Multimed Tools Appl (2019) 78:1399–1418
https://doi.org/10.1007/s11042-018-6186-z

Image quality assessment for inpainted images via learning to rank

Mariko Isogawa[1] · **Dan Mikami**[1,2] · **Kosuke Takahashi**[1] · **Hideaki Kimata**[1]

Received: 9 October 2017 / Revised: 5 April 2018 / Accepted: 22 May 2018 /
Published online: 21 June 2018

Abstract This paper proposes an image quality assessment (IQA) method for image inpainting, aiming at selecting the best one from a plurality of results. It is known that inpainting results vary largely with the method used for inpainting and the parameters set. Thus, in a typical use case, users need to manually select the inpainting method and the parameters that yield the best result. This manual selection takes a great deal of time and thus there is a great need for a way to automatically estimate the best result. Unlike existing IQA methods for inpainting, our method solves this problem as a learning-based ordering task between inpainted images. This approach makes it possible to introduce auto-generated training sets for more effective learning, which has been difficult for existing methods because judging inpainting quality is quite subjective. Our method focuses on the following three points: (1) the problem can be divided into a set of "pairwise preference order estimation" elemental problems, (2) this pairwise ordering approach enables a training set to be generated automatically, and (3) effective feature design is enabled by investigating actually measured human gazes for order estimation.

Keywords Learning to rank · Image inpainting · Image quality assessment (IQA)

✉ Mariko Isogawa
mariko.isogawa.kt@hco.ntt.co.jp

Dan Mikami
dan.mikami.vp@hco.ntt.co.jp

Kosuke Takahashi
kosuke.takahashi.rd@hco.ntt.co.jp

Hideaki Kimata
hideaki.kimata.yu@hco.ntt.co.jp

[1] NTT Media Intelligence Laboratories, 1-1 Hikarino-oka, Yokosuka, Japan

[2] NTT Communication Science Laboratories, 3-1 Wakamiya, Morinosato, Atsugi, Japan

1 Introduction

Photos sometimes include unwanted regions such as a person walking in front of a filming target or a trash can on a beautiful beach. Image inpainting is a technique to automatically remove such areas ("damaged regions" in this paper) and restore them [4, 5, 7, 8, 12, 15, 17, 34]. However, it is known that inpainted results vary largely with the method used and the parameters set.[1] For example, He et al.'s method effectively repairs images including horizontally or vertically repeated textures [12] and Huang et al.'s method is especially efficient for structural images [15]. In the conventional approach to obtain the best inpainting results, a user selects the inpainting technique and tunes the parameters by trial and error, while observing the inpainted results. Since this is time consuming and requires special knowledge, a method for automatically selecting the best result is required.

Unfortunately, no method has been established to determine the inpainting method and its parameters before conducting inpainting. Instead, we feel that methods for image quality assessment (IQA) that assess the quality of inpainted images have possibility to tackle this purpose. Assessing quality of inpainted images is widely acknowledged as a task that is difficult for automation because its judgment is quite subjective. To handle such tasks, many IQA methods for inpainting focus on how the gaze gathers when a human watches an unnatural image [3, 10, 23, 26, 28, 29, 31]. Through the assumption that unnatural removal of unwanted region gathers human attention, they tried to find a way to represent subjective quality by means of objectively measurable indicators. To obtain human attention, most of these IQA methods use a computational visual saliency map, which simulates human gaze density [3, 10, 23, 26, 29].

Although the basic idea of using human attention is reasonable, these existing methods are difficult to apply for comparing the qualities of two inpainted images due to the following two factors. One is the difficulty in estimating human attention. Actual human attention changes by contexts such as the reason for viewing. Isogawa et al. [16] revealed that the human gaze pattern while watching inpainted images is different from any computational saliency maps. The other one is the resolution of a saliency map, which is generally coarse. Thus, it is difficult to apply this method to the current task in which the difference resides in a locally ubiquitous way.

Estimation of subjective quality is not unique to inpainting image. In the research field of subjective-evaluation-estimation, learning-to-rank approaches have been investigated actively [1, 6, 11, 18–21, 33]. Although for learning based approaches, large and representative training data sets are essential to improve estimation accuracy, accumulating training data is difficult in view of annotation cost and fluctuation of user annotation. The learning-to-rank framework replaces the subjective evaluation tasks as an ordering task without estimating an absolute score. It is considered that it opens up a new era within the research field of subjective evaluation. The problem setting is quite reasonable and reduces preparation costs since it enables learning without absolute scores annotated by human subjects; selecting the better one is rather easier than providing the scores for images to be subjectively evaluated such as inpainted images. We consider that the learning-to-rank framework has the potential to further reduce the training data accumulation by making good use of ordering traits. Recently, some studies have automatically generated and/or augmented

[1] Inpainting quality is quite substantially affected by parameters such as multi-scale level or patch size to search appropriate regions for restoration. For variations depending on such parameters, please refer to Figs. 10 and 11.

training data by image processing [9, 24, 25]. We believe this concept is also applicable to IQA for inpainting.

In this paper, we show how we tackle the task of obtaining the best inpainted result among inpainted images obtained through various methods and parameters. We also propose a new learning-to-rank based ordering approach for inpainted images. Unlike existing IQA methods for inpainting, our method has a new feature that does not use a computational visual saliency map but uses our investigation of human gaze while watching inpainted images. Another important proposal is automatic generation of training data. By making good use of pair-wise learning, we propose automatic generation of training pairs to improve estimation accuracy. The contributions of this paper are as follows:

- This is the first trial for applying learning-to-rank for IQA of inpainted images.
- The proposed method enables automatically generated training data to be introduced by making good use of a ranking mechanism, although the learning target is quite subjective.
- It proposes new image features dedicated to inpainted image quality assessment on the basis of gaze measurement experiments.

This paper is based on our previous conference proceedings [16] and adds a comprehensive investigation on how the proposed features work and a novel method for accumulating training data automatically to improve estimation accuracy. The rest of this paper is organized as follows. In Section 2 we briefly review related work. Section 3 investigates actual human gazes to design effective image features for learning. Section 4 describes the learning based ranking method we propose, which was developed with the knowledge detailed in Section 3. In Section 5 we verify the method's effectiveness by comparing it with existing IQA methods. We also describe the effectiveness of introducing an auto-generated training set. In Section 6, we conclude the paper with a summary of key points and mention future work to be done.

2 Related work

This section reviews previous studies. Section 2.1 introduces IQA methods for image inpainting whose purposes are the same as ours. Then, in 2.2, we introduce a learning-to-rank approach that has attracted attention as a method for estimating subjective evaluations.

2.1 IQA methods for image inpainting

Estimating quality is one of the difficult issues for image inpainting. The main reasons are the ambiguity in subjective evaluations, and the cost for obtaining training data. Because of the former issue, although many effective IQA methods for degraded images, e.g., burred, compressed, or noised images have been proposed [11, 19–21, 32], these methods cannot be applied assessing inpainted images.

To overcome the former issue and to obtain subjective evaluations stably, there are three main approaches: reflecting human reactions such as gaze transition [23, 26, 28], asking subjects to provide their judgments [10], and combining them [29].

The basic concept of human reaction based IQA is that less natural inpainted regions will gather more gazes because of the unnaturalness for human perception. Thus, this method estimates inpainted image quality on the basis of gaze density before and after images are inpainted [28]. To reduce the cost for measuring actual human gazes, many metrics use

computational visual saliency maps instead of actual gazes [23, 26, 29]. A computational visual saliency map ("saliency map" for short), is a topographically arranged map that represents estimated visual saliency only from the image. If saliency maps well reflect actual human gaze patterns, substituting them for actual gazes will work well. However, the accuracy of saliency maps is unfortunately quite limited as we mention in Section 3, and thus the performance of saliency map-based IQA methods is also limited.

Learning based approaches that depend on support vector regression (SVR) have also been reported [10, 29]. For these methods, subjectively annotated rating scores are essential for training regression models. Because of the need for absolutely subjective scores, all of the training data should be manually annotated, which is the second issue for building an IQA method for inpainting. Thus generating training sets requires quite high annotation cost. To overcome this issue, our method generates training data automatically as described in Section 4.2.

2.2 Ranking based image evaluation for subjective judgment

In many subjective evaluation tasks, it is difficult to provide absolute scores. For example, scoring the degree of smiles is a quite difficult task and the scores may vary largely by question. Since estimating such varied subjective scores is quite difficult, sidestep methods have been widely considered. Learning-to-rank based approaches are now acknowledged as a promising solution. Rather than absolute scores, they provide a learning framework for merely ordering scores among target samples. Coming back to the above cited example, sorting the images by degree of smile is easier than giving smile scores to each image.

Among learning-to-rank approach variants, pairwise learning-to-rank methods have gathered attention due to the ease with which they can be implemented. They have been frequently applied for estimating preference order [1, 6, 18, 33]. Chang et al. estimated the age of a single face image [6]. Yan et al. obtained the most visually appealing color enhancement of an image [33]. Abe et al. estimated the surface qualities of an object, such as glossiness or transparency from its images [1], and Khosla et al. estimated the most memorable region inside images [18].

Learning-to-rank has also been introduced in IQA methods [11, 19–21]. Gao et al. [11] and Xu et al. [19, 20] proposed blind image quality assessment frameworks for degraded images, e.g., blurred or compressed images or images with white noise. In addition, Ma et al. introduced learning-to-rank to assessing retargeted images [21]. Although retargeted images are quite different from general degraded images that existing IQA methods deal with, they examined and investigated the effects of learning-to-rank based IQA for image retargeting.

Unlike existing methods, the method discussed in this paper focuses on assessing inpainted images. Since estimating the quality of inpainted images is a quite different task than assessing other deteriorated images, we designed new image features dedicated for assessing inpainted images. In addition, this paper shows that by using pairwise learning traits we can produce training data automatically and use the data to improve estimation accuracy.

3 Toward effective image features: eye gaze investigation

Many IQA methods use visual saliency maps as substitutes for actual gazes. However, we have doubts about the coherence of computational visual saliency and actual human gazes, especially when observing inpainted images. Therefore, before we go into the proposed

method, we will describe an eye gaze measurement experiment we conducted for two purposes. The first was to show the difference between measured gazes and the saliency map and to reveal the difficulty in using saliency maps instead of actual human gazes for IQA. The second was to analyze the region and features within inpainted images we should focus on to assess the quality of the images on the basis of measured gazes and the corresponding subjective evaluations.

3.1 Procedure and set-up of eye tracking experiment

We conducted this experiment with the aim of verifying coherence between the measured gazes and the saliency map. We also obtained subjective scores for each image to investigate how gazes affected the total subjective image quality. Figure 1 shows the test procedure, in which subjects repeated three tasks: (*a*) stare at a white cross on a black background for two seconds to fix their initial viewpoint, (*b*) observe images for 10 seconds, and (*c*) provide 5-point opinion scores representing image quality unnaturalness. The scores 1-5 respectively corresponded to *Very noticeable, Rather noticeable, Slightly noticeable, Hardly noticeable, and Unnoticeable*. Higher scores are better since they indicate that the unnaturalness that occurs with inpainting is unnoticeable.

The observed image in task (*b*) includes original images and inpainted images. Original images are images that are not inpainted. Inpainted images were generated with two methods, i.e., those reported by He et al. [12] and Huang et al. [15]. The subjects had no prior knowledge on the types of images displayed (i.e., whether they were original images or images that had been inpainted by using the methods reported by He et al. [12] and Huang et al. [15]). To prevent the subjects from having prior knowledge of the material, we generated three types of images generated from each of 100 original images. We asked 24 subjects (8 males and 16 females) with normal vision to report the image quality after observing the displayed images. These subjects were divided into three groups and the subjects in each group watched the same type of image. Each subject watched 100 images. We applied a stationary Tobii eye tracker for gaze measurement. The LCD monitor used for stimulus presentation was 21 inches (1280×1080 pixels). The monitor-observer distance was 60 cm.

3.2 Integrity between computational saliency maps and human visual attention

In Fig. 2, (*a*) shows an inpainting target image and (*b*) shows the inpainted result in which an undesired man standing in front of a boat was removed. Measured human attention is

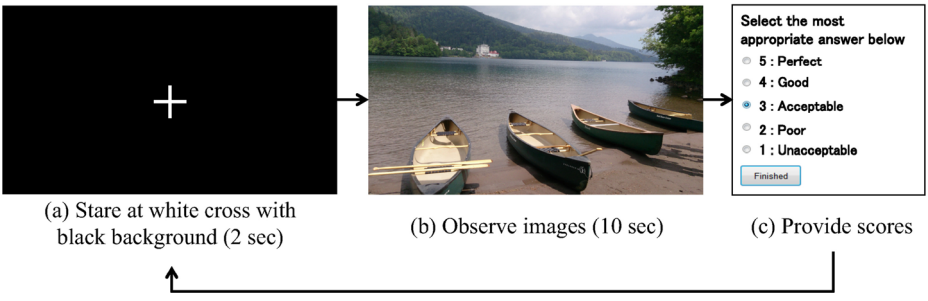

Fig. 1 Three step test procedure of the preliminary experiment conducted to elucidate the relationship between human gaze and subjective scores. Subjects are required to; **a** stare at the white cross to fix initial viewpoint, **b** observe an image, and **c** provide a 5-point opinion score (subjective assessment) of the image

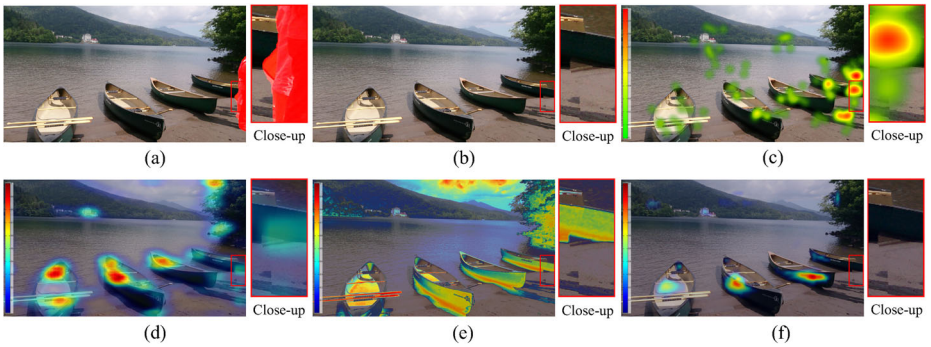

Fig. 2 Comparison between observed human visual attention and computational visual saliency. **a** Inpainting target image. **b** Inpainted image. **c** Human visual attention overlaid on (**b**) (red gathers more gazes). **e**–**f** Computational visual saliency overlaid on (**b**) (red gathers more gazes). Saliency maps are (**d**) Hou et al.'s [14] used in Voronin et al.'s metric[29], **e** Achanta et al.'s [2] used in Trung et al.'s metric [26], and (**f**) Walther et al.'s [30] used in Oncu et al.'s metric [23]

overlaid in (c). Calculated visual saliency maps obtained with calculation methods proposed by Hou et al. [14], Achanta et al.'s [2], and Walther et al.'s [30] are respectively shown in (d), (e), and (f). These maps have actually been used for assessing the image quality of inpainted images; they were respectively used by Voronin et al. [29], Trung et al. [26], and Oncu et al. [23]. From these maps, we can observe that the resolution of computational visual saliency maps is quite coarse and their results are significantly varied. Additionally, saliency maps are quite different from human visual attention. As shown in Fig. 2b, inpainting failed to fill the shape of the boat. Because this failure produces significant unnaturalness, the most salient areas for actual human gazes were those around the damaged boat (See Fig. 2c). The areas around the boat in (d) and (e) were somewhat salient, but were more salient in other areas (e.g., the other boats or the oars). In (f) no saliency around the boat was represented at all. These results suggest that it is difficult to use computational visual saliency maps as a substitute for human gazes. Thus, it is essential to come up with new image features that represent such unnaturalness.

3.3 Correlation between subjective quality and human visual attention to damaged region contours

Human gazes are potentially an excellent means for assessing inpainting quality, and a metric based on human gazes was proposed by Venkatesh et al. [28]. This metric categorizes eye gaze position into two categories, i.e., inside and outside damaged regions. They use the difference of amount of gazes between pre- and post- inpainted images. We were inspired by this simple and effective idea and so tackled further analysis of eye gaze patterns in categories other than inside and outside damaged regions. This section shows how we analyzed what we should focus on to assess the quality of the inpainted images on the basis of knowledge of human attention and corresponding subjective evaluations. We believe that this knowledge will be useful in developing an IQA method for image inpainting.

We analyzed the characteristics of observed gaze and corresponding MOS levels. The MOS values are the average of the 5-point annotated scores provided by subjects as described in the previous section. We first investigated on where human beings tend to watch for inpainted results. Fig. 3a shows an example gaze histogram for the inpainted

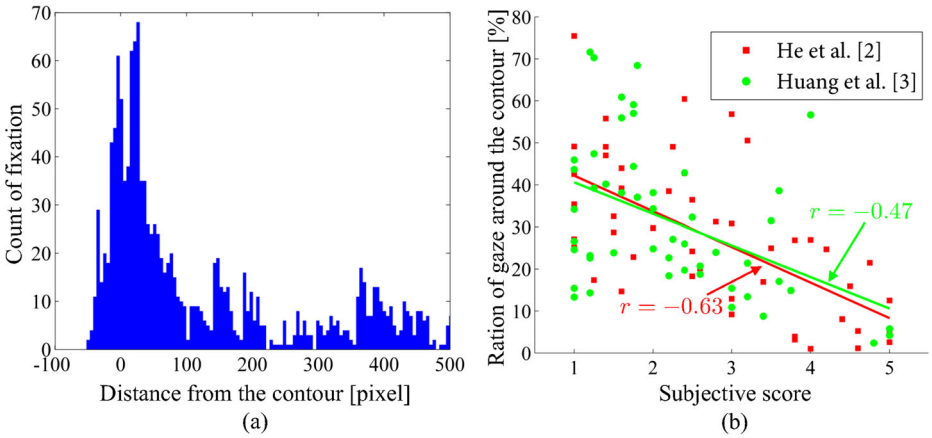

Fig. 3 Gaze measurement results; **a** gaze histogram with Fig. 2b, **b** relationship between the subjective score and density of gaze within the vicinity of damaged region's contour

image in Fig. 2b. Its vertical axis is the time the gaze was oriented and the horizontal axis is the distance from the contour of a damaged region, where a negative value means inside the damaged region. As shown in the histogram, around the contour, i.e., distance $= 0$, gathers more gazes, which indicates the contour of the damaged region tends to be salient.

To be more specific, Fig. 3b shows the relationship between subjective score and gaze density within the vicinity of a damaged region's contour for two different inpainting methods. Green and red points are for the methods proposed by He et al. [12] and Huang et al. [15]. Here, we set the contour vicinity to be within 30 pixels from the contour. This corresponds to the 1.0 degree view angle we used in our experimental setup. As shown in Fig. 3b, the correlation coefficients r for the inpainted results provided by He et al. and Huang et al. are respectively $r = -0.63$ and $r = -0.47$. These results indicate a high negative correlation exists between subjective scores and gaze density around the contour. Thus there is a high probability that the image features around the contour are important for preference estimation.

4 Proposed method

Now we are ready to describe our proposed method. Our goal is to obtain the best result among inpainted images that are generated with different inpainting methods and parameters. In Section 4.1, we describe our pairwise ordering method and in Section 4.2 show how we automatically generated large auto-generated training set in which manual intervention was not required. Then, in Section 4.3 we show effective image features for this approach, which include the knowledge given in Section 3.

4.1 Ranking by assessing image quality with learning

Aiming at tackling the difficulty in reflecting subjective evaluations for inpainted results to scores, we based our preference order estimation on a learning-to-rank approach. Figure 4 shows the overview of our proposed method.

Fig. 4 Proposed method overview

Before we explain the details, let us briefly explain a typical pairwise learning-to-rank algorithm. This algorithm premises a ranking function $f(x)$, which computes the strength of the target attribute for each sample, as described below. Hereafter, we use x_i to denote a feature vector extracted from sample image z_i. The $f(x)$ is trained so that the ordering of the output value from the function $f(x)$ reflects the user annotated preference order $z_i \succ z_j$ between image pairs. In a word, the function f should satisfy the following formula:

$$z_i \succ z_j \iff f(x_i) > f(x_j) \tag{1}$$

We modeled f with the linear function $f(x) = \omega^\top x$. Then inequalities (1) can be written as below.

$$z_i \succ z_j \iff \omega^\top (x_i - x_j) > 0 \tag{2}$$

This mirrors the problem of binary classification. To implement the formulation that uses binary classification to calculate the preference order of pairs of images, the pair-wise learning to rank approach is widely used. From among the various methods yielding pair-wise learning to rank, we adopted RankingSVM [13] as it is used widely [1, 18, 33] due to its effectiveness and ease of implementation.

In our method, function f is trained with the pair of image feature vectors described in Section 4.3 with their preferences. We call a training data set of this type a "query" (see Fig. 4e). Basically, all preferences for generating queries are manually annotated by subjects. First, inpainted images with several parameters are generated with a masked image as shown in Fig. 4i–a. These results are used to make a pair of train images like those in Fig. 4i–b. Then, as shown in Fig. 4i–c, subjects were asked to provide their preference judgment (x) to a pair of inpainted images (z). Subjects' preferences are reflected to all image pairs (Fig. 4i–d).

However, the training data shortage problem still remains. To solve this problem, we additionally propose a way to automatically generate training data as shown in Fig. 4ii. In this case, auto-generated images (Fig. 4ii–a) are used instead of inpainted images. The

difference is that we already know that preferences depend on the degradation level. Thus we can skip to annotation; images with preferences (Fig. 4ii–d) are directly generated with image pairs (Fig. 4ii–b). In the next section we will describe how we designed a way to automatically generate a training set for which manual annotation is not required.

4.2 Auto-generated training data

Existing learning-based IQA methods for inpainting learn the relationship between image features and corresponding scores provided by subjects. Thus, they require user annotated samples. Here, we propose one effective solution, i.e., an IQA method for inpainted images by pairwise ordering. It is effective because it does not need any absolute scores but only pairwise relationships.

We add some distortions, such as proportional changes in pixel values or applying a low pass filter, to the original images that tend to occur as the result of inpainting. Several levels of such distorted images and original images generate training data with the assumption that increased distortion lessens preference. Of course, the original image has better quality than the distorted image. Because our method requires only pairwise relationships, not absolute scores, this simple relationship in which images become more distorted can work as a training data source. Figure 5 shows examples of several levels of auto-generated images for training. The i-th auto-generated train image I_i is synthesized by combining the original image I_{orig} and the i-th distorted image D_i as below.

$$I_i(x, y) = \begin{cases} D_i^{\gamma}(x, y) & ((x, y) \in \Omega) \\ I_{orig}(x, y) & (otherwise) \end{cases} \tag{3}$$

$$\gamma = \begin{cases} c & (brighter\ or\ darker\ color\ distortion) \\ b & (blur\ distortion) \end{cases} \tag{4}$$

where Ω is a masked region to reflect distortion (see Fig. 5a).

The reason we apply color and blur distortion is to simulate typical failures occurring as a result of inpainting. Human attention is considered to be quite sensitive to unnaturalness that is produced by differences in brightness and frequency components in images. To represent unnaturalness of this type, we focus on three types of distortion, i.e., that caused by

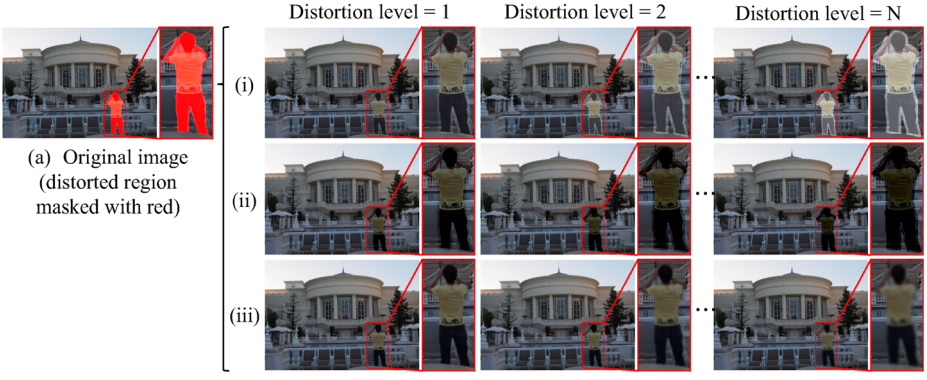

Fig. 5 Auto-generated images for training set. **a** Original image with region to be synthesized, which is masked in red. **b** Multi-levels of auto-generated images with (i) brighter color distortion, (ii) darker color distortion, (iii) blurred distortion

brighter colors, darker colors, and blurring. The former two distortions represent undesired inpainted results in which the colors of inpainted regions are brighter or darker color than those outside the contour of the inpainted region. The latter represents undesired inpainted results in which there is edge discontinuity around the inpainted region. Distorted images of these types are generated as follows.

Brighter and darker color distorted images

We define the i-th brighter/darker damaged image as below:

$$D_i^{(c)}(x, y) = I_{orig}(x, y) + \alpha\, \beta\, i \tag{5}$$

$$\beta = \begin{cases} +1 & (brighter\ color\ distortion) \\ -1 & (darker\ color\ distortion) \end{cases} \tag{6}$$

where α is a scalar parameter. In the work we report in this paper, we set $\alpha = 10$.

Blurred distorted image

We define the i-th blurred distorted image as below:

$$D_i^{(b)}(x, y) = \sum_{(k,l)} I_{orig}(x - k, y - l)\, G(k, l) \tag{7}$$

Here, G is a 2-dimensional Gaussian distribution with kernel size γ and we set $\gamma = 2i + 1$.

4.3 Features for learning-to-rank

Using the observation provided in Section 3, we designed image features for our framework. We call this image feature patch-based contour consistency ($PBCC$). As we described in Section 3, human perception is quite sensitive to color or edge discontinuities between in/out of damaged/distorted regions, which we combined to design the $PBCC$.

The $PBCC$ consists of the following two components: (1) differences between in/out of damaged/distorted region, and (2) normalized image around the contour. The former represents continuity across the contour of the damaged/distorted region. The latter represents the relative quality of images; the coherence of image quality between the inside and outside parts of a damaged region largely affects subjective quality. Thus, even if the image quality within the damaged region is the same, its perceptive quality varies depending on its surrounding region's quality.

To make the features dedicated for evaluating inpainted images, these components are computed along contours of the damaged/distorted region as shown in Fig. 6. We set the

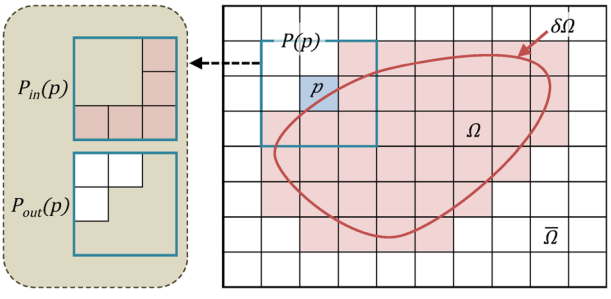

Fig. 6 Damaged/distorted region and its contour. $P_{in}(p)$ and $P_{out}(p)$ show masked or non masked regions in patch $P(p)$, which is centered at point p

features as $\mathbf{x} = (\mathbf{X_d}, \mathbf{X_s})$, where $\mathbf{X_d}$ and $\mathbf{X_s}$ respectively represent the first and the second components. $\mathbf{X_d}$ and $\mathbf{X_s}$ are computed as below;

$$X_d = ||S(P_{in}) - S(P_{out})||_2^2 \tag{8}$$

$$X_s = \frac{\sum_{p \in \delta\Omega} S(P_{out}(p))}{\sum_{p \in \delta\Omega} 1} \tag{9}$$

where Ω and $\delta\Omega$ respectively denote a damaged/distorted region and its contour. Equation (8) represents squared 2-norm. $P_{in}(p)$ and $P_{out}(p)$ show masked or non masked regions in patch $P(p)$, which is centered at point p (See Fig. 6). In addition, $S(P_{in}(p))$ and $S(P_{out}(p))$ represent average features of $P_{in}(p)$ and $P_{out}(p)$ as shown below.

$$S(P_{in}(p)) = \frac{\sum_{q \in P(p) \cap \Omega} \mathbf{s(q)}}{\sum_{q \in P(p) \cap \Omega} 1} \tag{10}$$

$$S(P_{out}(p)) = \frac{\sum_{q \in P(p) \cap \bar{\Omega}} \mathbf{s(q)}}{\sum_{q \in P(p) \cap \bar{\Omega}} 1} \tag{11}$$

To the extent of the work we report in this paper, we used $\mathbf{s(p)} = (u(p), v(p))$, where $\mathbf{u(p)} = (u_R(p), u_G(p), u_B(p))$ and $\mathbf{v(p)}$, each denoting RGB pixel values and edge strength.

5 Experiments

This section describes how we investigated the effectiveness of the proposed method. We will start by detailing the experimental setups used in Section 5.1. We will then verify the effectiveness our method by using the image features we derived and auto-generated training data as described in Sections 5.2 and 5.3, respectively. Section 5.4 then compares our method with the other existing IQA methods. For easy understanding, Fig. 7 shows the flow chart of the experiments conducted in this section.

Fig. 7 The flowchart for experiments conducted in Section 5. In Section 5.2, we investigated the efficacy of our proposed image representation. Section 5.3 verifies the effectiveness of auto-generated training data and Section 5.4 compares the estimation accuracy of our method and previous methods to assess their performance

5.1 Experimental setup

To generate training and test images, 111 images with manually masked damaged regions were prepared. The 111 images were inpainted with two inpainting methods [12, 15] and six parameters (3 patch sizes and 2 levels of multi-scale parameters) were used for each method. The quality of the images was evaluated by 24 subjects (12 males and 12 females) with normal vision. To make the users' judgment easy, we randomly displayed a pair of inpainted images side-by-side as shown in Fig. 8. Subjects were asked to choose one of three options: **r**: right image is better, **l**: left image is better, and **n**: no preference order (i.e., it is hard to decide which one is better or which one is worse). Excluding inpainted images with extremely poor quality, we obtained 2,466 image pairs. We excluded poor quality images because they might change subjects' judgement criteria during the experiment.

We implemented RankingSVM with SVM Rank [27] with Radial Basis Function (RBF) as the kernel function ($\gamma = 2^{-7}$), and the regularization parameter ($C = 2^{-5}$). We used a desktop PC (Intel Core i7, 3.4GHz CPU, 32GB memory) and used Matlab to implement the existing method, and Visual Studio 2012 to implement the proposed method.

5.2 Performance comparisons for different image features

This subsection verifies the effectiveness of $PBCC$, the proposed image feature described in Section 4.3. Table 1 compares the performances attained with seven different image features: F_{all} [11], GIST [22], EMD kernel [10, 29], $Saliency$, F_{in}, F_{out}, and $PBCC$. Note that in verifying performance with these image features we used the same training data and estimator; only the image features were different.

Here, we briefly introduce each of the compared features. F_{all} was originally used for learning-based-IQA of degraded images. GIST is one of the most commonly used global image features and is used for learning-based IQA for retargeted images. The EMD kernel is calculated on the basis of EMD and has been used in learning-based-IQA for inpainted images. F_{in} and F_{out} are the original features of this paper and represent the inside and outside parts of inpainted regions. The two methods represented by (12) and (13). were used for verifying the effectiveness of contour consistency on which $PBCC$ focuses. $Saliency$

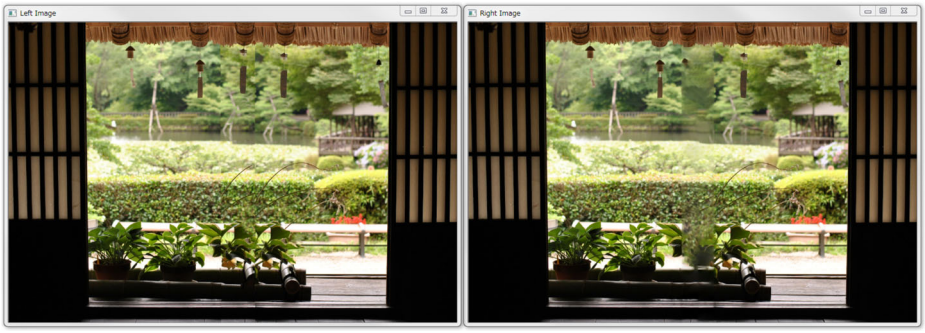

Fig. 8 Annotation interface for obtaining training data. Two different inpainted results are displayed side by side. Subjects annotate their preferences among three options: **r**: right image is better, **l**: left image is better, and **n**: no preference order

Table 1 Performance comparison for different image features [%]

Feature	Estimation accuracy
F_{all} [11]	44.47
GIST [22]	60.80
EMD kernel [10, 29]	53.27
$Saliency$ [30]	57.03
F_{in}	45.73
F_{out}	40.45
$PBCC$(Ours)	70.10

is for comparison of computational saliency maps. The same as $PBCC$, $Saliency$ is calculated so that it represents the inside and outside parts of inpainted regions and depends on how $PBCC$ is claculated. We used Walther et al.'s computational saliency map [30], as it is used by previous work [23].

$$F_{in} = \frac{\sum_{q \cap \Omega} \mathbf{s(q)}}{\sum_{q \cap \Omega} 1} \tag{12}$$

$$F_{out} = \frac{\sum_{q \cap \bar{\Omega}} \mathbf{s(q)}}{\sum_{q \cap \bar{\Omega}} 1} \tag{13}$$

Table 1 shows the estimation accuracy obtained, which is the ratio at which the preference order is correctly estimated among annotated pairs. As can be seen from the table, $PBCC$ correctly estimated the image pair preferences at 70.10%, as opposed to the 40.45% to 60.80% obtained with other methods. These results confirm the effectiveness of the proposed $PBCC$.

5.3 Verification of effectiveness depends on the amount of auto-generated training data

In this section we will show the results obtained in an investigation we conducted, which indicate that the system performance changes depending on the size of the auto-generated training set. First, we will show the results obtained with $N = 0$, in which no auto-generated images are included. Second, we will show how an auto-generated training set affects the performance. We used 100 original images to generate auto-generated image levels of $N =$ 1 to 10. Each level included the three types of distortion described in Section 4.2. N levels of auto-generated images make ${}_{N+1}C_2$ combinations of pair images, including comparisons with the original image. All of these ${}_{N+1}C_2$ pairs are used for training. We investigated the effect of the amount of data with $N = 1$ to 10. Figure 9 shows .a line graph in which the left y-axis shows estimation accuracy depending on N, which means the number of auto-generated images. Estimation accuracy without any auto-generated training set is annotated with a blue line for reference. The ratio of subjectively annotated training data to the data as a whole is shown by the orange line for the right y-axis. The training set in which N was 3 consists of 900 pairs of samples based on the 100 original images, with three types and three levels (six distortion combinations). In the same way, the training set in which N

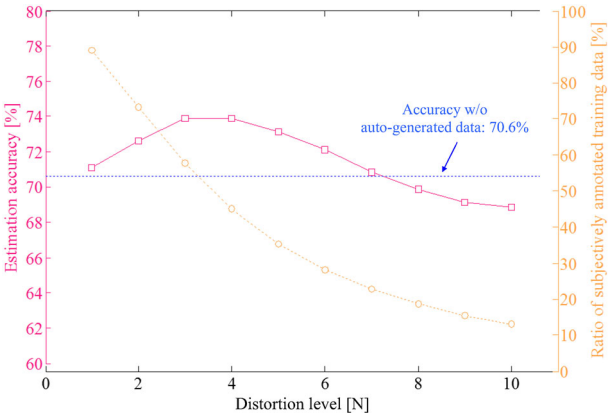

Fig. 9 Investigation of performance depending on the amount of auto-generated training set. Estimation accuracy depending on distortion level N is shown with magenta line with left y-axis. Accuracy without any auto-generated data is shown in blue line. Ratio of subjectively annotated training data to whole data is shown in orange bar graph with right y-axis

was 4 consists of 1800 pairs based on the 100 original images, with three types and four levels (10 distortion combinations). These results indicate that a larger training set is more effective; however, a set larger than certain levels of images results in worse performance.

One of the possible causes for this is that training sets with similar data may result in worse total prediction accuracy. If N is increased, a similar training set generated with the same original images will also increase. This additional data may decrease the effectiveness of learning. We also consider that another cause may be that the rate of subjective annotated data decreases as the amount of auto-generated data becomes larger. In the next section, we will show how we used two settings with different amounts of auto-generated data as our proposed methods: $N = 0$ with no auto-generated images, and $N = 3$, which was most effective for learning.

5.4 Comparison with existing methods

We conducted experiments in which we compared our method with other IQA methods for image inpainting, i.e., $ASVS$ and DN by Ardis et al. [3], $\overline{GD_{in}}$ by Venkatesh et al. [28],

Table 2 Prediction accuracy comparison with existing image quality assessment metrics [%]

Method	Estimation accuracy
ASVS [3]	58.04
DN [3]	60.22
$\overline{GD_{in}}$ [28]	55.88
BorSal [23]	50.46
StructBorSal [23]	51.55
Ours (w/o auto-generated data)	70.10
Ours (w/ auto-generated data)	73.87

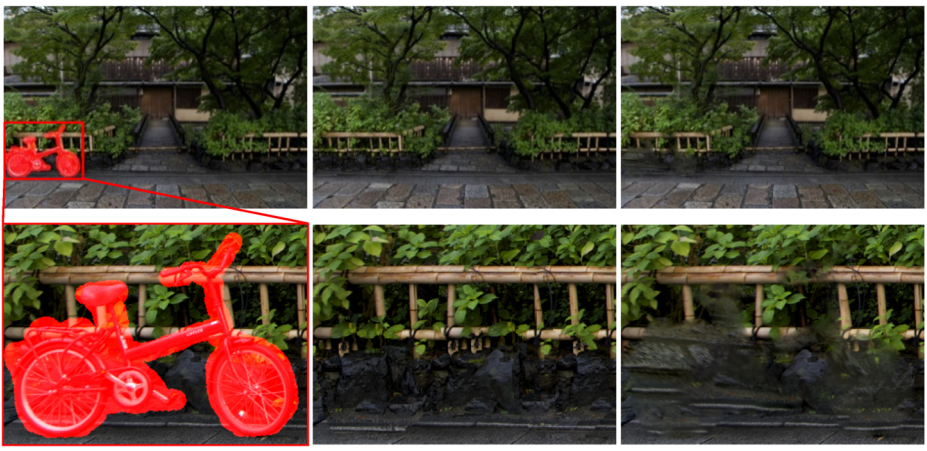

Fig. 10 Correctly ordered images with proposed method; **a** original image with damaged region masked in red while **b** and **c** are inpainted pairs of images that subjects annotated as (**b**) $\succ$ (**c**)

and $BorSal$, $StructBorSal$ by Oncu et al. [23]. Because we did not use an actual eye gazes for this experiment, we used a saliency map instead of human gaze for $\overline{GD_{in}}$; this is the same method that was used in the comparison experiment reported by Oncu et al. [23]. For our method we used $N = 0$ without an auto-generated training set and $N = 3$ with such a set; the results obtained were presented in Section 5.3.

Table 2 shows the prediction accuracy obtained for each metric. Our method without/with auto-generated training data correctly estimated the image pair preferences at 70.10% and 73.87% respectively, as opposed to the 50.46% to 60.22% obtained with other metrics. Thus,

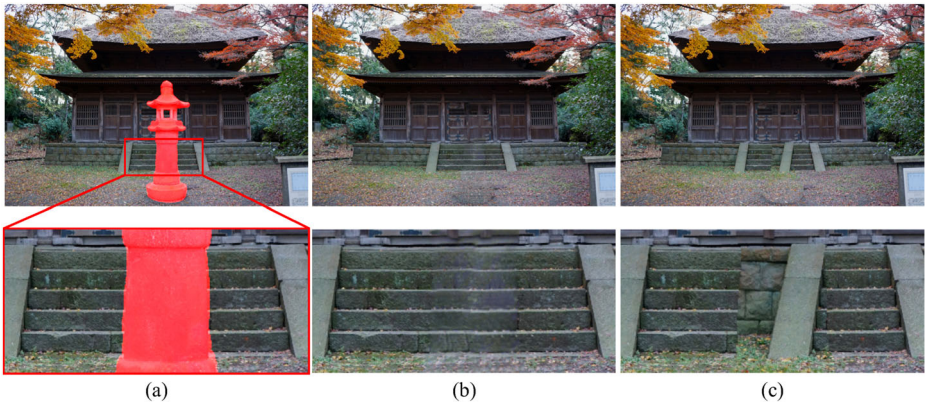

Fig. 11 Incorrectly ordered images with proposed method; **a** original image with damaged region masked in red while **b** and **c** are inpainted pairs of images that subjects annotated as (**b**) $\succ$ (**c**)

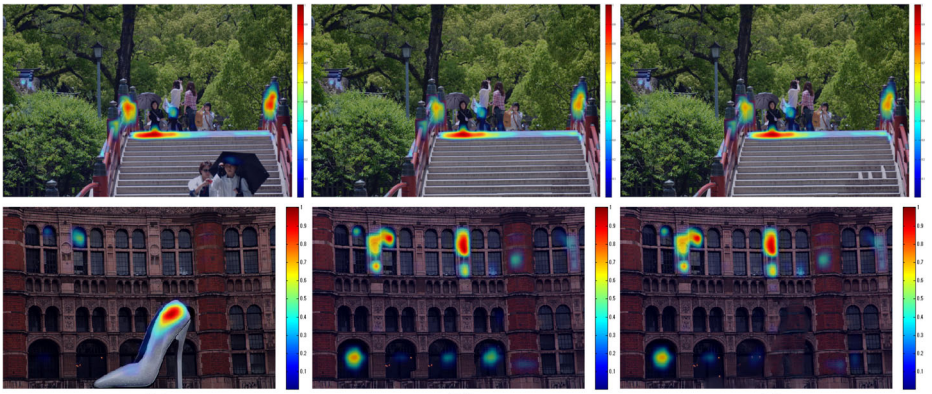

Fig. 12 To show the cause of the other existing methods' failure, a saliency map is overlaid on the left top and bottom images in Fig. 10. **a**–**c** are related to Fig. 10; original image and inpainted images. Upper images show that there are no significant differences between the two inpainted images. In the lower images, **b** gathered more gazes although subjects preferred (**b**)

the improvement our method achieved over existing methods was around 13 percentage points.

Figures 10 and 11 respectively show example outputs of correct and incorrect estimation obtained with our method. In both figures, (a) is an original image with a damaged region marked in red, and (b) and (c) are the inpainted results annotated by subjects as $(b) \succ (c)$. In Fig. 10, (b) images are successfully inpainted while example (c) was an inpainting failure in terms of both color consistency and structure. Our method successfully estimates preferences for such image pairs. We consider that because our image feature design focuses on the unnaturalness produced by color or structural discontinuity, it works well as the Fig. 10 results show. It is especially notable that none of the other existing metrics correctly estimated the preferences for the top left images in Table 2. Other than StructBorSal [23], these methods also failed to estimate preferences for the bottom left images.

We consider that these methods failed due to the uncertainty of the computed visual saliency maps. To demonstrate the cause of the previous method's failure, Fig. 12 shows a saliency map overlaid on the left half images of Fig. 10. In Fig. 12, (a) to (c) correspond to (a) to (c) in Fig. 10; the original image and inpainted pairs of images. All subjects answered that (b) was better than (c).

The upper (c) in Fig. 12 includes an inpainting failure around the stairs. However, neither (b) nor (c) gather saliency around the inpainted region. Also, the lower (c) in Fig. 12 has color and structural discontinuity, which generates huge unnaturalness. However (b) gathers more saliency on un-inpainted regions. We consider that this type of uncertainty and instability in saliency maps impede IQA quality.

Our method's limitation is shown in Fig. 11. In this figure, (b) has a blurred region and also has color discontinuity, but it is relatively natural in context. In (c), if we hide the left half of the picture, it is quite natural. However, structural unnaturalness occurs in the left half of the image and generates context unnaturalness. Because human perception is considered to be more sensitive for contextual failures, subjects preferred (b). Currently our method does not consider any semantic information, thus for such image pairs it generates ordering failures.

6 Conclusion

In this paper we described an image quality assessment method we developed for image inpainting. Three key ideas of our method are that (1) we use a ranking-by-learning algorithm to estimate the ordering of inpainted images on the basis of subjective quality, (2) our ranking system easily introduces auto-generated training data for more effective learning, and (3) we introducing image features that reflect differences around a contour of damaged regions on the basis of gaze measuring experiments which showed that a high negative correlation exists between subjective quality and gaze density around the contour. Unlike existing image quality assessment (IQA) methods for image inpainting, ours makes it possible to introduce auto-generated training data, due to introduction of our pairwise learning. Preference order estimation experiment results suggest the method's efficacy. Especially with auto-generated training sets, the estimation performance was about 13 percentage points higher than that of existing IQA methods. In future work, we will introduce other image features such as describing semantic unnaturalness inside inpainted images.

References

1. Abe T, Okatani T, Deguchi K (2012) Recognizing surface qualities from natural images based on learning to rank. In: International conference on pattern recognition (ICPR), pp 3712–3715
2. Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency-tuned salient region detection. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1597–1604
3. Ardis PaulA., Singhal Amit (2009) Visual salience metrics for image inpainting. In: Proc SPIE, vol 7257, pp 72571W–72571W–9
4. Barnes C, Shechtman E, Finkelstein A, Goldman DB (2009) PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Trans Graph (TOG) 28(3):24:1–24:11
5. Bertalmio M, Vese L, Sapiro G, Osher S (2003) Simultaneous structure and texture image inpainting. IEEE Trans Image Process 12(8):882–889
6. Chang K-Y, Chen C-S (2015) A learning framework for age rank estimation based on face images with scattering transform. IEEE Trans Image Process 24(3):785–798
7. Criminisi A, Perez P, Toyama K (2004) Region filling and object removal by exemplar-based inpainting. IEEE Trans Image Process 13(9):1200–1212
8. Darabi S, Shechtman E, Barnes C, Goldman DB, Sen P (2012) Image melding combining inconsistent images using patch-based synthesis. ACM Trans Graph (TOG) (Proceedings of SIGGRAPH 2012) 31(4):82:1–82:10
9. Deng J, Dong W, Socher R, Li jL, Li K, Li F-f (2009) Imagenet a large-scale hierarchical image database. In: The IEEE conference on computer vision and pattern recognition (CVPR), pp 248–255
10. Frantc VA, Voronin VV, Marchuk VI, Sherstobitov AI, Agaian S, Egiazarian K (2014) Machine learning approach for objective inpainting quality assessment. In: Proc SPIE, vol 9120, pp 91200S–91200S–9
11. Gao F, Tao D, Gao X, Li X (2015) Learning to rank for blind image quality assessment. IEEE Transactions on Neural Networks and Learning Systems 26(10):2275–2290
12. He K, Sun J (2014) Image completion approaches using the statistics of similar patches. IEEE Trans Pattern Anal Mach Intell 36(12):2423–2435
13. Herbrich R, Graepel T, Obermayer K (2000) Large margin rank boundaries for ordinal regression. In: Advances in large margin classifiers. The MIT Press, pp 115–132

14. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 1–8
15. Huang J-B, Kang SB, Ahuja N, Kopf J (2014) Image completion using planar structure guidance. ACM Trans Graph (TOG) 33(4):129:1–129:10
16. Isogawa M, Mikami D, Takahashi K, Kojima A (2016) Eye gaze analysis and learning-to-rank to obtain the most preferred result in image inpainting. In: IEEE International conference on image processing (ICIP), pp 3538–3542
17. Isogawa M, Mikami D, Takahashi K, Kojima A (2016) Image and video completion via feature reduction and compensation. Multimedia Tools and Applications 76(7):9443–9462
18. Khosla A, Xiao J, Torralba A, Oliva A (2012) Memorability of image regions. In: Advances in neural information processing systems (NIPS), pp 296–304
19. Long X, Li J, Lin W, Zhang Y, Zhang Y, Yan Y (2016) Pairwise comparison and rank learning for image quality assessment. Displays 44:21–26
20. Long X, Li J, Lin W, Zhang, Ma L, Fang Y, Yan Y (2016) Multi-task rank learning for image quality assessment. IEEE Trans Circuits Syst Video Technol PP(99):1–1
21. Ma L, Long X, Zhang Y, Yan Y, Ngan KN (2016) No-reference retargeted image quality assessment based on pairwise rank learning. IEEE Trans Multimedia 18(11):2228–2237
22. Oliva A, Torralba A (2001) Modeling the shape of the scene A holistic representation of the spatial envelope. Int J Comput Vis 42(3):145–175
23. Oncu A, Deger F, Hardeberg J (2012) Evaluation of digital inpainting quality in the context of artwork restoration. In: European conference on computer vision (ECCV) workshops and demonstrations, vol 7583, pp 561–570
24. Pishchulin L, Jain A, Andriluka M, Thormählen T, Schiele B (2012) Articulated people detection and pose estimation: reshaping the future. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 3178–3185
25. Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM (2016) The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: The IEEE conference on computer vision and pattern recognition (CVPR), pp 3234–3243
26. Thanh Trung A, Beghdadi B, Larabi C (2013) Chaker Azeddine DANG Perceptual quality assessment for color image inpainting. In: IEEE International conference on image processing (ICIP), pp 398–402
27. Tsochantaridis I, Joachims T, Hofmann T, Altun Y (2005) Large margin methods for structured and interdependent output variables. J Mach Learn Res 6:1453–1484
28. Vijay Venkatesh M, Cheung S-cS (2010) Eye tracking based perceptual image inpainting quality analysis. In: IEEE international conference on image processing (ICIP), pp 1109–1112
29. Voronin VV, Frantc VA, Marchuk VI, Sherstobitov AI, Egiazarian K (2015) No-reference visual quality assessment for image inpainting. In: Proc SPIE, volume 9399, pp 93990U–93990U–8
30. Walther D, Koch C (2006) Modeling attention to salient proto-objects. Neural Netw 19(9):1395–1407
31. Wang S, Li H, Zhu X, Li P (2008) An evaluation index based on parameter weight for image inpainting quality. In: International conference for young computer scientists (ICYCS), pp 786–790
32. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
33. Yan J, Lin S, Kang SB, Tang X (2014) A learning-to-rank approach for image color enhancement. In: IEEE Conference on computer vision and pattern recognition (CVPR), pp 2987–2994
34. Zongben X, Sun J (2010) Image inpainting by patch propagation using patch sparsity. IEEE Trans Image Process 19(5):1153–1165

Mariko Isogawa received her B.S. and M.S. degrees from Osaka University, Japan, in 2011 and 2013, respectively. She joined NTT Media Intelligence Laboratories in 2013 and has been working as a researcher. She is also working toward the PhD degree from the Graduate School of Engineering Science, Osaka University. Her research interests include computer vision and human-perception aware image/video processing.

Dan Mikami received his B.E and M.E degree from Keio University, Kanagawa, Japan in 2000 and 2002, respectively. He has been working for Nippon Telegraph and Telephone Corporation from 2002. He received his Ph.D. from Tsukuba University in 2012. He is currently a Senior Research Engineer of Cross Modal Computing Project, NTT Media Intelligence Laboratories, and a visiting associate professor of the graduate school of Keio University. His research activities are mainly focused on cross modal information handling for athletes. He was awarded the Meeting on Image Recognition and Understanding 2009 Excellent Paper Award 2009, the IEICE Best Papar Award 2010, the IEICE KIYASU-Zen'iti Award 2010, and the IPSJ SIG-CDS Excellent Paper Award 2013. He is a member of IEICE, IPSJ, and IEEE.

Kosuke Takahashi received his B.Sc. degree in engineering, M.Sc. and Ph.D. in informatics from Kyoto University, Japan, in 2010, 2012 and 2018 respectively. He joined NTT Media Intelligence Laboratories in 2012 and has been working as a researcher. His research interest includes computer vision and its applications. He received "Best Open Source Code" award Second Prize in CVPR 2012. He is a member of IPSJ.

Hideaki Kimata received the B.E. and M.E. degrees in applied physics in 1993 and 1995 and Ph.D. degree in electrical engineering in 2006 respectively from Nagoya University, Nagoya, Japan. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1995 and has been engaged in R&D of video processing of coding, realistic communication, computer vision, and recognition based on machine learning (deep learning). He is currently a Senior Research Engineer, Supervisor, in NTT Media Intelligence Laboratories. He is a member of the Institute of Electronics, Information and Communication Engineers of Japan (IEICE), and a chief examiner of special interest group on audio visual and multimedia information processing of Information Processing Society of Japan (IPSJ).