International Journal of Computer Vision
https://doi.org/10.1007/s11263-018-1132-0

Which is the Better Inpainted Image? Training Data Generation Without Any Manual Operations

Mariko Isogawa[1,2] · Dan Mikami[1,3] · Kosuke Takahashi[1] · Daisuke Iwai[2] · Kosuke Sato[2] · Hideaki Kimata[1]

Received: 27 February 2018 / Accepted: 13 November 2018

Abstract
This paper proposes a learning-based quality evaluation framework for inpainted results that does not require any subjectively annotated training data. Image inpainting, which removes and restores unwanted regions in images, is widely acknowledged as a task whose results are quite difficult to evaluate objectively. Thus, existing learning-based image quality assessment (IQA) methods for inpainting require subjectively annotated data for training. However, subjective annotation requires huge cost and subjects' judgment occasionally differs from person to person in accordance with the judgment criteria. To overcome these difficulties, the proposed framework generates and uses simulated failure results of inpainted images whose subjective qualities are controlled as the training data. We also propose a masking method for generating training data towards fully automated training data generation. These approaches make it possible to successfully estimate better inpainted images, even though the task is quite subjective. To demonstrate the effectiveness of our approach, we test our algorithm with various datasets and show it outperforms existing IQA methods for inpainting.

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

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 for automatically removing such areas ("damaged regions" in this paper) and restoring them (Criminisi et al. 2004; He and Sun 2014; Huang et al. 2014; Isogawa et al. 2017b; Bertalmio et al. 2003; Barnes et al. 2009; Darabi et al. 2012; Xu and Sun 2010; Yu et al. 2018). Although many effective algorithms have been proposed, it is known that inpainting results vary largely with the method used and the parameters set. In a typical use case, users iteratively repeat parameter tuning and result observation until desired results are obtained. Since this is time-consuming and requires special knowledge, a way to automatically select the best results is needed.

To achieve such automatic selection, we have identified two main issues. The first is that evaluating "correctness" of inpainted results is a task that requires subjective judgment. The second is that even with human judgment, providing absolute scores in a stable manner for inpainted images is a difficult task. Due to these issues, no definitive way has previously been found to estimate subjective quality on the basis of objectively measurable features.

Figure 1 explains these two issues with examples. In the figure, Fig. 1c–e show inpainted results obtained with different parameters. These depend on the original image and the original image with the damaged regions respectively shown in Fig. 1a, b. The former issue is explained by the results seen in Fig. 1c, d. Although neither of these results are different from the original one shown in Fig. 1a, both of them are perceptually natural. Results such as these are considered to be "correct" as long as they are perceptually natural for humans,

Communicated by Tae-Kyun Kim, Stefanos Zafeiriou, Ben Glocker and Stefan Leutenegger.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11263-018-1132-0) contains supplementary material, which is available to authorized users.

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

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

[1] NTT Media Intelligence Laboratories, Yokosuka, Japan

[2] Graduate School of Engineering Science, Osaka University, Toyonaka, Japan

[3] NTT Communication Science Laboratories, Atsugi, Japan

Published online: 26 November 2018

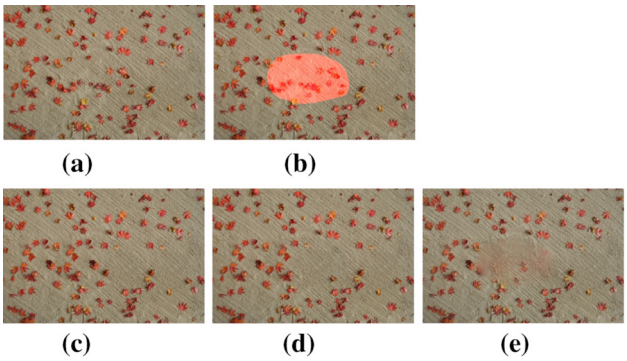

Fig. 1 An example that explains difficulty of evaluating inpainted images objectively. **a** Original image. **b** Original image with damaged region. **c** Inpainted image 1. **d** Inpainted image 2. **e** Inpainted image 3

even if they differ from the original one (He and Sun 2014; Huang et al. 2014; Isogawa et al. 2017b). The difficulty in defining this kind of "correctness" means human judgment must be relied on in many cases.

The latter issue is explained by the results seen in Fig. 1d, e. Although differences in quality can be seen between these images, it is quite difficult to stably give absolute scores to them because personal judgment plays such an important role in giving them. Because of these two difficulties, estimating inpainting quality has long been considered a challenging task.

To address the former issue, existing image quality assessment (IQA) methods have tried to find a way to represent subjective quality by means of objectively measurable indicators. Venkatesh and Cheung used observed gaze density inside and outside the damaged region in inpainted images (Venkatesh and Cheung 2010). Instead of observed gaze, many IQA methods using a computational visual saliency map, which simulates human gaze density, have also been proposed (Ardis and Singhal 2009; Oncu et al. 2012; Trung et al. 2013; Frantc et al. 2014; Voronin et al. 2015). However, actual human gazes vary by viewer and viewing context and their correspondence with saliency maps is quite limited. Isogawa et al. (2016) revealed that the pixel-wise unnaturalness that occurs in inpainted images is not suitable for saliency based methods because the resolution of visual saliency maps is coarse. They also proposed perceptually-aware image features focusing on the border area of mask regions where human gazes tend to gather.

To address the latter issue, let us look once more at the Fig. 1d, e results. In this case, although it is not easy to give stable scores, it is comparatively easy to choose which results are better. Current IQA methods mainly focus on providing absolute scores despite the difficulties involved in doing so Frantc et al. (2014), Voronin et al. (2015). They include support vector regression (SVR) based methods and require absolute scores for learning, which is difficult and tends to become unstable. Unlike these methods, that proposed by Isogawa et al. (2016) involves an ordering approach, which estimates the preference order of inpainted images. Therefore, to use the method for learning purposes it is only necessary to ascertain the preference order, which is a comparatively easy task. The method uses a learning-to-rank approach and accurately estimates the subjective quality of inpainted images by dividing problems into a set of pairwise preference order estimation tasks. Learning based methods such as these commonly require a subjective annotation step before the training step, which is considered essential. This labor-intensive annotation leads to both huge annotation cost and, what is worse, fluctuation of evaluation criteria.

Consequently, this paper proposes a new framework for estimating learning-to-rank based preference order with automatically generated training data. Such data is referred to as "auto-generated" data in the paper. The method simulates "failed" inpainting and assumes that a simulated sample has worse subjective quality than the method's best inpainted image. Thus, it generates training pairs automatically without any user intervention.

This paper is based on conference proceedings we previously published (Isogawa et al. 2017a) but also includes descriptions of novel techniques to further reduce intervention, including masking that leads to fully labor cost-free ranking function learning. We also describe an investigation we conducted to determine how well the method works for test images with unknown inpainting methods.

Contribution The main contribution we show in this paper is that the proposed method achieves learning based preference estimation of inpainted images without annotated training data. To the best of our knowledge, this is the first study that tackles learning based estimation of subjective attributes without manually annotated training data. The other contributions include a way to generate degraded inpainted results and a way to generate masked regions as means to fully generate automated training data.

The rest of this paper is organized as follows. In Sect. 2 we review related work. Section 3 describes the learning based ranking method we propose, which is trained with auto-generated data. Section 4 proposes a way to generate masked regions with the aim of generating fully automated training data. Section 5 verifies the effectiveness of the proposed method and Sect. 6 concludes the paper with a summary of key points and describes the subjects for future work.

2 Related Work

This section introduces related work. First, Sect. 2.1 overviews existing IQA methods for image inpainting. In Sect. 2.2 we focus on machine learning, especially as a means to prepare training data. Studies on automatic generation or augmentation of training data are introduced.

2.1 Learning Based IQA Methods for Inpainted Image

With the aim of selecting the best one from a plurality of results among varied inpainted images, many IQA methods have been proposed (Ardis and Singhal 2009; Oncu et al. 2012; Trung et al. 2013; Frantc et al. 2014; Voronin et al. 2015; Isogawa et al. 2016). Among these methods, learning based approaches have demonstrated effective performance (Frantc et al. 2014; Voronin et al. 2015; Isogawa et al. 2016). Frantc et al. (2014) and Voronin et al. (2015) proposed SVR based IQA methods. These approaches estimate an absolute subjective score for each test image. For training regression models, subjectively annotated rating scores are essential. Thus, they used data annotated by subjects who were asked to provide scores on a 5-point scale.

Another learning based approach to tackle this problem is the learning-to-rank approach. It learns and estimates rank order on the basis of a trained ranking function. The important advantage of this approach is that it can learn only on the basis of rank order. Because of this advantage, this approach has been the focus of considerable attention, especially when it is applied to tasks where it is difficult to estimate subjective preference objectively Chang and Chen (2015), Yan et al. (2014), Abe et al. (2012), Khosla et al. (2012). Isogawa et al. (2016) proposed a learning-to-rank based IQA method by pairwise preference estimation. This method focuses on the premise that the preference order, rather than absolute scores, is good for selecting the best one from a plurality of results, which is the method's primary goal. For training data, the method requires image pairs with annotated preference order.

As described above, one difficulty that commonly exists in learning based methods is the need for a labor-intensive annotation step for obtaining training data. These manual annotations require huge annotation cost. In addition, the judgment criteria of subjects fluctuate occasionally. To overcome these problems, the proposed method enables automatic generation of training data. It generates pairwise training data automatically and applies a learning-to-rank based algorithm to the preference order estimation.

2.2 Learning with Auto-generated Training Set

Larger amounts of training data generally lead to higher performance in learning-based methods. Therefore, in recent years some studies have improved learning accuracy by augmenting the learning data with automatic generation (Pishchulin et al. 2012; Ros et al. 2016). Pishchulin et al. (2012) proposed a human detection and pose estimation by using automatically generated training sets. The main advantage of their method is that it enables human poses and shapes to be controlled explicitly on the basis of existing training sets. They also combine various background images to increase training data. Ros et al. (2016) proposed learning based pixel-wise semantic segmentation that uses automatically generated training data. Since annotation data for this task must be provided on a pixel-by-pixel basis, having humans provide the data is labor-intensive. To overcome the problem, they use realistic synthetic images of urban views in a virtual world that can provide annotation data on a pixel-by-pixel basis.

In these tasks, it is apparent that an image generation model can be obtained from annotation data. Thus, generation of training examples, i.e., a set of annotated data and the generated images, is rather easy. In contrast, modeling the relationship between inpainted images and their annotated subjective quality is quite difficult. The reason is shown in Fig. 1, where both inpainted images have subjectively good quality. This makes it difficult to create the auto-generated training data. To the best of our knowledge, this study is the first trial of making auto-generated training data for subjectively assessing inpainted image quality.

3 Proposed Method

3.1 Overview

Figure 2 shows the overview of the proposed preference ordering framework with automatic generation of training data. Our framework consists of training and estimation phases.

In the training phase, the proposed method first generates a simulated training set. Then, a ranking model is trained with these auto-generated images. As the ranking model, we utilize the pairwise learning-to-rank based method previously used by Isogawa et al. (2016). This is because unlike SVR-based methods (Frantc et al. 2014; Voronin et al. 2015), a pairwise method only requires a set of preference orders between two images and does not require absolute scores of subjective quality. The learning process is detailed in Sect. 3.2. Because the proposed method uses pairwise learning-to-rank, the proposed generation of training data yields inpainted image pairs with known preference orders. This is described in more detail in Sect. 3.3.

The estimation phase procedure is fairly clear. With inpainted image pair input, the method extracts feature vectors that focus on the unnaturalness produced by color or structural discontinuity around the inpainted region contours. Then, the ranking function's scalar output values are calculated. The magnitude relationship between pairs of images shows their preference order.

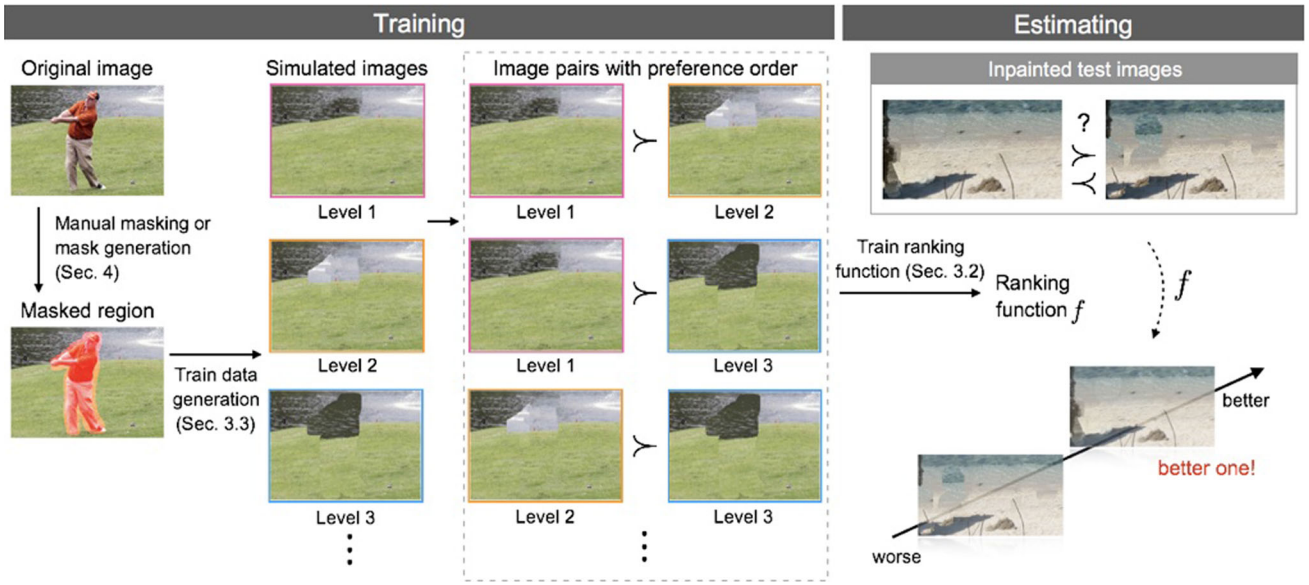

Fig. 2 Overview of our proposed learning framework

3.2 Model Learning with Auto-generated Training Data

This subsection describes a learning-to-rank based learning algorithm for ordering pairwise images. The method is trained with auto-generated simulation images that represent degraded inpainted results.

This algorithm premises a ranking function $f(x^i)$ that projects feature vector x^i, which is obtained by $x^i = g(I^i)$ from image I^i, to a one-dimensional axis in accordance with the subjective quality of inpainted results, where $g(\cdot)$ is a feature extraction function. For simplicity, we use "$I^i \succ I^j$" to express that "I^i is preferred to I^j".

For easy understanding, before we describe the ranking algorithm trained with auto-generated training data, let us briefly explain the training algorithm for $f(x)$. We define the function $h(x^i, x^j)$ that denotes preference order as follows.

$$h(x^i, x^j) = \begin{cases} +1 \ (I^i \succ I^j) \\ \ \ 0 \ \ \text{(no preferences)} \\ -1 \ (I^j \succ I^i), \end{cases} \tag{1}$$

The $f(x)$ is trained so that the difference of outputs $f(x^i) - f(x^j)$ has the same sign as $h(x^i, x^j)$. In a word, the function f should satisfy the following formula:

$$\text{sign}(h(x^i, x^j)) = \text{sign}\,(f(x^i) - f(x^j)). \tag{2}$$

The goal is to learn f, which is concordant with the training samples. We modeled f with the linear function $f(x) = \omega^\top x$. Then Eq. 2 can be rewritten as

$$\text{sign}\,(h(x^i, x^j)) = \text{sign}\,(\omega^\top (x^i - x^j)). \tag{3}$$

The error function is defined on the basis of Eq. 3 and is optimized with respect to ω. This is the same problem as that of binary classification. We use a pairwise learning-to-rank algorithm called RankingSVM (Herbrich et al. 2000) to solve it.

Now we are ready to introduce auto-generated training set into the ranking algorithm. Let I^l_{sim} be a simulated inpainted image with l degraded level, with which larger l indicates more degradation. In accordance with degraded level, the preference order among such images is

$$l_1 < l_2 \rightarrow I^{l_1}_{sim} \succ I^{l_2}_{sim}. \tag{4}$$

Such auto-generated images are used to train the ranking function. The way to generate degraded images I_{sim} is described in the next subsection.

3.3 Automatic Training Data Generation

The proposed method relies on existing inpainting methods and devices for them to obtain degraded inpainted images that well simulate inpainting failures. Section 3.3.1 briefly reviews typical inpainting algorithms and then Sect. 3.3.2 describes how degraded data are generated on the basis of the existing inpainting algorithms.

3.3.1 Patch Based Image Inpainting Algorithms

Among various inpainting methods, patch-based algorithms are widely acknowledged as promising approaches. Typi-

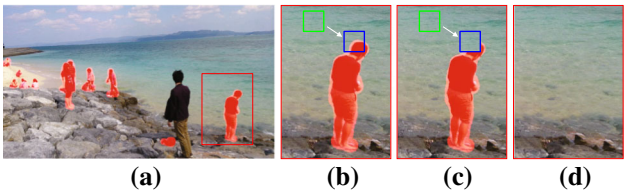

Fig. 3 Typical patch-based approach for image inpainting

cally they comprise three steps, which we will explain by using Fig. 3. For the damaged region masked with red in Fig. 3a, (1) a patch that includes both the source and the damaged region is set as a target, the blue rectangle in Fig. 3b, (2) a similar patch for the target patch, the green rectangle in Fig. 3b, is retrieved in the source region, and (3) the damaged region in the target patch is replaced in accordance with similar patches as in Fig. 3c. The resultant restored image is shown in Fig. 3d.

For proposing our method in subsequent part, we will introduce the following notation for the above patch-based inpainting concept. Here, $P(p)$ is a target patch whose center pixel is p, and $\widehat{P}_p^N$ denotes the Nth most similar patch from $P(p)$. Since total inpainting quality highly depends on the quality of retrieved patches, it is basically acknowledged that the more similar the retrieved patch is, the better the inpainted quality becomes. That is, for "fine" inpainted results, the most similar patch (i.e., with $N = 1$) from $P(p)$, which is denoted as $\widehat{P}_p^1$, is used as in Eq. 5 to restore a missing region.

$$\widehat{P}_p^1 = P(q') = \arg\min_{P(q)} \ dist(P(p), P(q)) \tag{5}$$

$dist(\cdot)$ represents distance function. The proposed method uses the assumption in an inverse way, i.e., dissimilar patches generate unnatural inpainted images.

3.3.2 Auto-generated Inpainted Images as a Training Set

In simulating failed inpainted images, we found that if we selected the Nth most similar patch $\widehat{P}^N$ having larger N, it would apparently correspond to the cases in which good patches for inpainting cannot be found. This is a typical case of inpainting failure. Therefore, as the value N gets larger, the patches become dissimilar and the inpainting results get worse. That is, simulated inpainted images are generated so that their relationships depend on the level of patch similarity as $I_{sim}^{l_1} \succ I_{sim}^{l_2}$ when $l_1 < l_2$, where I_{sim}^N represents a simulated image inpainted with Nth similar patches. We propose incorporating this patch retrieval into existing inpainting algorithms. The following shows our simulated image generation with two types of algorithms as examples.

With Patch-Retrieval Based Method Here we explain our data simulation for patch retrieval based inpainting methods

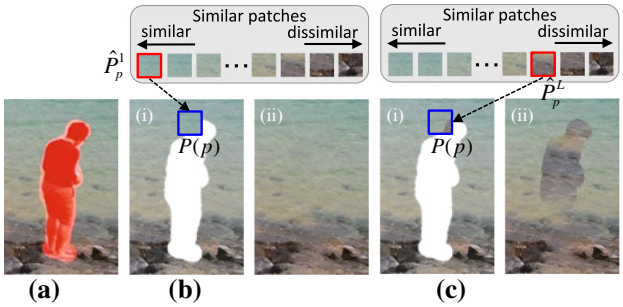

Fig. 4 Patch based degraded inpainted image generation. **a** Original image with masked region. **b** Method's best. **c** Lth degraded

with Criminisi et al.'s method (2004) as a base algorithm. This simulation can also be applied to other patch-based algorithms. The original method uses simple patch retrieval as shown in Eq. 5. Our method can be easily incorporated into this patch retrieval; instead of retrieving the most similar patches $\widehat{P}_p^1$, we obtain $\widehat{P}_p^L$ with $L > 1$. Figure 4 illustrates this in more detail. The original image with damaged region is shown in (a). The resultant inpainted images are shown in (b) and (c). Here, (b) is the method's best result and (c) is a simulated deteriorated result obtained using our proposed method. As the typical procedure of patch-based inpainting, target patch $P(p)$ is determined in (i), and then a similar patch is retrieved and used for filling in the hole. In case (b), the most similar patch is used. Unlike this, the proposed method uses a dissimilar patch depending on degraded level L and obtains the degraded result as shown in (c-ii).

With Image-Retrieval Based Method Some current studies extend the basic algorithms by using patch retrieval indirectly. Our training data simulation method can also be applied to such methods without loss of generality. Here we explain an extension using He and Sun's method (He and Sun 2014) as the base algorithm. He and Sun improve the basic algorithm on the basis of two ideas; extension of the patch P to the whole image I, and treating an inpainting task as a Photomontage problem (Agarwala et al. 2004). Figure 5 illustrates this in more detail. The original image with damaged region and the target region to be inpainted are shown in Fig. 5a, b. The resultant inpainted images are shown in Fig. 5c, d. Figure 5c is the method's best result and Fig. 5d is a simulated deteriorated result obtained using our proposed method. Let $\widehat{I}^N$ be the Nth most similar image for damaged image I. To generate the method's best inpainting result, they retrieve the K most similar images $\widehat{I}_K = \{\widehat{I}^1, \widehat{I}^2, \ldots, \widehat{I}^K\}$ and the missing region is filled by combining a stack of these images. Our method modifies this image retrieval part; instead of retrieving the K most similar images, we obtain $\widehat{I}_{\alpha L}$ that excludes $\widehat{I}_{\alpha(L-1)}$. That is, we obtain $\widehat{I}_{\alpha L} - \widehat{I}_{\alpha(L-1)} = \{\widehat{I}^{\alpha(L-1)+1}, \widehat{I}^{\alpha(L-1)+2}, \ldots, \widehat{I}^{\alpha L}\}$ to generate an Lth level of a degraded image.

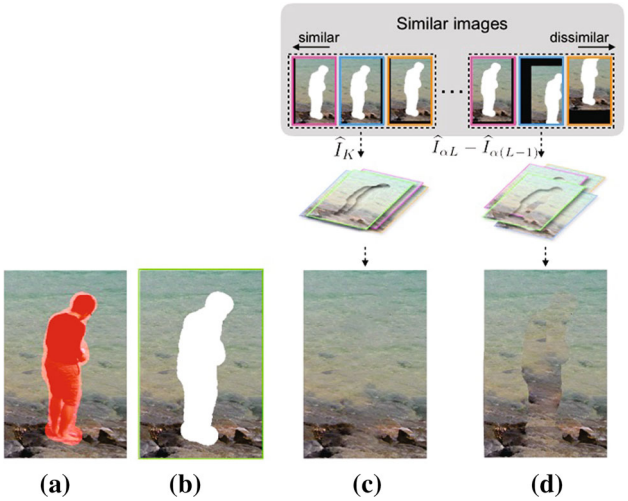

Fig. 5 Image retrieval based degraded inpainted image generation. **a** Original image with masked region. **b** Target image to be filled in. **c** Method's best. **d** *L*th degraded

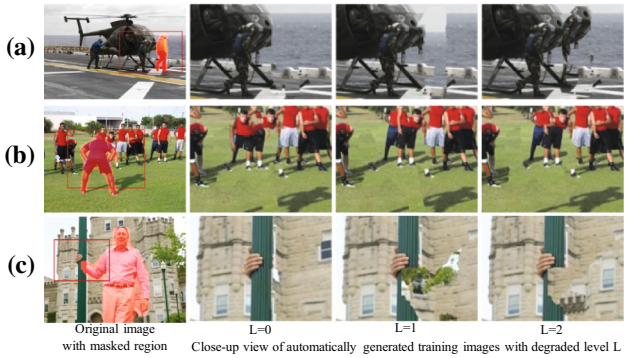

Fig. 6 Example of auto-generated data quality depending on masked region; **a** desirable multi-degraded auto-generated data, **b**, **c** undesirable auto-generated data with masked region that do not satisfy the requirements

4 Towards Completely Automatic Training Data Generation: Masked Region Generation

The previous section focuses only on the labor cost involved in annotating preference orders. Since training data generation requires huge annotation cost, our approach with auto-generated data can significantly reduce the cost. However, there is still a process that must be done manually in this method. That is to designate the region to be inpainted as a masked region. This operation requires manual intervention and hinders larger training set generation.

To improve our method, we also propose a method to generate masked regions for effective auto-generated data by utilizing semantic segmentation. By additionally using this automatic designation of mask regions, the method eliminates any manual work needed to generate training data. In other words, it makes it possible to increase the amount of training data with no labor cost.

However, it is known that the quality of inpainted results varies largely depending on their masked region. Since our training data generation method assumes that simulated image pairs with multi-levels of degradation have orders in terms of quality, the method's best inpainted images with no degradation should have good enough quality with its masked region. In case the method's best inpainting results with degradation level $L = 0$ do not have good enough quality, the quality of all degraded image is biased toward poor directions and thus it might break our assumption.

Thus, we consider that desirable masked regions for effective auto-generated data should satisfy the following three requirements; (1) the images for training data include varied scenes for high versatility training sets, (2) other objects are not adjacent to the contours of the masked regions, and (3) the region of the object to be inpainted does not protrude from the masked region.

Figure 6 explains how requirements (2) and (3) affect auto-generated data. In the figure, (a) shows the masked region that satisfies the requirement and its auto-generated data with degradation level 0 to 2. Multi-levels of degraded images become worse if degradation level L increases, as expected. Fig. 6b, c respectively show failure cases in which requirements (2) or (3) are not satisfied. In Fig. 6b another object touches a masked human region, and in Fig. 6c the people's region to be inpainted is revealed from the masked region. In both cases, unnatural inpainted results are generated in all degradation stages and there are no definitive preferences between them. This is due to the fact that the designated masked region makes it difficult to find the source region to be used for filling the hole. Restoring missing objects such as adjacent humans or revealed human regions is rather difficult.

To satisfy the requirements described above, we use people's regions in images found by semantic segmentation. Humans appear in a lot of images and they are less likely to be adjacent to other objects compared to other objects such as desks in a classroom. Therefore, using such regions makes it possible to meet the first and second requirements. In order to meet the second point better, we use regions that do not have any adjacency with other objects. To satisfy the third requirement, we dilate the extracted region since the human region extracted by semantic segmentation is often smaller than that of an actual human region.

The detailed process of masked region generation is described below (see Fig. 7). With original image I_{orig} obtained from a dataset, the Internet, etc (see Fig. 7a), semantic segmentation results are calculated as shown in Fig. 7b. We use Mask R-CNN (He et al. 2017) as a semantic segmentation method. If the detected people's region has no overlap with any other object and its size is 1–20% of I_{orig}, the initial masked image I^{init}_{mask} is generated with the segmented region (see Fig. 7c). For I_{orig} with multiple people's

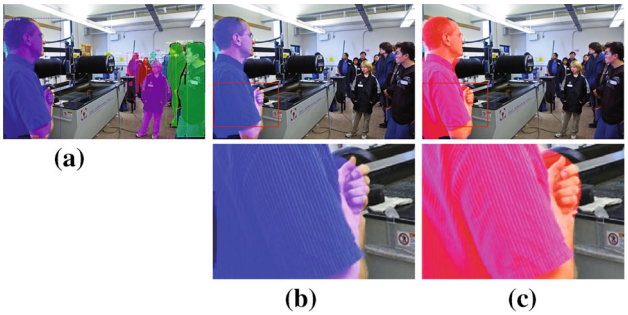

Fig. 7 Proposed masked region generation. The method first detects people's regions as shown in **a** and dilates initial masked region **b**–**c** so that the region satisfies the mask requirements

regions, only the region having the largest area is adopted. The final masked image I_{mask} shown in Fig. 7d is calculated by dilating the masked region in I_{mask}^{init}.

5 Experiment

This section reports the efficacy of the proposed method. In Sect. 5.1 we will show the experimental setup we used, including training data preparation. In Sect. 5.2 we will show the efficacy of our auto-generated data as a training set. In Sect. 5.3 we will demonstrate the efficacy of the proposed masked region generation.

5.1 Experimental Setup

5.1.1 Ranking Learning

The proposed method uses Isogawa et al.'s rank learning framework (Isogawa et al. 2016). The characteristics and advantage of the framework are as follows. The framework uses RankingSVM, implemented using SVM Rank (Tsochantaridis et al. 2005) with a radial basis function (RBF) kernel, whose parameters are well tuned. For training and testing, it uses ten-dimensional image features that focus on the unnaturalness around the contours of the inpainted regions produced by color or structural discontinuity. Since the features are normalized for the size of the masked region contour, the rank learning is relatively robust for the shape complexity of the masked region or non-uniformity of the texture (see Fig. 9 for image pairs, the preference orders of which were correctly estimated).

5.1.2 Preparing Manually Annotated Data

Manually annotated data were basically used for test data with ground-truth preference annotation. They were also used as training data for comparing the estimation accuracy obtained with an auto-generation data based model to that obtained with a manual annotation data based one.

Subjective Annotation We prepared 100 publicly available images obtained from the Web. Damaged regions in these images were manually masked. For each masked image, we generated a fixed number of inpainted results with different parameters. The number differed depending on the experiments we conducted. The experiments discussed in Sects. 5.2.1 and 5.2.2 and the first experiment discussed in Sect. 5.2.3 involved generating six inpainted results by a combination of three options of patch size and two options of number of similar images to be retrieved. The second experiment discussed in Sect. 5.2.3 involved preparing three inpainted results by changing the pre-trained model for inpainting. The third experiment discussed in Sect. 5.2.3 also involved preparing three inpainted results by changing the inpainting method.

The quality of the images was evaluated by eight subjects (four males and four females) with normal vision. To make the users' judgment easy, we randomly displayed a pair of inpainted images side-by-side. Subjects were asked to choose one of three options: right image is better, left image is better, and no preference order (i.e., it is hard to decide which one is better or which one is worse). As inpainting methods, we basically used He and Sun's method (He and Sun 2014) throughout this section, but in Sect. 5.2.3 we add Huang et al.'s method (2014), Herling et al.'s method (2014), and Yu et al.'s method (2018).

Notation and Reliability of Annotated Data We used $T_a^{(M,S)}$ to denote annotated datasets with inpainted method M, for which a consensus was obtained for at least S subjects. For example, $T_a^{(He,8)}$ indicates the dataset with He et al.'s inpainting method, which got a unanimous answer from all eight subjects. With the dataset $T_a^{(He,5)}$ consensus was obtained for from five to eight subjects. Thus, more than half of the subjects gave the same preference order to pairs in $T_a^{(He,5)}$. This indicates that S reflects the difficulty humans have in making judgments in such cases.

5.1.3 Auto-generated Training Data

This subsection describes how we got an auto-generated training set.

Degraded Image Simulation We gathered the same images of annotated data. But, please note that we excluded auto-generated data simulated with identical images as test data. Damaged regions in these images were manually masked. Since the position, size, and shape of the damaged region are normalized during the learning-to-rank process, we were able to set the damaged regions arbitrarily regardless of the objects in the target images. We set five degradation levels for simulating inpainted images I_{sim}^{L}, i.e., $L = 0, 1, 2, 3$, and 4, where $L = 0$ indicates an image without any intentional

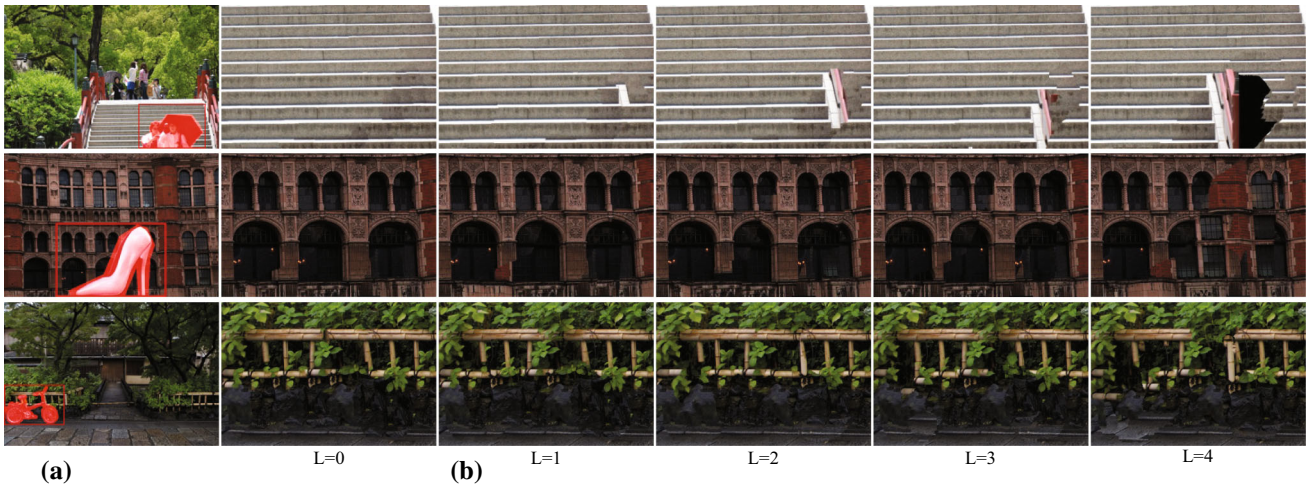

Fig. 8 Simulated inpainted images. **a** Original image with masked region. **b** Close-up view of automatically generated training images with degraded level L

degradation, i.e., the method's best inpainted image. These five image levels are generated from one original image. By combining these five images, we generated ${}_5C_2 = 10$ pairs of training data, with preference orders $I^x_{sim} \succ I^y_{sim}$ $(\forall x < y)$, i.e., the inverse of degradation level.

Figure 8b shows degraded images depending on the degraded level L ($L = 0, 1, 2, 3,$ and 4). All degraded images are inpainted with the masked region shown in (a). Figure 8 shows that our method simulates degraded images well; each degraded image gets worse quality as L increases. Though the deterioration is subjective, it well simulates the failures that typically occur in ordinary inpainting methods having inappropriate parameters such as patch size.

Notation We denote the auto-generated degraded images with inpainting method M for training T^M_d. For example, auto-generated data with He et al.'s inpainting method is denoted as T^{He}_d.

5.2 Investigation to Ascertain Effectiveness of Auto-generated Training Data

5.2.1 Comparison with Existing IQA Methods

We conducted experiments comparing our method to other IQA methods for image inpainting, i.e., $ASVS$ and DN by Ardis and Singhal (2009), $\overline{GD}_{in}$ by Venkatesh and Cheung (2010), $BorSal$, $StructBorSal$ by Oncu et al. (2012) as non learning-based methods, and Isogawa et al.'s method (2016) as a learning-based method. We also verified RankIQA (Liu et al. 2017), the rank-learning-based IQA method with a deep neural network (DNN). Although the method is not for IQA of inpainting, we argue that the comparison with the DNN-based IQA method is informative. Note that although the $\overline{GD}_{in}$ originally uses measured human gaze, we used a saliency map instead. This is the same evaluation approach used in Oncu et al. (2012). For training with Isogawa et al.'s method we used the annotation data of $T^{(He,8)}_a$. Our proposed learning method trained with auto-generated data T^{He}_d is denoted as $Ours(T^{He}_d)$.

Table 1 shows the prediction accuracy for all test data $T^{(He,S)}_a$ $(5 \leq S \leq 8)$ obtained for each metric. Excluding inpainted images with extremely poor quality, the amounts of test data $|T^S_a|$ of T^S_a with $S = 5$ to 8 were $(|T^5_a|, |T^6_a|, |T^7_a|, |T^8_a|) = (184, 136, 71, 38)$. Our method $Ours(T^{He}_d)$ correctly estimated the preference order within image pair with the highest score for all test data; the improvement our method achieved over Isogawa et al.'s method was 6.52, 8.09, 4.23, and 2.63 points for test data $T^{(He,5)}_a$, $T^{(He,6)}_a$, $T^{(He,7)}_a$, and $T^{(He,8)}_a$.

Figure 9 shows examples of image pairs, the preference orders of which were correctly estimated even with non-uniformity texture (see (a)) and shape complexity of masked region (see (b)). Please refer our supplemental material for more results.

5.2.2 Verifying Effectiveness of Auto-generated Training Data Depending on Varied Conditions

This subsection describes three more investigations we conducted to verify the validity of the auto-generated data. The first one investigates the effects of the volume of auto-generated data on estimation accuracy, which is the ratio of estimation success of preference order among annotated pairs. Figure 10 shows the estimation accuracy obtained when the amount of auto-generated data is increased from 50 to 990 in 50 increments. Training data are randomly selected from T^{He}_d. As shown in the graphs in the figure, estimation accuracy increased as the amount of training data increased. In addition, to investigate whether auto-generated data can be used as a substitute for annotated

Table 1 Prediction accuracy comparison with existing image quality assessment metrics (%)

	$T_a^{(He,5)}$	$T_a^{(He,6)}$	$T_a^{(He,7)}$	$T_a^{(He,8)}$
ASVS (Ardis and Singhal 2009)	45.11	44.85	43.66	44.74
DN (Ardis and Singhal 2009)	53.26	53.68	56.34	57.89
$\overline{GD}_{in}$ (Venkatesh and Cheung 2010)	43.48	44.85	40.85	39.47
BorSal (Oncu et al. 2012)	42.39	43.38	42.25	44.74
StructBorSal (Oncu et al. 2012)	46.74	45.59	42.25	52.63
RankIQA (Liu et al. 2017)	65.79	60.53	63.16	42.11
Isogawa et al. (2016)	60.33	62.5	71.83	76.32
Ours(T_d^{He})	<u>66.85</u>	<u>70.59</u>	<u>76.06</u>	<u>78.95</u>
Ours($T_d^{He} + T_a^{He}$)	65.22	68.38	<u>76.06</u>	<u>78.95</u>

The highest scores are underlined

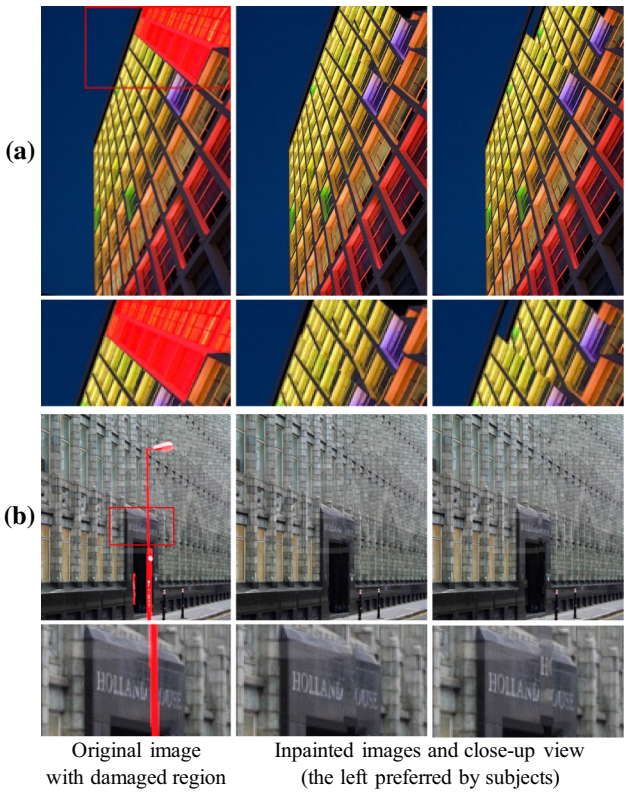

Fig. 9 Inpainted image pairs, the preference orders of which were correctly estimated with our model even with **a** non-uniformity background textures, and **b** complex shape of masked region

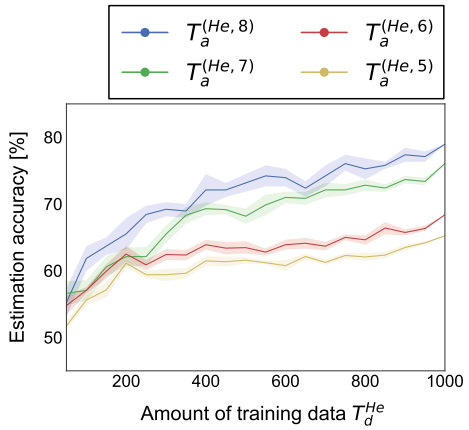

Fig. 10 Prediction accuracy with each T_a^{He} depending on the amount of T_d^{He}

data, the performances depending on the proportion of auto-generated data were tested. Figure 11 shows performances depending on the proportion of auto-generated data with $T_a^{(He,5)}$. Here, the number of training data was fixed to 990 in all cases; only the proportion of annotated and auto-generated data was changed. The amount of annotated data was decreased from 180 to 0 in 10 decrements. Even though the amount of subjectively annotated training data was changed, the estimation accuracies were almost constant in all cases. These results suggested that auto-generated training data could be a substitute for manually annotated data.

We also investigated how the amount of auto-generated data affected estimation accuracy. Figure 12 shows estimation accuracy for each T_a^S when the auto-generated training data levels L were changed between $L = 1$ to 4. The available training data amounts were respectively 99, 297, 594, and 990 for $L = 1, 2, 3$, and 4. However, to focus on the affect of data levels, we set it to the smallest number 99, i.e., that for $L = 1$. Note that the results are average performances with 10 trial runs and that training data for each trial are randomly selected. As the figure shows, the estimation accuracy increases as L increases, which suggests that multi-levels of auto-generated data work effectively. These two kinds of investigations suggest that the auto-generated data works as expected and using multi-levels of data works effectively.

The third one verifies the performance when subjectively annotated data is added to auto-generated data. We added $T_a^{(He,S)}$ to T_d^{He} for further verification of auto-generated data performance. Hereafter, in this section we denote these two data as T_a^S and T_d for simplify the explanation. We denote

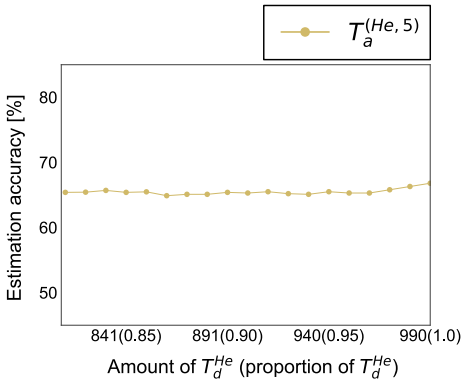

Fig. 11 Prediction accuracy depending on the proportion of T_d^{He}

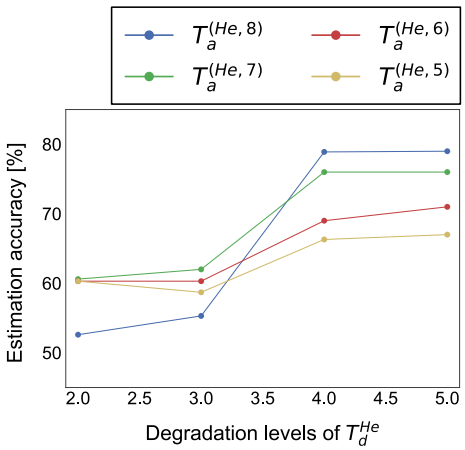

Fig. 12 Prediction accuracy depending on the levels of T_d^{He}

Table 2 Prediction accuracy with or without subjectively annotated data (%)

	T_a^5	T_a^6	T_a^7	T_a^8
Ours(T_d)	66.85	70.59	76.06	78.95
Ours($T_d + T_a$)	65.22	68.38	76.06	78.95

The higher scores are underlined

our learning method with such data as $Ours(T_d + T_a)$. The comparison between $Ours(T_d)$ and $Ours(T_d + T_a)$ with all test data T_a^S with $S = 5, 6, 7$, and 8 is shown in Table 2.

With this table, we found that the use of annotated training data does not show significant changes on prediction accuracies of for all cases (T_a^5, T_a^6, T_a^7, and T_a^8). However, in case of low consensus data such as $S = 5$ and 6, the use of annotated training data deteriorated the prediction accuracy. "Low consensus" means the subjective judgement varies by the subject and may not suit for machine learning. To verify this consideration, we conducted the next experiment.

We divide the auto-generated data, T_d into two groups; reliable data set consisted by auto-generated data of $L = 1$ and 4, T_d^{re}; unreliable data set consisted by that of $L = 2$ and 3, T_d^{un}. We subjectively confirmed that T_d^{un} have small difference in subjective quality and are difficult to be judged by a large margin compare to T_d^{re}, as in Fig. 8. The prediction accuracy is shown in Table 3. For all test data, $Ours(T_d^{re})$ excels $Ours(T_d^{un})$. It also suggests that subjectively similar data like T_d^{un} is not a good data for training. Thus, we should consider the balance between the number of auto-generated data and the quality of them as a future work.

Table 3 Prediction accuracy with or without unreliable data (%)

	T_a^5	T_a^6	T_a^7	T_a^8
Ours(T_d^{re})	66.30	69.85	76.06	81.58
Ours(T_d^{un})	63.04	67.65	73.24	78.95

The higher scores are underlined

5.2.3 Effectiveness for Test Image with Unknown Inpainting Method

Up to the preceding sections, we have examined the effect of auto-generated training data under the condition that the inpainting method used for test and training data generation is same. However, since inpainted results vary depending on their inpainting method, auto-generated training data are also varied by their inpainting methods. Thus, this section investigates how ranking function trained with a certain method works for the test data generated by another method. Hereafter, we define "pre-trained" and "re-trained" ranking functions as that are trained without/with auto-generated training data inpainted with same method as test data. We also denote test data whose inpainting method is not used for pre-training as "unknown test data" or "test data with unknown method", and its inpainting method as "unknown inpainting method".

With Patch-Based Inpainting Methods For investigation, we used Huang et al.'s (2014) and Herling et al.'s (2014) methods as unknown patch-based inpainting methods. Since these two are patch-retrieval based methods, our patch-retrieval based simulation method is applied for auto-generated data preparation. Such auto-generated data with these two methods were not used for pre-training and we used He et al.'s method (2014) for pre-training same as previous section. From here we denote auto-generated training data with these three inpainting methods as T_d^{He}, T_d^{Sc}, and T_d^{Pix}. Samely, we denote test data with new two methods as T_a^{Sc} and T_a^{Pix}. As a ground truth, preference orders for T_a^{Sc} and T_a^{Pix} are annotated by 8 subjects. T_a^{Sc}, T_a^{Pix} with at least S subject's consensus is denoted as $T_a^{(Sc,S)}$ and $T_a^{(Pix,S)}$. The amounts of test data of $T_a^{(Sc,S)}$ and $T_a^{(Pix,S)}$ with $S = 5$ to 8 were (321, 199, 108, 47) and (712, 602, 461, 317), respectively. Same as previous section, 100 original images for auto-generated training data were prepared for five levels of degraded images. Each of T_d^{He}, T_d^{Sc}, T_d^{Pix} includes ${}_5C_2 \times 100 = 1000$ pairs of training data.

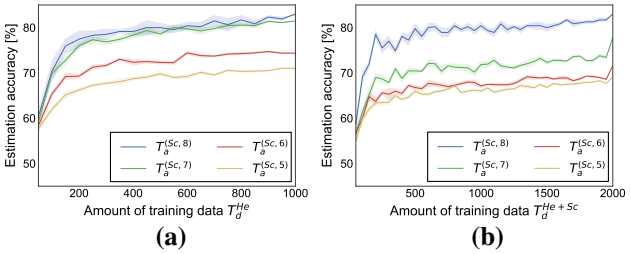

Fig. 13 Prediction accuracy for T_a^{Sc} with **a** pre-trained model with T_d^{He} and **b** re-trained model with T_d^{He+Sc}

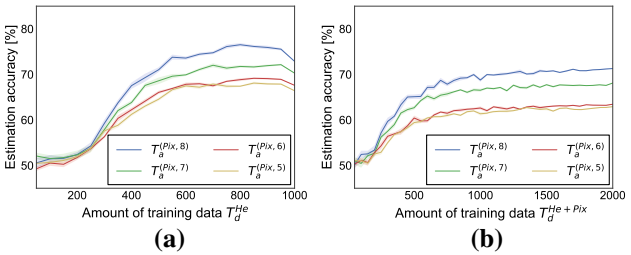

Fig. 14 Prediction accuracy for T_a^{Pix} with **a** pre-trained model with T_d^{He} and **b** re-trained model with T_d^{He+Pix}

Table 4 Prediction accuracy with or without auto-generated data inpainted with unknown Huang et al.'s method (%)

	$T_a^{(Sc,5)}$	$T_a^{(Sc,6)}$	$T_a^{(Sc,7)}$	$T_a^{(Sc,8)}$
Ours(T_d^{He})	71.03	74.37	81.48	82.98
Ours(T_d^{He+Sc})	69.16	71.86	77.78	82.98

Table 5 Prediction accuracy with or without auto-generated data inpainted with unknown Herling et al.'s method (%)

	$T_a^{(Pix,5)}$	$T_a^{(Pix,6)}$	$T_a^{(Pix,7)}$	$T_a^{(Pix,8)}$
Ours(T_d^{He})	66.41	67.61	70.28	72.87
Ours(T_d^{He+Pix})	62.92	63.46	68.11	71.29

Test data with unknown inpainted method T_a^{Sc}, T_a^{Pix} are evaluated under two conditions; one is with pre-trained ranking model with T_d^{He}, and the other is re-trained model with same inpainting method as test data, i.e., mixture training data includes T_d^{He} and T_d^{Sc}, or T_d^{He} and T_d^{Pix}. We call such adjacent data as "mixture training data" and denote them as T_d^{He+Sc} and T_d^{He+Pix}, respectively. These mixture training data have twice samples of each training data, i.e., 2000 pairs.

Figures 13 and 14 show estimation accuracy for each T_a^S where $S = 5, 6, 7$, and 8 with pre-trained ranking function with T_d^{He} (see (a)) and re-trained it with mixture training data T_d^{He+Sc} or T_d^{He+Pix} (see (b)). In the figures, vertical axis shows estimation accuracy and horizontal axis shows amount of training data, which is auto-generated. That are increased from 100 to 1000 in 100 increments for (a) and from 100 to 2000 in 100 increments for (b). The training data are randomly selected and the plotted accuracy is an average of 10 runs with these standard deviations. Tables 4 and 5 show the estimation accuracy value we obtained for each T_a^S, where $S = 5, 6, 7$, and 8 with the pre-trained ranking function with T_d^{He} with 1000 data elements and re-trained it with mixture training set T_d^{He+Sc} or T_d^{He+Pix} with 2000 data.

As the results show, it was possible to estimate preference orders with high accuracy, i.e., 82.97% for T_a^{Sc} and 72.87% for T_a^{Pix} even for these unknown test data. In addition, regarding result with T_d^{He+Sc}, although estimation accuracy was slightly decreased with mixture training data, significant differences could not be observed. These results show that our method has a certain amount of robustness against the patch-based inpainting method used.

With GAN-Based Inpainting Methods So far, we have used specific types of inpainting methods, i.e., image retrieval based and patch based. However, in addition to these effective conventional methods, generative adversarial network (GAN) based inpainting methods have achieved remarkable progress in recent years. We verified how our IQA method trained with auto-generated data works for GAN-based inpainting methods that may have different types of degradation.

We used Yu et al.'s (2018) GAN-based inpainting method as the unknown method. The method was trained with the three database, i.e., places2 (Zhou et al. 2018), CelebA (Liu et al. 2015), and imageNet (Deng et al. 2009). As in Sects. 5.2.1 and 5.2.2, we used He et al.'s method (2014) for training our preference-ordering model. The denotation for auto-generated data is the same as previous one; T_d^{He}. Test data with Yu et al.'s GAN-based inpainting method are denoted as T_a^{GAN}, and T_a^{GAN} with the consensus of at least S participants are denoted as $T_a^{(GAN,S)}$. The amount of $T_a^{(GAN,S)}$ with $S = 5$ to 8 was (212, 204, 151, 139). As discussed in the previous section, 100 original images for auto-generated training data were prepared for five levels of degraded images; thus, T_d^{He} includes ${}_5C_2 \times 100 = 1000$ pairs of training data.

Figure 15a shows the estimation accuracy for each $T_a^{(GAN,S)}$, where $S = 5, 6, 7$, and 8 with the pre-trained ranking function with T_d^{He}. The vertical axis shows estimation

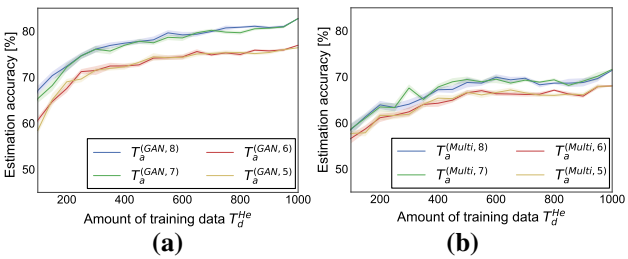

Fig. 15 Prediction accuracy for **a** T_a^{GAN} and **b** T_a^{Multi}. T_d^{He} was used as trained model

Table 6 Prediction accuracy for test sets generated with unknown inpainted methods, i.e., Yu et al.'s GAN-based inpainting method (%)

	$T_a^{(GAN,5)}$	$T_a^{(GAN,6)}$	$T_a^{(GAN,7)}$	$T_a^{(GAN,8)}$
Ours(T_d^{He})	76.42	76.96	82.78	82.73

Table 7 Prediction accuracy for test sets generated with multiple unknown inpainted methods, i.e., Yu et al.'s GAN-based, He et al.'s image retrieval based, and Huang et al.'s patch based methods (%)

	$T_a^{(Multi,5)}$	$T_a^{(Multi,6)}$	$T_a^{(Multi,7)}$	$T_a^{(Multi,8)}$
Ours(T_d^{He})	68.02	68.04	71.63	71.43

accuracy and the horizontal axis shows the amount of training data, which were auto-generated. The amount of training data were increased from 100 to 1000 in 100 increments. The training data were randomly selected, and the plotted accuracy is an average of 10 runs with these standard deviations. Table 6 shows the estimation accuracy score for each $T_a^{(GAN,S)}$, where $S = 5, 6, 7$, and 8 with the pre-trained ranking function with T_d^{He}. Even with the GAN-based inpainted images, it was possible to estimate preference orders with high accuracy, i.e., 82.73%. The results also indicate that our method can handle test images with different types of inpainting methods with a certain amount of robustness.

With Multiple Types of Inpainting Methods for Test Images So far, we have focused on one inpainting algorithm for one experiment to generate test images. That is, our ranking algorithm estimates preference orders of images inpainted *with the same algorithms.* However, there may be situations in which users want to find the orders between images inpainted *with different algorithms.* To generate test images, this experiment uses three different types of inpainted algorithms, i.e., patch based (Huang et al. 2014), image retrieval based (He and Sun 2014), and GAN based (Yu et al. 2018), trained with the places2 dataset (Zhou et al. 2018). As in the previous experiments, we used He et al.'s method (2014) for auto-generated training data. The auto-generated training and test sets consisting of images with the three different inpainting methods are respectively denoted as T_d^{He} and T_a^{Multi}. The T_a^{Multi} with the consensus of at least S participants is denoted as $T_a^{(Multi,S)}$. The amount of test data of $T_a^{(Multi,S)}$ with $S = 5$ to 8 was (197, 194, 141, 140). Training data consisted of ${}_5C_2 \times 100 = 1000$ pairs, as in the previous experiments.

Figure 15b shows the estimation accuracy for each $T_a^{(Multi,S)}$ where $S = 5, 6, 7$, and 8. The vertical axis shows estimation accuracy and the horizontal axis shows the amount of training data, which were auto-generated. That are increased from 100 to 1000 in 100 increments. The training data were randomly selected, and the plotted accuracy is an average of 10 runs with these standard deviations. Table 7 shows the estimation accuracy score for each $T_a^{(Multi,S)}$, where $S = 5, 6, 7$, and 8. Despite the task's difficulty in estimating preference orders between images generated using different inpainting methods, our proposed method estimated the orders with 71.43%. Note that the test images contained inpainted results with two unknown and fundamentally different algorithms from that for training-data generation. The results also indicate that our method can work well even for a test data set consisting of images inpainted using different types of methods.

5.3 Effectiveness Investigation for Masked Region Generation

This section investigates the effectiveness of proposed masked region creation towards completely human labor-free training data generation. Note that our main proposal is a training-data-generation method with the degraded inpainting introduced thus far. It is not restricted to fully automatic masking, which is introduced in this section.

Experimental Setup For experiments, we additionally gathered new images from ImageNet (Deng et al. 2009). The mask type, i.e., the masked images of the proposed method and comparison targets, are as follows.

1. The proposed method is denoted by I_{mask}^{people}, which uses automatically segmented people regions without contacting other objects (proposed in Sect. 4).
2. The first method for comparison is denoted by I_{mask}^{rect}, which is masked by rectangle of 200×200 (pixels) centered at the original image.
3. The second method for comparison is denoted by $I_{mask}^{people(ad)}$, which is similar to the proposed method but includes images that are adjacent to other objects.
4. The third method for comparison is denoted by I_{mask}^{all}, which is also similar to the proposed method. Although the proposed method only used people's region, it uses all kinds of objects that are automatically tagged.

Training Data Generation Inpainted images of five levels of degradation with use of He et al.'s method were generated for each mask type. These training data that corresponds to I_{mask}^{people}, I_{mask}^{rect}, $I_{mask}^{people(ad)}$, and I_{mask}^{all} are denoted by T_d^{people}, T_d^{rect}, $T_d^{people(ad)}$, and T_d^{all} respectively. The amount of original images used for auto-generated data were 900 for each type of mask and thus each auto-generated training set consisted of ${}_5C_2 \times 900 = 9000$ pairs of training data. Examples of training data T_d^{people} with degraded levels L are shown in Fig. 16. As these examples show, degraded inpainting images as training data are appropriately generated. Exam-

Fig. 16 Auto-generated training data with proposed masked region generation; **a** generated I_{mask}^{people}, **b** close-up view of multi levels of simulated inpainted images

ples of other training set, i.e., I_{mask}^{rect}, $I_{mask}^{people(ad)}$, and I_{mask}^{all}, are shown in Fig. 17.

Results Figure 18 shows estimation accuracy for each T_a^S with He et al.'s method where $S = 5, 6, 7$, and 8 with each dataset T_d^{people}, T_d^{rect}, $T_d^{people(ad)}$, and T_d^{all}. In all of these graphs, vertical axis shows estimation accuracy and horizontal axis shows amount of data used for training, that are increased from 100 to 9000 in 100 increments. The training data are randomly selected and the plotted accuracy is an average of 10 runs with standard deviation. As reference, black dotted lines indicate estimation accuracy with auto-generated data with manually designated masked region, i.e., T_d^{He}, which is considered to be the most effective training data proposed in Sect. 3.3. Table 8 compares estimation accuracy for each T_a^S with four types of training sets.

As shown in the graphs in Fig. 18 and Table 8, estimation accuracy with T_d^{people} was 79.47%. Although this training set requires no human labor for both masked region designation nor training data generation, the estimation accuracy was rather higher than that of T_d^{He} described in Sect. 5.2.1, which requires annotated masked regions. Regarding other types of masked regions, i.e., T_d^{rect}, $T_d^{people(ad)}$, and T_d^{all}, the estimation accuracies were 59.74 to 74.47, which were far below the accuracy with T_d^{people}. These results support our assumption that designation of the masked region is important to generate effective auto-generated training data.

Figure 17 explains possible reasons for the accuracy with T_d^{rect}, $T_d^{people(ad)}$, and T_d^{all} did not reach that of T_d^{people} with failure examples of auto-generated data. The failure case in (b) shows an example where the masked people's region is adjacent to another people's region, and the failure case shown in (c) shows an example where the object's region to be inpainted is revealed from the mask. In either case, since the masked regions are not appropriate for inpainting, auto-generated data results with degradation level $L = 0$ have lower quality, even though ideally they should not include any degradation. Therefore, we consider that the effectiveness of training data might be reduced because the quality of each level of simulated data was biased towards the worse quality direction and it was difficult to get definitive preference relationships between them.

In addition, although auto-generated training set T_d^{people} with proposed masked region showed higher accuracy than T_d^{He}, which requires annotated masked regions, we consider that there is still room for improvement. Our masked region generation strategy avoids adjacency with other objects in images and protruding object regions. However, we found that it was not always satisfied. Figure 19 shows examples of

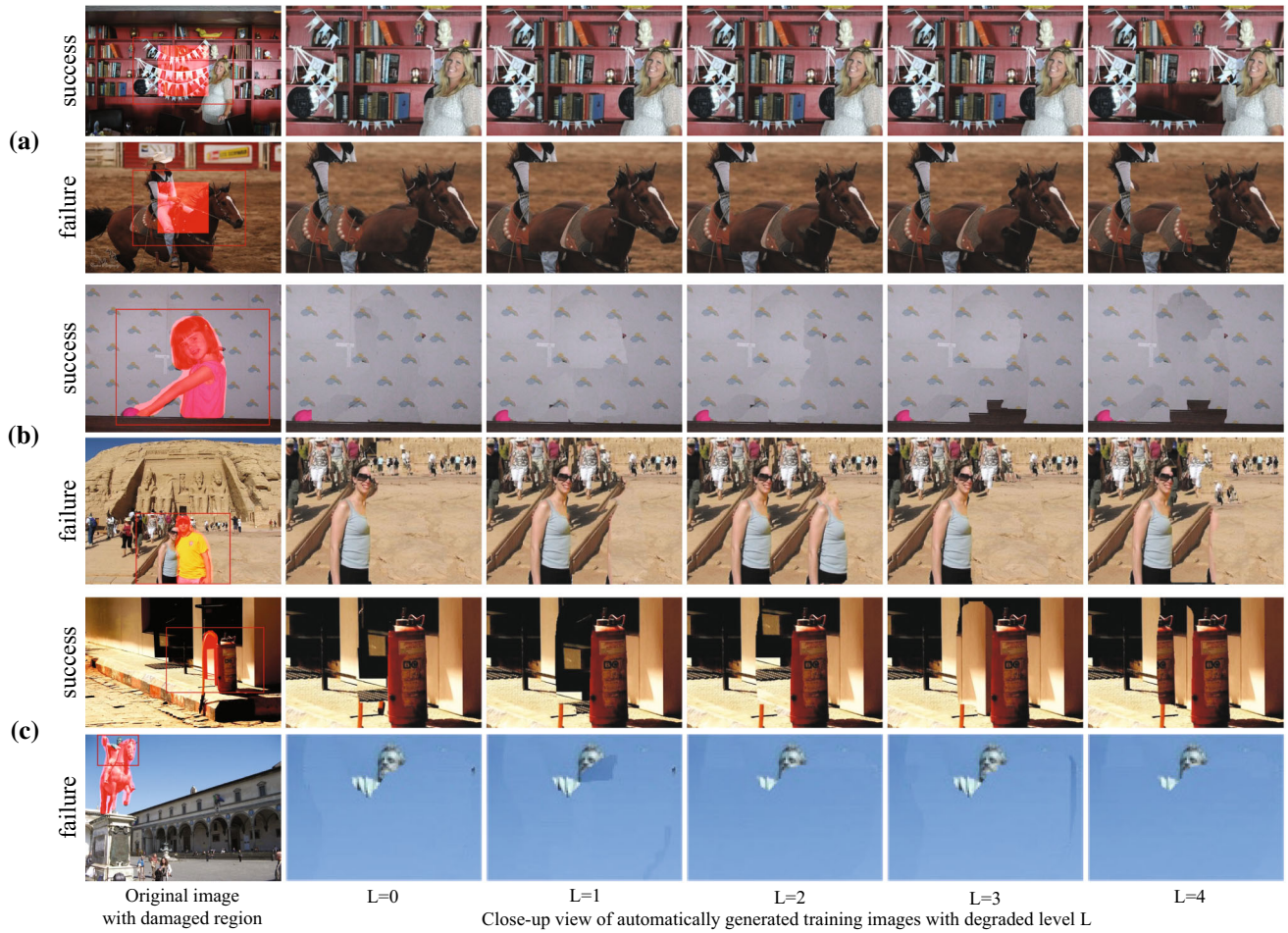

Fig. 17 Success or failure case of auto-generated training data with automatically generated masked region (**a**) I^{rect}_{mask}, **b** $I^{people(ad)}_{mask}$, and **a** I^{all}_{mask}

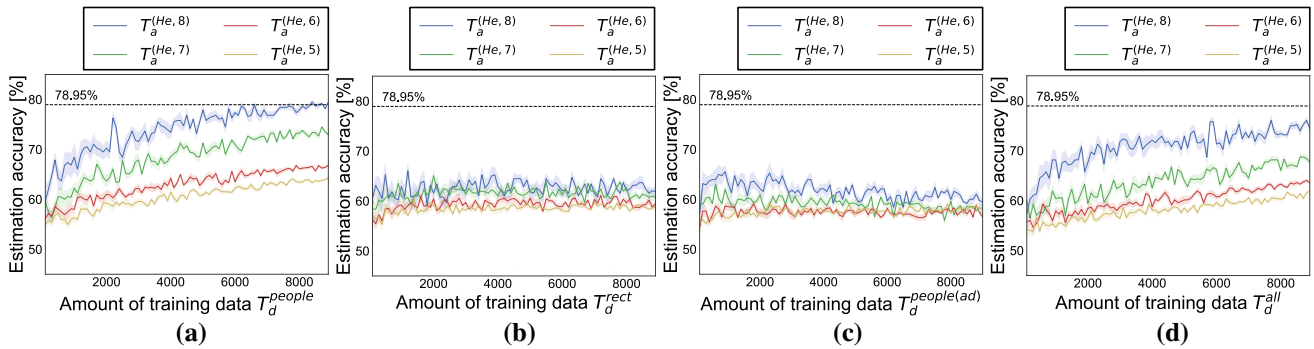

Fig. 18 Prediction accuracy for each T^S_a $(S = 5, 6, 7, 8)$ with **a** T^{people}_d, **b** T^{rect}_d, **c** $T^{people(ad)}_d$, and **d** T^{all}_d

failure cases of our mask generation. The masked people's region is protruded from the masked region and degraded inpainted images are biased towards worse quality direction. This is the current limitation of our proposed method. We are planning to optimize the masked region towards effective inpainting as a subject for future work.

6 Conclusion

This paper describes a learning-based ranking framework for image inpainting. Unlike existing learning-based IQA methods, our method trains without using subjectively annotated data by using auto-generated data; we used simulated "failed" inpainted images by focusing on inpainting algorithms. In

Table 8 Prediction accuracy comparison for auto-generated training sets with different types of masked regions (%)

	$T_a^{(He,5)}$	$T_a^{(He,6)}$	$T_a^{(He,7)}$	$T_a^{(He,8)}$
Ours(T_d^{people})	<u>64.35</u>	<u>66.84</u>	<u>72.96</u>	<u>79.47</u>
Ours(T_d^{rect})	58.80	59.41	60.70	61.84
Ours($T_d^{people(ad)}$)	57.72	57.35	58.31	59.74
Ours(T_d^{all})	61.96	63.60	67.75	74.47

The highest scores are underlined

Fig. 19 Failure case of our masked region generation. Since the people's region is protruded from masked region, multi levels of auto-generated data are biased towards worse quality direction

addition, we also proposed an automatic masked region generation method for auto-generated data, with the aim of generating completely effortless training data. Preference order estimation experiment results suggest the method's efficacy and several investigations suggest the validity of using auto-generated data instead of subjectively annotated data.

In future work we will optimize the balance between the amount of auto-generated data and their quality for our proposed system and optimize masked region towards more effective masked region generation. Applying neural network (NN)-based rank learning is also for our future work. Since our proposed data-generation method can increase the amount of training data, we argue that the NN-based ranking algorithm has a high affinity with our method. Also, we believe that the idea of generating training data by daringly generating failed images can be widely applied to other tasks requiring subjective evaluations such as image colorization (Levin et al. 2004) or image transfer (Hertzmann et al. 2001). Investigating the efficacy for these other tasks is also a subject for our future work.

Acknowledgements The authors would like to thank Dr. Shohei Mori for sharing the image inpainting implementation.

References

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).

Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., et al. (2004). Interactive digital photomontage. *ACM Transactions on Graphics*, *23*(3), 294–302.

Ardis, P. A., & Singhal, A. (2009). Visual salience metrics for image inpainting. In *Proceedings of the SPIE* (vol. 7257, pp. 72571W–72571W-9).

Barnes, C., Shechtman, E., Finkelstein, A., & Goldman, D. B. (2009). PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics*, *28*(3), 24. https://doi.org/10.1145/1531326.1531330.

Bertalmio, M., Vese, L., Sapiro, G., & Osher, S. (2003). Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing*, *12*(8), 882–889.

Chang, K., & Chen, C. (2015). A learning framework for age rank estimation based on face images with scattering transform. *IEEE Transactions on Image Processing*, *24*(3), 785–798.

Criminisi, A., Perez, P., & Toyama, K. (2004). Region filling and object removal by exemplar-based inpainting. *IEEE Transactions on Image Processing*, *13*(9), 1200–1212.

Darabi, S., Shechtman, E., Barnes, C., Goldman, D. B., & Sen, P. (2012). Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on Graphics (TOG) (Proceedings of SIGGRAPH 2012)*, *31*(4), 82:1–82:10.

Deng, J., Dong, W., Socher, R., Jia Li, L., Li, K., & Fei-fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *The IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 248–255).

Frantc, V. A., Voronin, V. V., Marchuk, V. I., Sherstobitov, A. I., Agaian, S., & Egiazarian, K. (2014). Machine learning approach for objec-

tive inpainting quality assessment. In *Proceedings of the SPIE* (vol. 9120, pp. 91200S–91200S-9).

He, K., Gkioxari, G., Dollár, P., & Girshick, R. B. (2017). Mask r-cnn. In *IEEE international conference on computer vision (ICCV)* (pp. 2980–2988). IEEE Computer Society.

He, K., & Sun, J. (2014). Image completion approaches using the statistics of similar patches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(12), 2423–2435.

Herbrich, R., Graepel, T., & Obermayer, K. (2000). *Large margin rank boundaries for ordinal regression* (chap. 7 pp. 115–132). Cambridge: MIT Press.

Herling, J., & Broll, W. (2014). High-quality real-time video inpainting with pixmix. *IEEE Transactions on Visualization and Computer Graphics*, *20*(6), 866–879.

Hertzmann, A., Jacobs, C. E., Oliver, N., Curless, B., & Salesin, D. H. (2001). Image analogies. In *Proceedings of the ACM SIGGRAPH* (pp. 327–340).

Huang, J. B., Kang, S. B., Ahuja, N., & Kopf, J. (2014). Image completion using planar structure guidance. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2014)*, *33*(4), 129:1–129:10.

Isogawa, M., Mikami, D., Takahashi, K., & Kimata, H. (2017a). Which is the better inpainted image? learning without subjective annotation. In *British machine vision conference (BMVC)* (pp. 472:1–472:10).

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).

Isogawa, M., Mikami, D., Takahashi, K., & Kojima, A. (2017b). Image and video completion via feature reduction and compensation. *Multimedia Tools and Applications*, *76*, 9443–9462.

Khosla, A., Xiao, J., Torralba, A., & Oliva, A. (2012). Memorability of image regions. In *Advances in neural information processing systems (NIPS)* (pp. 296–304).

Levin, A., Lischinski, D., & Weiss, Y. (2004). Colorization using optimization. In *Proceedings of the ACM SIGGRAPH* (pp. 689–694).

Liu, X., van de Weijer, J., & Bagdanov, A. D. (2017). Rankiqa: Learning from rankings for no-reference image quality assessment. In *IEEE international conference on computer vision (ICCV)*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *IEEE international conference on computer vision (ICCV)*.

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).

Pishchulin, L., Jain, A., Andriluka, M., Thormählen, T., & Schiele, B. (2012). Articulated people detection and pose estimation: Reshaping the future. *IEEE Conference on Computer Vision and Pattern Recognition*, *CVPR*, 3178–3185.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. *The IEEE Conference on Computer Vision and Pattern Recognition*, *CVPR*, 3234–3243.

Trung, A. T., Beghdadi, B., & Larabi, C. (2013). Perceptual quality assessment for color image inpainting. In *IEEE international conference on image processing (ICIP)* (pp. 398–402).

Tsochantaridis, I., Joachims, T., Hofmann, T., & Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *The Journal of Machine Learning Research*, *6*, 1453–1484.

Venkatesh, M. V., & Cheung, S. C. S. (2010). Eye tracking based perceptual image inpainting quality analysis. In *Proceedings of the IEEE international conference on image processing (ICIP)* (pp. 1109–1112).

Voronin, V. V., Frantc, V. A., Marchuk, V. I., Sherstobitov, A. I., & Egiazarian, K. (2015). No-reference visual quality assessment for image inpainting. *Proceedings of the SPIE*, *9399*, pp. 93990U–93990U-8.

Xu, Z., & Sun, J. (2010). Image inpainting by patch propagation using patch sparsity. *IEEE Transactions on Image Processing*, *19*(5), 1153–1165.

Yan, J., Lin, S., Kang, S. B., & Tang, X. (2014). A learning-to-rank approach for image color enhancement. *IEEE Conference on Computer Vision and Pattern Recognition*, *CVPR*, 2987–2994.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. *Intelligence*, *40*, 1452–1464.