A Dual-Domain Perceptual Framework for Generating the Visual Inconspicuous Counterparts

ZHUO SU, Sun Yat-sen University
KUN ZENG, Sun Yat-sen University
HANHUI LI, Sun Yat-sen University
XIAONAN LUO, Sun Yat-sen University

For a given image, it is a challenging task to generate its corresponding counterpart with visual inconspicuous modification. The complexity of this problem reasons from the high correlativity between the editing operations and vision perception. Essentially, a significant requirement that should be emphasized is how to make the object modifications hard to be found visually in the generative counterparts. In this paper, we propose a novel dual-domain perceptual framework to generate the visual inconspicuous counterparts, which applies the Perceptual Bidirectional Similarity Metric (PBSM) and Appearance Similarity Metric (ASM) to create the dual-domain perception error minimization model. The candidate targets are yielded by the well-known PatchMatch model with the strokes-based interactions and selective object library. By the dual-perceptual evaluation index, all the candidate targets are sorted to select out the best result. For demonstration, a series of objective and subjective measurements are used to evaluate the performance of our framework.

Additional Key Words and Phrases: Object Manipulation, Image Editing, Visual Perception, Bidirectional Similarity, Image Quality Assessment

ACM Reference Format:
Zhuo Su, Kun Zeng, Hanhui Li, Xiaonan Luo, 2017. A Dual-Domain Perceptual Framework for Generating the Visual Inconspicuous Counterparts. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, Article 1 (January 2017), 21 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Automatically generating the visual inconspicuous counterparts of an image is a challenging vision-perceptual task. And it provides a feasible way to rapidly generate the image pairs for “Find the Difference (also called Find X)” game. Find X game appeals lots of players because of its strong playability and low barriers. However, a funny game contains a significant and challenging task that is how to easily obtain a visual inconspicuous counterpart for a given image. Empirically, there exist two feasible ways to yield the difference counterparts (for convenience, we call “target image” instead in the following). One is to shoot a pair of photos with replacing some unimpressive objects in the same scene. The other is to modify an existed image by some interactive

This work is supported by the National Natural Science Foundation of China (61320106008, 61502541, 61370186), the Natural Science Foundation of Guangdong Province (2016A030310202, 2015A030313129), the Science and Technology Planning Project of Guangdong Province (2015B010106005), and the Fundamental Research Funds for the Central Universities (Sun Yat-sen University).
Author's addresses: Z. Su, K. Zeng, H. Li are with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China; X. Luo is with the National Engineering Research Center of Digital Life, Sun Yat-sen University, Guangzhou, China.

DOI: 0000001.0000001

(a) (b) (c)

Fig. 1: The representative image pairs are shown in the popular "Find the Difference" game. (a) A reference image. (b) A corresponding counterpart with some visual inconspicuous modifications. (c) The visual differences between the reference and the target are shown.

editing tools (e.g. Photoshop). As illustrated in Fig. 1, it is not easy to find the variances in the target images (Fig. 1(b)) without some visual difference hints (Fig. 1(c)).

Imaginably, the aforementioned ways are low efficient and high-required experimental skills for generating the visual inconspicuous counterparts. Essentially, the real difficulty of the task is how to effectively modify the given image, and further requires these modifications are unimpressive that hard to be found. For the former, with the recent developments of gradient domain processing, Markov random fields and graph theory, a series of advanced frameworks were developed to tackle the image editing tasks, such as the Poisson editing [Perez et al. 2003], shift-map editing [Pritch et al. 2009], PatchMatch [Barnes et al. 2009], Sketch2Photo [Chen et al. 2009], and RepFinder [Cheng et al. 2010]. These frameworks obviously alleviate the editing complexity and achieve a high-quality processing. Exploiting these frameworks, we can effectively tackle the editing tasks in the construction of visual inconspicuous counterparts. However, for how to make the variances hard to be found, it has not been addressed yet.

To tackle this problem, some vision perceptual factors should be considered to quantify the inconspicuous variances. In this paper, we propose a dual-domain perceptual framework, which integrates the Perceptual Bidirectional Similarity Metric (PBSM) and Appearance Similarity Metric (ASM) into a PatchMatch-like editing paradigm under the dual-domain perception error minimization measurement. The evaluation index in our model is regarded as the selection criteria for the best result. The perceptual bidirectional similarity metric takes advantage of the object saliency information with respect to the visual attention. It combines the completeness metric and the coherence metric in the bidirectional similarity discrimination, which evaluates the distinctions between the original image and the target image. Furthermore, the appearance similarity metric evaluates the variances on the object appearance from the aspect of sta-

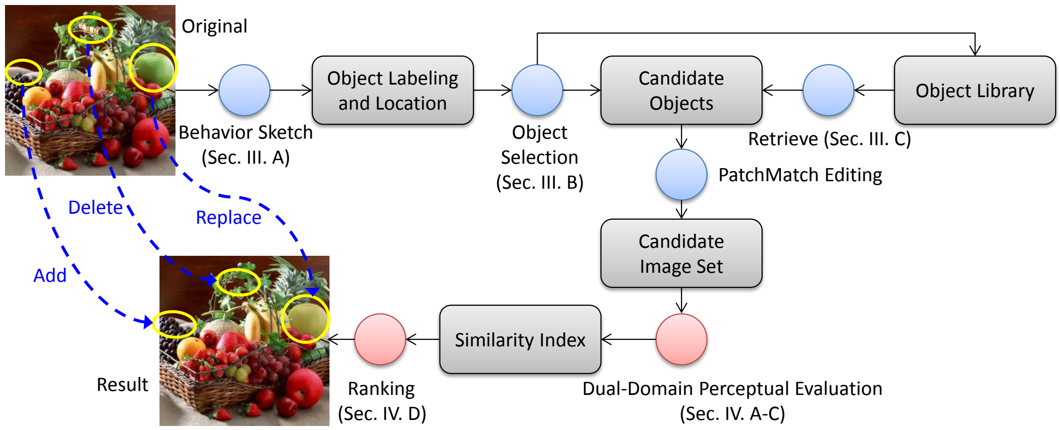


Fig. 2: The pipeline of our framework illustrates the generation of the visual inconspicuous counterparts. This pipeline could be roughly divided into two phases, one is the procedure of the data preparation, and the other is the dual-domain perceptual evaluation for the candidate targets. The basic behaviors for object manipulations are illustrated and the corresponding variances are labeled out in the original image and the result.

tistical colors distribution. By the logarithmic operation, a suitable evaluation index is obtained to sort the candidate target images, and finally yields the best result. The overview of our framework is illustrated in Fig. 2.

In summary, our main contributions are as follows.

— Construct a data generation framework for the visual inconspicuous counterparts, which applies the strokes-based image editing with visual perceptual measurement to yield the suitable targets.
— Propose the dual-domain perceptual evaluation framework, which includes the perceptual bidirectional similarity metric (PBSM) and appearance similarity metric (ASM) to measure the variances from the viewpoints of geometry and statistics, respectively.
— Demonstrate the effectiveness of our framework by the analysis of parameter setting, operations and performance. Especially, our perceptual index has a superior performance on guiding to avoid the visual attentions.

The rest of this paper is organized as follows. The related work is reviewed in Sec. II. The procedure of the data preparation is presented in Sec. III. The dual-domain perceptual editing framework with visual metrics is detailed explained in Sec. IV. The experimental results are shown in Sec. V, and some characteristics and summarizations are discussed in the last section.

2. RELATED WORK

In this section, we will summarize the related state-of-the-art theories and technologies with respect to photo manipulation, content-aware analysis, and perceptual quality assessment.

2.1. Photo manipulation

In the past decades, local photo manipulation technologies (including composition, recomposition, and completion) had a rapid development. In a wide range of novel manipulation technologies, Poisson equation based gradient domain reconstruction is proposed by [Perez et al. 2003] that provides a new inspiration for efficient and effective image editing. Following this way, [Jia et al. 2006], [Farbman et al. 2009], and [Tao et al. 2013] improved the Poisson model from the aspects of efficiency and effects, respectively. In the field of composition, [Eitz et al. 2009] first proposed a sketch-based image query and compositing system to create artificial image contents progressively. Soon after, [Chen et al. 2009] developed a Sketch2Photo system. In their system, users first sketch out the draft, then exploit a series of filters (for region, contour, and scene) to obtain the candidate objects and scenes from the Internet, and compose all the selections to yield the final result. Inspired by the Sketch2Photo system, [Goldberg et al. 2012] proposed a data-driven object manipulation framework in the recent.

From the aspect of the recomposition, [Cheng et al. 2010] sought to find the repeated elements in the scene and extracted them to rearrange by the proposed boundary band method. Similarly, [Zhang et al. 2012] adopted a semi-automatic element separation framework to perform individual structured object replacement from groups. Recently, [Bhattacharya et al. 2011] introduced a holistic approach to aesthetic enhancement. And [Zhang et al. 2013] further made a consideration with the aesthetic influence in the object recomposition. Meanwhile, in the field of content completion, [Criminisi et al. 2003] developed a new way called the example-based region filling. After that, [Sun et al. 2005] proposed graph structure propagation with belief propagation (BP) to complete the missing regions, which made a consideration to the geometry structure of the structural objects. Recently, [Huang et al. 2013] presented a structure-driven image completion with tele-registration.

Generally speaking, aforementioned local manipulation models concentrate on the performance in specific vision tasks. However, image editing is not an independent operation. It generally requires effective collaboration with kinds of related manipulations. The similarity-based metric model provides a feasible way to construct a generalized editing framework. [Simakov et al. 2008] creatively proposed the Bidirectional Similarity Distance (BSD) Model, which evaluated the variance between the source and the target from the aspect of both completeness and coherence. Inspired by the concept of bidirectional similarity, [Barnes et al. 2009] proposed a structural image editing framework with the randomized correspondence algorithm, called PatchMatch model. Subsequently, they proposed a generalized PatchMatch correspondence algorithm [Daniilidis et al. 2010] to improve their model, and applied it to a wide range of image editing tasks [Barnes et al. 2011]. In addition, other generalized editing frameworks are proposed recently, e.g. [Cho et al. 2008]'s Patch Transform, [Pritch et al. 2009]'s Shift-Map and [Hu et al. 2013]'s PatchNet .

2.2. Content-aware analysis

Content-aware analysis rises in the research of the non-homogeneous retargeting problem [Shamir and Sorkine 2009], due to that the retargeting quality mainly depends on the image saliency. The saliency detection started from [Itti et al. 1998]'s rapid scene analysis with regard to the visual attention. After that, [Harel et al. 2007] proposed a Graph-based Visual Saliency (GBVS) model from the bottom-up vision. Meanwhile, [Liu et al. 2007] regarded the salient object detection as a special image segmentation issue. They exploited Conditional Random Field (CRF) learning to complete the saliency detection tasks and published their dataset for saliency test. [Cheng et al. 2011] simultaneously evaluated global contrast differences and spatial coherence to construct a regional contrast based saliency extraction algorithm. Inspired by the combination of the low- and mid-level visual cues, [Xie et al. 2013] applied Laplacian sparse subspace clustering method and convex hull with the Bayesian inference to evaluate the object saliency. And [Wu et al. 2013] combined an improved principle component analysis (PCA) and the edge information by the L_0-norm smoothing to obtain the saliency. Limited to the length of the paper, more saliency models and their performance analysis could be found in [Borji et al. 2012]'s and [Cheng et al. 2015]'s benchmarks.

2.3. Perceptual quality assessment

Image quality assessment (IQA) is meaningful for analyzing the processing performance of a wide range of visual applications. However, there exists a certain limitation in some visual error metrics such as classical mean squared error (MSE) and peak signal-to-noise ratio (PSNR) [Wang et al. 2004]. Reason for the perception characteristic in human vision system, [Wang et al. 2004] proposed the well-known structural similarity (SSIM) index to evaluate the pixel-wise or patch-wise distortion/quality measures in image space. Then, [Zhou and Qiang 2011] proposed the concept of information content weighting to improve the perceptual image quality assessment. [Lin and Jay Kuo 2011] summarized the perceptual visual quality metrics (PVQMs) from kinds of aspects.

In recent years, perceptual-based blind (no-reference) image quality assessment [Moorthy and Bovik 2011] had a rapid development. Because of the important of visual perception, perceptual elements made a wide range of application in kinds of vision applications. For example, [Doersch et al. 2012] exploited the perceptual geo-information to automatically find visual elements in the street view. [Song et al. 2010] and [Lu et al. 2014] introduced the visual perception into color-to-gray problem, and emphasized the contrast preservation in the procedure of gray transform. [Zeng et al. 2009] and [Zhao and Zhu 2013] presented how to apply the visual perception in the painterly rendering.

3. DATA PREPARATION

To generate the corresponding visual inconspicuous counterparts (for convenience, we name "target images" instead in the following paragraphs) from the original one, we need to address the following three challenging problems. 1). What kinds of user's manipulations are performed? 2). Where are the candidate objects to be edited? 3). How to guarantee the harmony of the target images to make the modification unimpressively? In this section, we will address these problems one by one.

3.1. Editing operations group

Before the editing, first, we need to make the definition for the feasible operations. In our framework, we summarize four types of major image operations as follows.

Add, including the non-overlap adding, which inserts the new object into a scene without any occlusions; the full-overlap adding, which inserts the new object to fully cover some existing objects; the partial-overlap adding, which inserts the new object with partial occlusion.

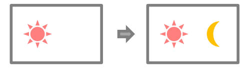
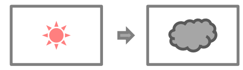
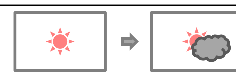
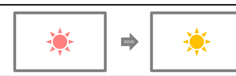
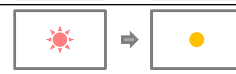
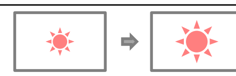
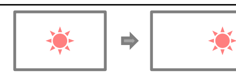
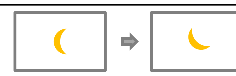
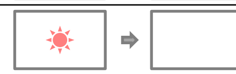
Replace, including replacing the appearance of the objects and replacing one object to other candidate objects. The full-overlap adding could be regarded as a special case of the object replacing, but not exactly equivalent if the new object is smaller than that of the original object.

Shift, some classical geometric transformations, including the scaling, translation and rotation.

Delete, that means to remove the selected object from the original image.

The above operations are summarized in Table I with a series of concise illustrations. In the procedure of the target generation, the amounts of the required operations mainly depend on the amounts of the variances. Empirically, suitable amounts

Table I: The relationship of the operation levels for the object manipulations.

Basics	Refined Behaviors	Operations Description	Illustrations
Add	Non-Overlap Adding	Insert obj. B into a scene that obj. A exists.	
	Full-Overlap Adding	Insert obj. B to completely overlap obj. A.	
	Partial-Overlap Adding	Insert obj. B to partially overlap obj. A.	
Replace	Color Replacing	Specify the appearance of obj. A to a new color	
	Object Replacing	Replace obj. A to obj. B	
Shift	Scaling	Change the size of obj. A	
	Translation	Move the obj. A	
	Rotation	Rotate the obj. A	
Delete	——	Remove the obj. A	

of variances range from 3 to 5. Too much or too less variances would be lower the enjoyment. To simplify the interaction procedure, we assume each performed operation is independent.

3.2. Candidate objects localization and selection

After defining the types of operations, we come to the second problem, how to find the candidates which could be edited. Image parsing [Tighe and Lazebnik 2013] provides a way to analyze and recognize the content of image scene. However, our goal requires elaborative object and scene manipulation to make the variances hard to be found visually. This is a major barrier in automatic image parsing techniques to obtain the satisfactory recognition performance for the region boundaries and the objects. Therefore, we alternate to take the strokes-based interaction to indicate the candidate objects and locations instead.

Strokes-based interaction effectively addresses lots of challenging problems in automatic processing, and it provides a flexible and concise way for complicated operations. Therefore, strokes-based interaction has a lot of successful applications in image editing [Cheng et al. 2010], such as graph-cut segmentation [Rother et al. 2004], colorization [Xu et al. 2013]. In our framework, we utilize strokes-based interaction to complete the object selection task.

However, as aforementioned, it refers to nine refined behaviors with regard to creating the target image. So it is not feasible to just apply binary strokes for the foreground and background strokes as the graph-cut segmentation model [Rother et al. 2004]. Inspired by the work for the colorization problem, in our framework, different operations are described by the corresponding strokes with the distinctive colors. However, abun-

(a) (b) (c)

Fig. 3: User strokes for object selection and localization. (a) Original image. (b) The corresponding operations with the defined color strokes. (c) A result with highlights.

dant colors would cause inconvenience in practical use. Hence, for convenience, we limit four colors to correspond to the basic operations, see Table I.

Besides the advantage of the object selection, strokes could be also applied to locate the new positions of the manipulated objects. Due to the complexity of the spatial directions and the relative locations, automatic operation would face amounts of computational burden. By contrast, strokes-based model could effectively simplify the selection procedures of the object directions and positions. Four defined operations with specific colors are presented in Fig. 3. Note that, we clearly note the dual functions of the strokes-based model for selecting and locating the editable object.

3.3. Scene-driven candidate object retrieval

The final problem that should not be ignored is how to guarantee the harmony between the local modification and global object impression in the target image. This is the key to establish the inconspicuous effect, because adding or generating an object significantly different from its context would expose it easily. In particular, operations like adding or replacing might introduce objects that should not appear in certain scenes, like adding a tree in an indoor image. Hence, we utilize a scene-driven candidate object retrieval method to retrieve suitable objects in a given scene.

First, we set up our object library by extracting object from SUN Database [Russell et al. 2008] [1]. Each specific type of scene contains some common object categories, for instance, as illustrated in Fig. 4, an image of bedroom might contain objects such as books, lamp and doll, while these objects are uncommon in outdoor images. Since these objects are closely related with human cognition, we can use them as a high-level feature to categorize images into various scenes, and retrieve similar candidate objects in the same category. In our framework, for each type of scene, we exploit the aforementioned strokes-based interaction to extract its common objects and store them in our object library. And we implement the Bag of Objects [Yang et al. 2012] model to achieve scene classification.

After locating the candidate objects in the original image, we project these objects into the vocabulary of the Bag of Objects model by the nearest neighbor method, which calculates the similarity between two objects based on their shape of strokes and other low-level image features (e.g. SIFT, HOG [Yang et al. 2012], [Sun et al. 2015], etc.). Then the category of the original image could be predicted, and objects in the same scene category would be retrieved for further operations.

[1] http://groups.csail.mit.edu/vision/SUN/

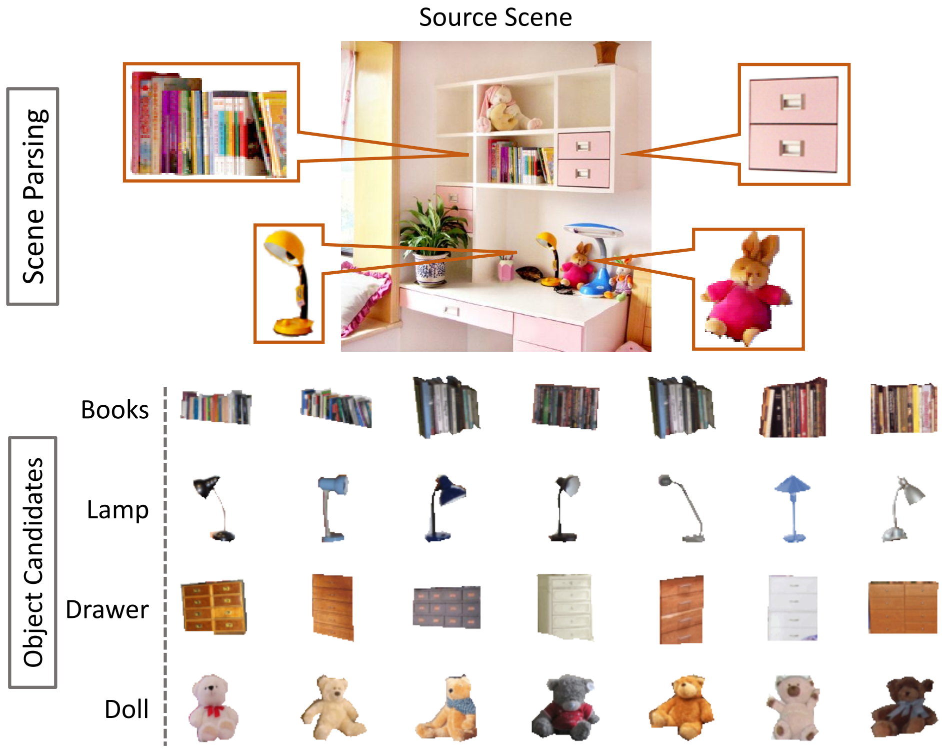


Fig. 4: New candidate objects could be found from object library. Analyzing the scene, we easily indicate some objects and retrieve corresponding similar candidates by the shape matching. All the object data are collected from MIT SUN Dataset with LabelMe toolbox in our experiments.

4. DUAL-DOMAIN PERCEPTUAL EVALUATION FRAMEWORK

As aforementioned in Sec. 1, there are some state-of-the-art frameworks in image editing. By exploiting these tools, the basic image operations could be completed. However, the challenge is how to make the modifications in the target image that are hard to be found with respect to human visual perception. To achieve this goal, we propose a Dual-domain Perceptual Evaluation Framework (DPEF), which evaluates the perception errors to measure the differences between the original and the target.

4.1. Dual-domain perception error minimization model

According to the characteristic of the human visual perception, we are sensitive to the high-contrast intensity variance, color saturation, and object orientation when we observe an image. On the basis of this observation, we consider to construct a perception error model from the aspect of the image saliency theory. In our goal, the variances would not be found out easily. Therefore, we apply the saliency information inversely to keep the modifications away from the visual attentions. In other words, the fewer distinctions between the corresponding saliency map of the original and that of the target, the harder to find them. In addition, the color histogram reflects the distribution of the color amounts. Hence, it can be used to measure the variances of the object appearance.

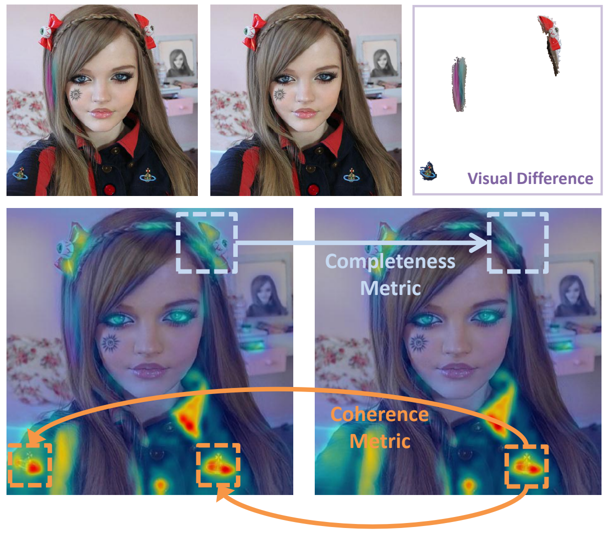


Fig. 5: Perceptual bidirectional similarity metric (PBSM), including completeness metric and coherence metric for selected objects.

Refer to the evaluation in the saliency domain and pixel domain, let E_D be the dual errors with respect to the saliency and appearance measurement, we formulate the following dual-domain perception error model,

$$E_D = E_S + E_A, \tag{1}$$

where E_S denotes the accumulative error in the saliency domain, and E_A denotes the accumulative error in the pixel domain with respect to the object and scene appearance.

4.2. Perceptual bidirectional similarity metric

To measure the "visual similar" between the original image and the target image, we introduce a new model inspired by the bidirectional similarity measurement. This model evaluates the distinction by cumulating the error distance from the corresponding sampled patch pairs. The key observation is that, the completeness assures that all patches from the original image are found in the output, while the coherence assures their coherent merging, see Fig. 5.

However, although the distinctions could be recognized in the pixel domain, this model does not satisfy the requirement with respect to avoiding the salient objects. To further remedy this defect, we proposed a perceptual bidirectional similarity metric in the saliency domain to evaluate the target image. In our model, F denotes the original image, and T denotes the target image. Let P and Q represent patches in F and T, respectively. Then, the cumulative error is defined in the saliency domain as

$$\begin{aligned} E_S(F,T) &= E_S^{Complete}(F,T) + E_S^{Coherent}(F,T) \\ &= \frac{1}{N_F^{\kappa,\sigma}} \sum_{P \subset F} \min_{Q \subset T} E_s(P,Q) \\ &+ \frac{1}{N_T^{\kappa,\sigma}} \sum_{Q \subset T} \min_{P \subset F} E_s(Q,P), \end{aligned} \tag{2}$$

where $N_F^{\kappa,\sigma}$ and $N_T^{\kappa,\sigma}$ denote the number of patches in F and T with patch size κ and sampling interval σ, respectively. And E_s is a distance for the difference between patches Q and P.

The perceptual bidirectional similarity metric is illustrated in Fig. 5. The distinction between the original image and the target image could be recognized in the pixel level. As aforementioned, this visual difference cannot reflect the degree of the visual attention. By exploiting the saliency information, the perceptual difference will be measured. Meanwhile, through various patch scales, the bidirectional similarity index of specific regions will be obtained.

4.3. Appearance similarity metric

Perceptual bidirectional similarity metric evaluates the similarity between the original image and the target image from the aspect of the salient visual perception. This metric can effectively reflects the variances of the salient objects at the spatial locations. However, for the variance on the object appearance, PBSM has obvious limitation. To evaluate the appearance variance, we introduce the Kullback-Leibler distance [Su et al. 2014] to measure the distinction from the color distribution. The minimization of K-L distance means the color appearance of the target close to that of the original image. Let $\rho\left(P\right)$ and $\rho\left(Q\right)$ denote the distributions of corresponding patches of the original image and the target image, respectively, we have

$$\begin{aligned} E_A &= \min_{P\subset F, Q\subset T} E_{\text{KL}}\left(\rho\left(Q\right)\|\,\rho\left(P\right)\right) \\ &= \min_{P\subset F, Q\subset T} \sum \rho\left(Q\right)\ln\frac{\rho\left(Q\right)}{\rho\left(P\right)}. \end{aligned} \tag{3}$$

In Fig. 6, with kinds of color variance ratios, the subjective illustrations, one-dimensional statistical color distributions, and corresponding K-L distance value are shown. Observing the figure, we will know two important points. One is that we can catch the subtle appearance changing under the statistical measurement; the other is that the different ratios of color changing can be reflected by K-L distance.

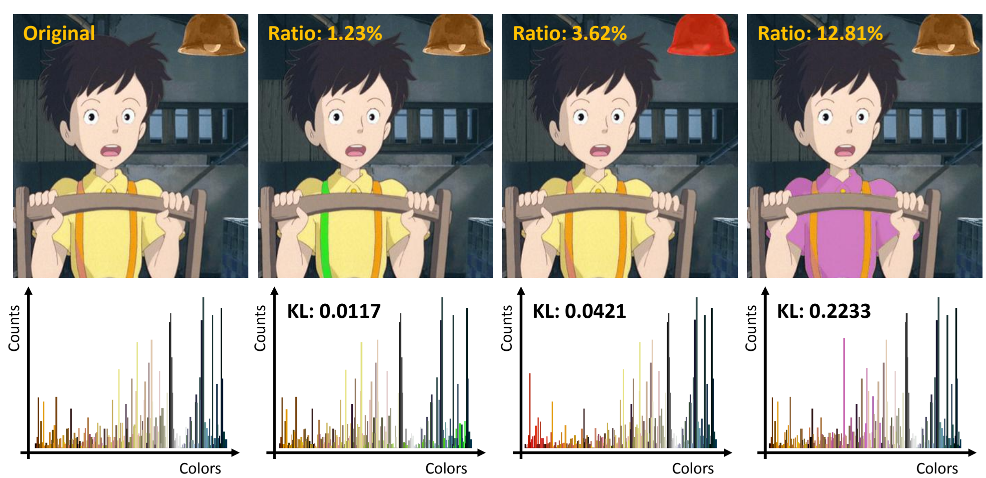


Fig. 6: Appearance similarity metric with K-L distance. The variance ratios are presented in the up-left location of each sub-figure. And the corresponding K-L distance is presented in the sub-figure of color histogram.

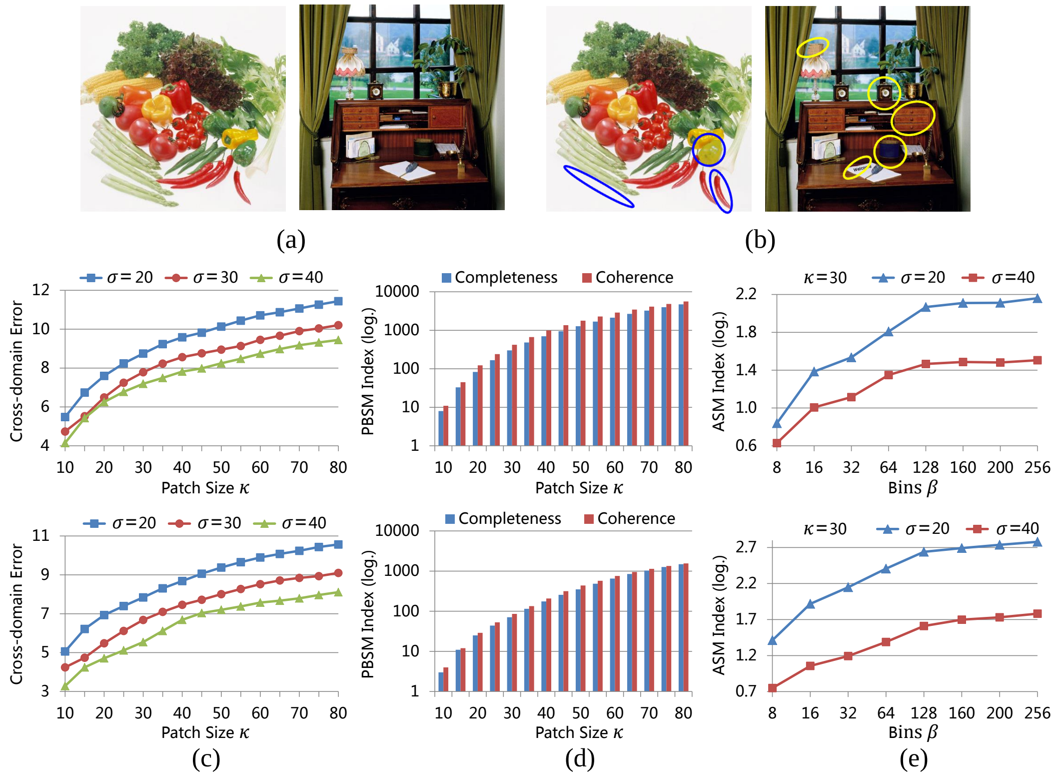


Fig. 7: Parameter analysis. (a) Original image. (b) Target image with variances labeling. (c) Relationship of changing the patch size κ and the sampling interval σ. (d) The completeness and coherence indexes with σ=20 as κ increases. (e) ASM index with different bins (intensity levels) in the statistical procedure.

4.4. Evaluation ranking

As aforementioned in Sec. 3, due to the complexity of the object size, orientation, and location, various possibilities would be yielded in the procedure of the object manipulations. To generate the satisfactory results, we measure the quality of the candidate objects by making a combination to perceptual bidirectional similarity metric (PBSM) and appearance similarity metric (ASM). However, it is important to note that the PBSM values E_S and the ASM values E_A have different dimensions of quantity. In our model, we take a logarithmic operation to produce a robust index E_C, that is,

$$E_C = \log\left(E_S + \mathbf{1}\right) + \log\left(E_A + \mathbf{1}\right), \tag{4}$$

where 1 is a unit constant for avoiding the negative indexes. At last, by performing a ranking algorithm to all the candidate target images, the error minimization one would be the best output. Further, prior k candidates also would be regarded as the selected output if required. The implementation of the pesudo code is presented in Algorithm 1.

5. EXPERIMENTAL RESULTS

In this section, from aspects of parameter setting, independent operation, multiple operations, and time performance, we will demonstrate the effectiveness and applicabil-

Algorithm 1: Dual-Domain Perceptual Evaluation

Input: F: original image, T: target image, κ: patch size, σ: sampling interval.
Output: E: measurement index.

Initialization:
  $S_F \leftarrow Sal(F), S_T \leftarrow Sal(T), H_F \leftarrow Hist(F), H_T \leftarrow Hist(T)$.
Completeness metric:
for i **to** $Num(P)$
  $E_{com} = \textbf{sum}(\min{(S_F(P_i) - S_T(Q))^2})/N_F$.
end
Coherence metric:
for j **to** $Num(Q)$
  $E_{coh} = \textbf{sum}(\min{(S_T(Q_j) - S_F(P))^2})/N_S$.
end
K-L metric:
  $E_{kl} = \min(\textbf{sum}(H_T(Q) \cdot \ln(H_T(Q)/H_F(P))))$.
Normalization:
  $E = \log(E_{com} + E_{coh} + 1) + \log(E_{KL} + 1)$.
Return

ity of our proposed method for generate the visual inconspicuous targets. All the experiments were tested on PC with Intel i7-4770 3.4GHz CPU, NVIDIA GTX660 (without GPU acceleration), 32GB DDR3 Ram, and MATLAB 2015a.

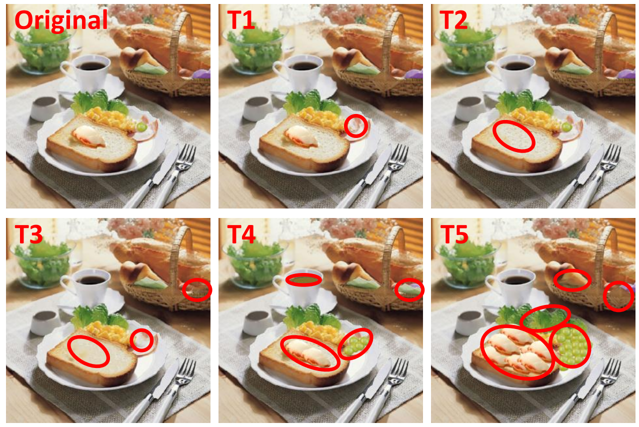
5.1. Relationship of parameter settings

In our implementation, we try to reduce the required parameters as possible. In the stage of the object manipulation, the strokes-based interaction addresses the troubles with respect to the parameter settings. Meanwhile, without the parameter adjustment, the preprocessing object category provides a convenient way to select the candidate new objects. In addition, PatchMatch operation refers to the strokes-based interaction without the complicate parameter control. Known from Sec. 4, our dual-domain perceptual evaluation framework refers to three key parameters, including the patch size κ and the sampling interval σ in PBSM, and the statistical bins β of pixel intensity in ASM.

In Fig. 7, we apply two groups of test data and their corresponding experimental records to analyze the evaluation performance with different parameter settings. The original images and corresponding target images with labels are shown in Fig. 7(a) and (b), respectively. With the patch size κ, the variation tendencies of the total errors are presented in Fig. 7(c). Meanwhile, we record the situations of the sampling step $\sigma = (20, 30, 40)$. According to Eq. 2 and Eq. 3, the larger error index corresponds to the larger variance in the original image. In Fig. 7(d), it presents the situations of the completeness metric and coherence metric with the changing of the patch size. For better observation, we take a logarithmic operation to the indexes. The statistical bins β corresponds to the quantity levels of the image intensity, and their influences to the ASM indexes are illustrated in Fig. 7(e).

Besides the control of the parameter setting, the variance ratio of the image content also produces obvious influence to the measurement indexes. In Table II, we test five samples, and record the indexes in each stage of our framework. Observing the samples and the records, we find that the object changing may be found when the variance ratio is over 2%; and it is very obvious to human vision if the variance ratio is over 10%.

Table II: The influence of various variance ratios to dual-domain perceptual measurements.



Items	T1	T2	T3	T4	T5
Variance Ratio	0.22%	1.08%	1.87%	4.36%	14.48%
Completeness Distance	9	21	30	183	200
Coherence Distance	5	15	20	154	185
Similar Distance	14	36	50	337	385
K-L Distance	0.8785	5.5853	9.0029	11.497	28.644
Log PBSM Index	2.7081	3.6109	3.9318	5.8230	5.9558
Log ASM Index	0.6305	1.8848	2.3029	2.5255	3.3893
Dual-Domain Error	3.3385	5.4958	6.2347	8.3486	9.3451

From the records, we find that our evaluation framework could perceive the slight variances in the image. Under the same scenes, there exist a directly proportional relationship between the variance ratio and the perception error.

5.2. Independent operation analysis

In Table I, we summarize 4 types of basic user behaviors, including 9 refined object operations. In Fig. 8, we make an independent analysis to each user behavior. The replace behavior, containing color replacing (middle) and object replacing (right), are presented at the first row of Fig. 8. By analyzing, we know that the appearance variance would be more obvious to human vision, that is, it is easy to be found.

The add behavior is shown in the second row, including the non-overlap adding (left), the partial-overlap adding (middle) and the full-overlap adding (right). Generally speaking, the non-overlap adding may saliently change the original content structure or the object distributions, so it would be stronger to the visual perception. On the contrast, the partial-overlap adding is usually similar to the neighbor objects, which is not easy to be found. Moreover, the full-overlap adding requires the new object larger than the covered object. Therefore, the error would become larger in practice.

The third row presents the shift behavior, including scaling, translation, and rotation. The small variances appear in the images and the lower errors are recorded, since the scaling and the rotation usually perform on the original object locations. By contrast, the translation refers to the information complement at the original loca-

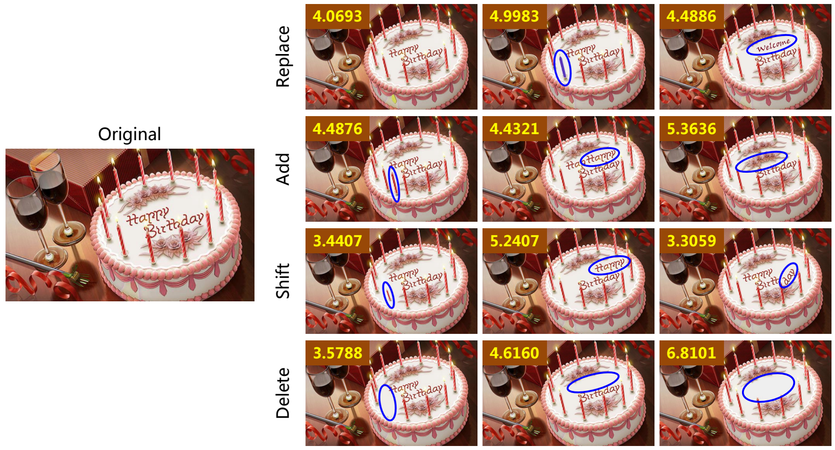


Fig. 8: Independent operation analysis. The original image and the corresponding 4 types of basic behaviors are presented. The recorded indexes are evaluated with κ=30, σ=30, β=256.

tions and the content fusion at the new locations. In other words, it would make larger variance to the image that caused higher index.

The delete behavior is presented in the last row. In fact, the delete behavior equals to the procedure of image complement, that is, filling some visual similar pixels to the missing regions from the neighbor domain. Here, three scales of the operation are shown. From the measurement records, we know that the changing index exactly reflects the variance degree of the image content.

5.3. Multiple operations analysis

Actually, a candidate target usually contains multiple operations for objects or scenes. As aforementioned, the amounts of the edited objects should be limited into 3 to 5. It would reduce the enjoyment and the quality of the original image if the amount is beyond or under this range.

In Fig. 9, we make a comparison between the hand-craft copy in Photoshop tools and our result in the proposed framework. In Fig. 9(a), five selected regions/objects are labeled in expected behaviors. Intuitionally, inconspicuous regions/objects are preferred to be selected. Fig. 9(b) is a well editing sample in common Photoshop operations. Generally speaking, the salient variances have been avoided in the most extent. By con-

Fig. 9: Instance analysis. (a) Original. (b) Photoshop editing. (c) Our result. The recorded indexes are evaluated with κ=30, σ=30, β=256.

Fig. 10: Multiple operations analysis. In each group, the first row contains the original image, the best target image, and the corresponding visual difference with operation labels. The second row lists the sorted candidate target images with their corresponding evaluation indexes.

trast, our result in Fig. 9(c) has a lower evaluation index than the Photoshop output, while keep a comparable visual result to Fig. 9(b).

In Fig. 10, we present 4 groups of images. In each group, it shows the original image, the best target image, the corresponding visual difference, and the sorted candidate image sets. By analyzing these experimental results, we know that the evaluated indexes effectively reflect the perceptive degree to the object variances in the human vision. However, we should note that, this index is just a quantity evaluation, rather than the determined visual perceptual procedure. Through our experiments, we demonstrated that a sound result could be obtained under this indication.

5.4. Performance analysis

Known from Eq. 2 and Eq. 3, our method performs the computation by patch sampling. Therefore, the computational cost is related to the original image size, patch size and sampling interval, rather than the amounts of the edited objects. A naïve brute-force implementation of our method is $O((\kappa M/\sigma)^2)$ for image regions of size M that contains patches of κ^2 pixels with σ steps. As illustrated in Fig. 7(c), the error would be affected by the adjustments of the patch size and the sampling interval. Fortunately, these adjustments do not affect the evaluation results. In addition, the adjustment of the patch size and sampling interval may adapt to the various size of the input image, and further raise the computational efficiency.

5.5. Quantitative analysis and user study

To demonstrate our framework whether indicates the extent of the difference or not, we perform the quantitative analysis and the user study to help to analyze our framework. We select 10 group images and their corresponding visual inconspicuous counterparts,

Fig. 11: Test samples are used for user study. Each group contains three modifications and has been labeled on the corresponding counterpart.

including portrait, landscape, furnishing, food, and cartoon. Each group contains three inconspicuous modifications and has been labeled on the corresponding counterpart. Illustrated as Fig. 11, the small subfigures are the original images, and the large ones are the modified counterparts.

For these test samples, we select some common image quality metrics to objectively measure the variance between the original image and its modified counterpart. We adopt these metrics as the comparison baselines, including Mean Square Error (MSE), Structural SIMilarity (SSIM) [Wang et al. 2004], Normalized Cross Correlation (NCC) [Shen et al. 2015], and Area Error (A-Err) by the amount of intensity variance. In fact, MSE reflects the difference between the original-modified pairs under the Euclidian distance. SSIM reflects the similarity of the original-modified pairs from the content structures. NCC describes the content relationship in the pairs. And A-Err sums the amounts of modified pixels to present the scale of modification. See from Table III, the SSIM and NCC indexes are close in all test samples that means hard to represent the scale of the modification. And the irregular MSE and A-Err indexes do not represent the direct relationship with the visual perception. Note the records, our method may represent the scale of the modification effectively.

On the basis of the quality metrics, we further simulate the procedure of the “Find the Difference” game. The specified different regions/objects are found and the con-

Table III: Objective baselines vs. our metric. **MSE**: Mean Square Error. **SSIM**: Structural SIMilarity [Wang et al. 2004]. **NCC**: Normalized Cross Correlation [Shen et al. 2015]. **A-Err**: Area Error by the amount of intensity variance. The tabs (S.1-S.10) represent the corresponding tested samples in Fig. 11.

Items	S.1	S.2	S.3	S.4	S.5	S.6	S.7	S.8	S.9	S.10
MSE	5.960	0.445	1.159	1.747	0.512	5.208	0.654	2.050	2.651	1.898
SSIM	0.978	0.991	0.995	0.970	0.989	0.983	0.990	0.972	0.973	0.991
NCC	0.991	1.000	0.999	0.997	0.999	0.965	0.965	0.998	0.994	0.997
A-Err	6905	1593	1438	2163	675	908	1289	2843	2767	1805
Ours	8.422	5.649	5.753	6.791	4.822	6.752	6.708	4.798	5.871	7.413

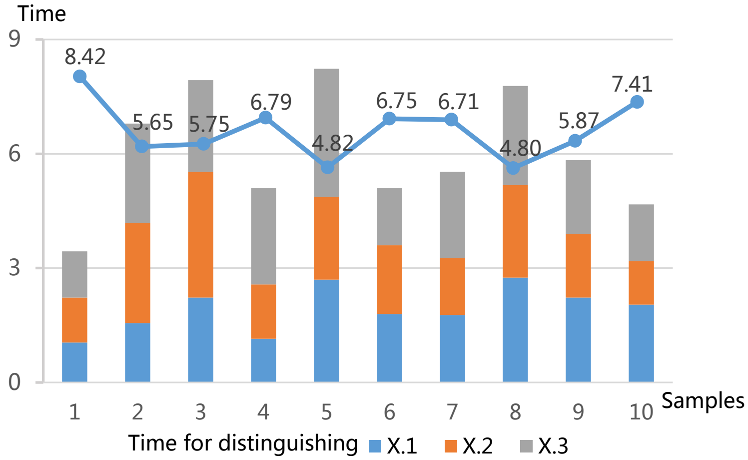


Fig. 12: User study. The piecewise bar graph records the average time consuming for finding the differences, and the curve labels the evaluation index of the test samples.

suming time is recorded. In our test, we invited 32 undergraduate students to attend. To eliminate some possible abnormal situation (e.g. overtime to find a modification, wrong selection), all the samples are tested in multiple random validations (time $>$ 3). Then we perform statistical average to analyze the recorded data. In addition, to suppress the perceived acceleration resulted from the location recall, the sample itself cannot continuously appear (sample interval > 2).

As illustrated in Fig. 12, we apply the piecewise bar graph and the curve graph to analyze the recorded data, respectively. The piecewise bar graph presents the triple average time cost in each group, and the additional label along the curve indicates the evaluation index with our measurement. As known from Eq. (1), larger index represents more inconsistent between the original and the modified counterpart. In other words, we need less time to find out the difference. Otherwise, the small index means more time for finding. In Fig. 12, we demonstrate the relationship between our frame-

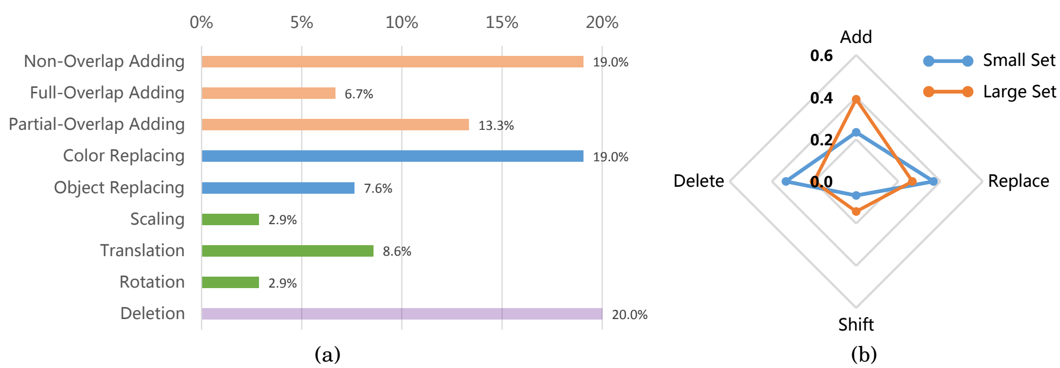


Fig. 13: The frequency of specific behaviors. (a) The refined behaviors performed on 35 tested samples are recorded. (b) The case comparison between the small set (10 samples) and the large set (35 samples).

work and the time consuming of finding, and further demonstrate the effectiveness of ours.

To explore the potential relationship between each operation and the visual perception, we collected 35 original-modified pairs (including 10 pairs in Fig. 11), and then recorded the frequency of utilization with 9 refined behaviors in 4 types of operations. According to the statistical analysis in Fig. 13(a), we found that the non-overlap adding, the color replacing and the deletion have higher frequency in the editing procedure. That means people tend to regard these three operations may yield unimpressive modification. In addition, as illustrated in Fig. 13(b), we apply the radar chart to help analyze the frequency of utilization with the small set (10 samples) and the large set (35 samples) in our framework.

5.6. Extensive applications

The dual-domain perception error model can be regarded as a kind of the content similarity metrics. It may be applied to evaluate the quality of both scale-consistent and scale-inhomogeneous image pairs. This is an superior characteristic that the model could be extended to some visual applications for quality measurement. In Fig. 14, we present tow extensive cases. One is the blurring evaluation, the other is the retargeting measurement. In the blurring case, we exploit the index to discriminate the blurred extent. In the retargeting case, the index may evaluate the quality of the results of different methods [Rubinstein et al. 2010]. The larger index means the more distortion at the salient objects.

Fig. 14: Extensive applications. The dual-domain perception error model is applied to evaluate the blurring situation and the quality of retargeting results. The retargeting samples are collected from RetargetMe dataset [Rubinstein et al. 2010].

6. DISCUSSIONS AND CONCLUSIONS

To generate the visual inconspicuous counterparts, we not only face the problem of how to edit and manipulate the objects. A more challenging problem is how to make the variances hard to be found in the target image. The latter has two potential characteristics. One requires considering the effect of the image editing, the other emphasizes the visual perception in the human's observation. Essentially, the key point in this problem is how to establish an objective measurement index to evaluate the distinctions between the original image and the target image. This index should not only recognize the slight variances in the target image, but also make certain estimation to the visual perception.

In this paper, we propose a dual-domain perceptual editing framework to generate the corresponding visual inconspicuous images. We first exploit the object saliency information to simulate the visual perception, and then create a perceptual bidirectional similarity metric to evaluate the geometry variances of the image contents. The combination of the saliency and bidirectional similarity metrics provides a feasible way to exactly evaluate the object variances. From the viewpoint of statistical colors distribution, the K-L distance is introduced to evaluate the variance of the object appearance. This metrics could reflect the influence of the color changing to the human vision system.

Since the object manipulation refers to complicated procedures of the image editing, we summarize the possible behaviors and apply the strokes-based interaction with PatchMatch model [Barnes et al. 2011] to simplify the editing operations. For the data sources, we not only consider extract the edited objects from the original image itself, but also take advantage of the well-known databases. By analyzing the experimental results, we demonstrate the effectiveness and applicability of our framework.

Relationship. Actually, to generate the visual inconspicuous counterparts, it refers to lots of computer vision and graphics operations from aspects of data preparation, image editing, and quality measurement. Distinguished to the previous work (e.g. PatchMatch [Barnes et al. 2011], Sketch2Photo [Chen et al. 2009], RepFinder [Cheng et al. 2010]), we emphasize on the distinctions evaluation between the original image and the target image. This evaluation would be close to human visual perception by exploiting the object saliency information. In our experiments, we demonstrate that the index conforms to the variance tendency of object amounts and scales. However, we should note that this evaluation is just a feasible quantity index, rather than an absolute estimation to the human vision.

Strengths. In one side, this framework exploits the visual attentions information in the saliency domain to simulate the procedure of the perception. In the other side, it exploits the statistical colors distribution to reflect the object appearance in the pixel domain. The combination of these characteristics could effectively raise the evaluation accuracy with respect to the visual distinctions. The reason is that our dual-domain perceptual evaluation framework could describe the slight variances in the target image.

Limitations. However, there still exists a certain limitation in the current model. This is reflected in that some structural shapes are not well defined in the current framework, such as the circularity and symmetry. Usually, we are sensitive to the structural information when we observe a scene. The rearrangement of the object structures may be salient to vision. However, it is hard to directly reflect in the evaluation index, because the current framework just depends on the patch sampling without structural guidance.

Future. In the future, we will continue to improve the recognition ability with respect to the geometry information in the current framework, such as the symmetry. In addition, we will mine the potentials of the framework to apply to some high subjective-dependent fields, such as the image/video retargeting, edge-preserving smoothing, and image enhancement.

ACKNOWLEDGMENTS

We appreciate all data collectors for providing the experimental materials, and all testees in our user studies. Kun Zeng is the corresponding author of this paper (E-mail: kunzeng@outlook.com).

REFERENCES

Connelly Barnes, Dan Goldman, Eli Shechtman, and Adam Finkelstein. 2011. The PatchMatch randomized matching algorithm for image manipulation. *Commun. ACM* 54, 11 (2011), 103–110.

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan Goldman. 2009. PatchMatch: a randomized correspondence algorithm for structural image editing. In *proceedings of ACM SIGGRAPH 2009 papers*, Vol. 28. 1–11.

Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2011. A Holistic Approach to Aesthetic Enhancement of Photographs. *ACM Trans. Multimedia Comput. Commun. Appl.* 7S, 1 (2011), 21:1–21:21.

Ali Borji, Dicky Sihite, and Laurent Itti. 2012. Salient object detection: A benchmark. In *proceedings of 12th European Conference on Computer Vision*. 414–429.

Tao Chen, Mingming Cheng, Ping Tan, Ariel Shamir, and Shimin Hu. 2009. Sketch2Photo: Internet image montage. In *proceedings of ACM SIGGRAPH Asia 2009*, Vol. 28. 124:1–124:10.

Mingming Cheng, Fanglue Zhang, Niloy Mitra, and Xiaolei and Huang. 2010. Repfinder: Finding approximately repeated scene elements for image editing. *ACM Transactions on Graphics* 29, 4 (2010), 1–8.

Mingming Cheng, Guoxin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shimin Hu. 2011. Global contrast based salient region detection. In *proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 409–416.

Mingming Cheng, Guoxin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shimin Hu. 2015. Global contrast based salient region detection. *to appear in IEEE Transactions on Pattern Analysis and Machine Intelligence* (2015).

Taeg Cho, M. Butman, S. Avidan, and W. Freeman. 2008. The patch transform and its applications to image editing. In *proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

A. Criminisi, P. Perez, and K. Toyama. 2003. Object removal by exemplar-based inpainting. In *proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2. 721–728.

Kostas Daniilidis, Petros Maragos, Nikos Paragios, Connelly Barnes, Eli Shechtman, Dan Goldman, and Adam Finkelstein. 2010. The Generalized PatchMatch Correspondence Algorithm. In *proceedings of European Conference on Computer Vision*. 29–43.

Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. 2012. What makes Paris look like Paris? *ACM Transactions on Graphics* 31, 4 (2012), 1–9.

Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. 2009. PhotoSketch: a sketch based image query and compositing system. In *SIGGRAPH 2009: Talks*. New Orleans, Louisiana, 1–4.

Zeev Farbman, Gil Hoffer, Yaron Lipman, Daniel Cohen-Or, and Dani Lischinski. 2009. Coordinates for instant image cloning. In *proceedings of ACM SIGGRAPH 2009 papers*, Vol. 28. 1–9.

Chen Goldberg, Tao Chen, Fanglue Zhang, Ariel Shamir, and Shimin Hu. 2012. Data-Driven Object Manipulation in Images. *Computer Graphics Forum* 31, 2 (2012), 265–274.

Jonathan Harel, Christof Koch, and Pietro Perona. 2007. Graph-based visual saliency. In *proceedings of 20th Annual Conference on Neural Information Processing Systems*. 545–552.

Shimin Hu, Fanglue Zhang, Miao Wang, Ralph Martin, and Jue Wang. 2013. PatchNet: a patch-based image representation for interactive library-driven image editing. *ACM Transactions on Graphics* 32, 6 (2013), 196:1–12.

Hui Huang, Kangxue Yin, Minglun Gong, Dani Lischinski, Daniel Cohen-Or, Uri Ascher, and Baoquan Chen. 2013. "Mind the gap": tele-registration for structure-driven image completion. *ACM Transactions on Graphics* 32, 6 (2013), 1–10.

L. Itti, C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.

Jiaya Jia, Jian Sun, Chikeung Tang, and Heungyeung Shum. 2006. Drag-and-drop pasting. In *proceedings of ACM SIGGRAPH 2006 Papers*. 631–637.

Weisi Lin and C. Jay Kuo. 2011. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation* 22, 4 (2011), 297–312.

Tie Liu, Jian Sun, Nanning Zheng, Xiaoou Tang, and Heungyeung Shum. 2007. Learning to Detect A Salient Object. In *proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 33. 1–8.

Cewu Lu, Li Xu, and Jiaya Jia. 2014. Contrast Preserving Decolorization with Perception-Based Quality Metrics. *International Journal of Computer Vision* 110, 2 (2014), 222–239.

Anush Moorthy and Alan Bovik. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing* 20, 12 (2011), 3350–3364.

Patrick Perez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. In *proceedings of ACM SIGGRAPH 2003 Papers*. 313–318.

Yael Pritch, Eitam Kav-Venaki, and Shmuel Peleg. 2009. Shift-map image editing. In *proceedings of 12th International Conference on Computer Vision*. 151–158.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. “GrabCut” - interactive foreground extraction using iterated graph cuts. In *proceedings of ACM SIGGRAPH 2004 Papers*. 309–314.

M. Rubinstein, D. Gutierrez, O. Sorkine, and A. Shamir. 2010. A Comparative Study of Image Retargeting. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2010)* 29, 5 (2010), 160:1–10.

Bryan Russell, Antonio Torralba, Kevin Murphy, and William Freeman. 2008. LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision* 77, 1-3 (2008), 157–173.

Ariel Shamir and Olga Sorkine. 2009. Visual media retargeting. In *ACM SIGGRAPH ASIA 2009 Courses*. 1–13.

X. Shen, C. Zhou, L. Xu, and J. Jia. 2015. Mutual-Structure for Joint Filtering. In *2015 IEEE International Conference on Computer Vision (ICCV)*. 3406–3414.

D. Simakov, Y. Caspi, E. Shechtman, and M. Irani. 2008. Summarizing visual data using bidirectional similarity. In *proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

Mingli Song, Dacheng Tao, Chun Chen, Xuelong Li, and Chang Chen. 2010. Color to Gray: Visual Cue Preservation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (2010), 1537–1552.

Z. Su, K. Zeng, L. Liu, B. Li, and X. Luo. 2014. Corruptive Artifacts Suppression for Example-Based Color Transfer. *IEEE Transactions on Multimedia* 16, 4 (2014), 988–999.

Jian Sun, Lu Yuan, Jiaya Jia, and Heungyeung Shum. 2005. Image completion with structure propagation. *ACM Transactions on Graphics* 24, 3 (2005), 861–868.

Shaoyan Sun, Wengang Zhou, Qi Tian, and Houqiang Li. 2015. Scalable Object Retrieval with Compact Image Representation from Generic Object Regions. *ACM Trans. Multimedia Comput. Commun. Appl.* 12, 2 (2015), 29:1–29:21.

Michael Tao, Micah Johnson, and Sylvain Paris. 2013. Error-Tolerant Image Compositing. *International Journal of Computer Vision* 103, 2 (2013), 178–189.

Joseph Tighe and Svetlana Lazebnik. 2013. Superparsing: Scalable nonparametric image parsing with superpixels. *International Journal of Computer Vision* 101, 2 (2013), 329–349.

Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

Pohung Wu, Chienchi Chen, Jianjiun Ding, Chiyu Hsu, and Yingwun Huang. 2013. Salient Region Detection Improved by Principle Component Analysis and Boundary Information. *IEEE Transactions on Image Processing* 22, 9 (2013), 3614–3624.

Yulin Xie, Huchuan Lu, and Minghsuan Yang. 2013. Bayesian Saliency via Low and Mid Level Cues. *IEEE Transactions on Image Processing* 22, 5 (2013), 1689–1698.

Li Xu, Qiong Yan, and Jiaya Jia. 2013. A sparse control model for image and video editing. *ACM Transactions on Graphics* 32, 6 (2013), 197:1–10.

Yang Yang, Linjun Yang, Gangshan Wu, and Shipeng Li. 2012. A bag-of-objects retrieval model for web image search. In *proceedings of 20th ACM International Conference on Multimedia (SIGMM 2012)*. 49–58.

Kun Zeng, Mingtian Zhao, Caiming Xiong, and Songchun Zhu. 2009. From image parsing to painterly rendering. *ACM Transactions on Graphics* 29, 1 (2009), 1–11.

Fanglue Zhang, Mingming Cheng, Jiaya Jia, and Shimin Hu. 2012. ImageAdmixture: Putting Together Dissimilar Objects from Groups. *IEEE Transactions on Visualization and Computer Graphics* 18, 11 (2012), 1849–1857.

Fanglue Zhang, Miao Wang, and Shimin Hu. 2013. Aesthetic Image Enhancement by Dependence-Aware Object Recomposition. *IEEE Transactions on Multimedia* 15, 7 (2013), 1480–1490.

Mingtian Zhao and Songchun Zhu. 2013. Abstract painting with interactive control of perceptual entropy. *ACM Transactions on Applied Perception* 10, 1 (2013), 1–21.

Wang Zhou and Li Qiang. 2011. Information Content Weighting for Perceptual Image Quality Assessment. *Image Processing, IEEE Transactions on* 20, 5 (2011), 1185–1198.

Received February 2007; revised March 2009; accepted June 2009