004. He is a member of ACM and the IEEE.